THE THESAURUS APPROACH TO INFORMATION RETRIEVAL

T. Joyce. R.M. Needham.

An article by Dr. Vannevar Bush (1) which appeared in 1945 can be considered as the beginning of the literature on mechanised information retrieval. In this article Dr. Bush described an imaginary machine, the "Memex", *in* which a research worker would store his personal library (principally on microfilm) together with other reports, papers, and records, and from which he would be able to select instantly all references relevant to the information he desired.

Dr. Bush's article is chiefly of interest today in its account of the inadequacies of the conventional systems of library classification and the resulting tendency to neglect existing information in research work. "Even the modern great library is not generally consulted: it is nibbled at by a few ... our ineptitude at getting at the record is largely caused by the artificiality of systems of indexing."

Since the importance of this problem became widely appreciated a number of retrieval systems have been designed or suggested. These systems normally incorporate the following stages, of which some may be wholly or partly mechanised:

1) The documents or other records which are to be added to the system are processed in a certain way, and information about them which will assist in retrieval is recorded.

2) The requests for information are processed in a similar, though not necessarily identical, manner. *

3) The data resulting from (1) and (2) are then compared or matched in such a way as to distinguish those documents which are relevant to the information requests. Alternatively several lists, or an ordering of documents, may be prepared according to the degree of relevance.

4) Access to the documents referred to is then possible, or copies may be provided.

Broadly speaking, there have been two basic approaches to the first stage of information retrieval, namely the scanning or processing of the documents which it is proposed to incorporate in the system:

1) Classifying or grouping in a particular order - possibly physically on bookshelves - according to a normally preplanned classification. This is the principle upon which almost every library operates, and upon which the conventional library classifications (Dewey, Bliss, etc.) are based.

However, conventional classifications run into difficulties which do not appear to be soluble by a process of constant revision of whatever classification is employed. It is by no means always clear into which class or sub-class a document should go, and rules have to be devised to enable the arbitrary selection of one class from two or more which may be applicable. Unless there are duplicate copies or an adequate system of cross-references, this may mean that documents which are relevant to the subject dealt with within a given class are not to be found in that class, and may not be retrieved when they are wanted.

2) Indexing the documents by selecting terms (also known as descriptors, concepts, aspects) which provide a sufficient indication of the subject-matter of the document to ensure that documents will be retrieved according to specifications derived from the information requests. In most systems *of* this type documents are retrieved which have been indexed by all the terms specified, although there are alternative possibilities (see e.g. (2), (3)).

Retrieval is often carried out by superimposing punched cards or metal plates representing the terms or descriptors, in which the holes represent the documents to which the terms apply. In Zatocoding, (4), a set of random numbers is assigned to each descriptor to be encoded, and each number corresponds to a hole to be punched in a given field on cards representing the documents.

The principal difference between the various systems is in the nature of the terms which are chosen to index the documents. These vary from several thousand in the Uniterm system, with little or no attempt to remove or cater for synonyms, to a few hundred or even less than a hundred descriptors in the Zatocoding system. Uniterms have the advantage that they can be selected easily - terms which actually appear in the documents are employed and can be 'posted' on the Uniterm cards without difficulty. Descriptors (in Zatocoding, e.g.) have the advantage that they can be used consistently - the same notion will always be represented by the same descriptor or set of descriptors.

There are two main difficulties which are encountered in the application of any system of 'multiple aspect indexing':

1) It may not be clear whether a particular term should be associated with a document or not. To an extent, any decision as to what a document may bo said to be about will be subjective and may be biased according to expectations of future information requests. Admittedly it may be possible to index a research paper completely and satisfactorily by means of certain terms. In the case of a more discursive document, the decision as to which terms to employ must necessarily be more difficult. If one errs on the side of generosity in the allocation of terms, there is the likelihood of 'false drops' when any information request is dealt with; if on the other hand they are allocated more strictly, there is the danger that some documents may not be retrieved even though they are definitely relevant to the information request.

It is also clear that the phrasing of the information request will affect the nature of the problem in certain important respects. A loosely worded request may produce no documents at all, or it may produce an impossibly large number. One may say that the enquirer is getting no more than he deserves: he should learn from this experience and come back with a properly phrased request. On the other hand it may be said that it should be one of the functions of a retrieval system to produce work which has been done in a field of which the enquirer has no knowledge but is nevertheless relevant to his enquiry. One may compare the situation with that of a well-organised index, from which it is often possible to get new ideas as to relevant information. Something of this sort ought to find a place in any information retrieval system; while the enquirer should be encouraged to bring forward a tightly worded request, if he does not get what he wants at the 'first go' the system should be capable of producing a secondary list of references less directly relevant, and so on.

The random superimposed coding of Zatocoding does do something of the sort, but not in an organised way. Briefly, if the mechanical selector is set to work for a number of descriptors - say A, B and C - it will produce all the cards bearing the three descriptors together with a small number bearing only two of them, in a random way. Mooers claims that these 'false drops' have a certain value: (5)

"These samplings are very useful because they lead to information the existence of which in the file might not otherwise have been discovered. They also permit reformulations of better search prescriptions. Because the random samplings are usefully biased to the desired subjects, they have been called 'subject induced extra selections'."

2) Multiple aspect indexing will tend to produce false drops which could have been avoided if the structural content of the documents was in some way taken into account. Thus, there will be no distinction made between a paper on exports from Britain to the U.S.A., and another on exports from the U.S.A. to Britain. Also, a document giving information upon the 'economic comparison of steam railway locomotives and diesel highway truck tractors' may be retrieved in response to a request for material on the cost of operation of railway diesel locomotives (example due to Mooers). This problem is considered further below.

The Thesaurus Approach

As mentioned above, where a large number of terms are employed for the indexing process synonyms or near-synonyms are bound to occur. These may lead to the non-retrieval of relevant documents unless allowance is in some way made for them. This is also the case with general terms. For example, it is desirable that a paper dealing with 'personal income tax and social security payments' should be retrieved when there is a request for material on 'the financing of social security'. This will not happen unless all documents relating to income tax are also indexed under 'taxation', or there is a procedure for bringing out documents on income tax where 'taxation' is referred to, and no on.

The problems arising from synonymity with a large number of terms do not arise where only a few terms are employed, provided care has been taken to make them mutually exclusive. This is so with Zatocoding. On the other hand, selecting the Zatocoding descriptors when indexing each individual document requires more intelligence than selecting Uniterms from a considerably longer list, when the Uniterms selected are for the most part words which actually occur in the document.

In order to combine the advantage of the systems which employ a large number of terms with that of the systems employing a small number the suggestion has been made that a thesaurus of some kind should be employed. For example, Bernier (6) writes: "A limited thesaurus would seem to be another effective way of bringing the relevant terms to the attention of the searcher if the vocabulary proves too large to be read completely each time for selection."

A pilot scheme for a retrieval system at the library of the Radar Research Establishment at Malvern* employs about 75 terms which appear on a list which is effectively a thesaurus. The following entries, selected more or less at random, will illustrate this:

(*) Information supplied by Mr. S. Whelan of the R.R.E.

2. Add (gain, superimpose, sum, application, join, towards)

10. Calculate (compute, analog, digital, count, enumerate)

28. Generate (excitation, construct, make, produce, prepare, design)

41. Micro (miniature, small, narrow)

60. Square (area, surface, mean, square, field, plane)

73. Star (solar flares, prominences, eclipse, meteors, sun)

An alphabetic dictionary of terms which occur in the reports and requests has been drawn up, giving references to one or more of the listed head-words. For each report that is indexed, the relevant terms are selected (these can often be derived from the title or at any rate from the abstract), and the corresponding head-numbers looked up and noted. The reports are represented by holes punched in plates representing the heads. Requests are dealt with in the same way, the plates corresponding to the relevant heads being held in register so that reports are indicated by spots of light,

In a paper by H.P. Luhn. (7) the thesaurus approach is carried still further. Luhn believes that it can be applied as follows:

1) Words of similar or related meaning would be grouped into 'notional families', similar to the heads in Roget's Thesaurus.

2) The encoding of documents in terms of notional elements is then carried out by moans of the dictionary of notions, the end result being a mechanically prepared notional abstract.

3) For retrieval, the enquirer is asked to prepare an essay giving as many details as come to his mind concerning the problem. This is then encoded in the same manner, and the question notional pattern is then compared with the notional patterns of the documents. "Since an identical match is highly improbable, this process would be carried out on a statistical basis by asking for a given degree of similarity."

The Thesaurus Approach of the Cambridge Language Research Unit

As we have seen, the employment of a large number of terms when indexing a collection of documents must somehow take account of the existence of synonyms; on the other hand, the employment of a comparatively small number, particularly if the notions represented by the terms are not supposed to overlap, makes the indexing process considerably more difficult. These disadvantages can be avoided if a thesaurus is employed together with an alphabetic index which includes all the terms by which one might wish to index a document.

The thesaurus approach of the Cambridge Language Research Unit originated in three papers by M. Masterman, M.A.K. Halliday, and A.F. Parker-Rhodes presented to the M.I.T. Conference on Machine Translation in October 1956. They regarded language as consisting of words which necessarily, and as a normal thing, derive much of their significance from their context; this was in opposition to the view that words have precise meanings, some words unfortunately having several.

The developments of these ideas, which were the first use of the thesaurus in a mathematical way, may be seen in their papers (8), (9), and (10), and in a number of other papers and notes by members of the Unit (11). It is not easy to test these ideas as applied to Machine Translation without a large vocabulary and carrying through a complete and successful treatment of syntax; various tests of such procedures, which are described in the papers mentioned above, contain several kinds

of intuitive simplification and are therefore not entirely satisfactory. Nevertheless the approach obviously contains immense possibilities.

It has been widely remarked that there is a strong analogy between Machine Translation and information retrieval (R.A. Fairthorne, G. King, A. Uttley), and after conversations with, in particular, Dr. Uttley, investigations were started into an application of the approach to retrieval work, to make a useful test without depending on the results of other researches not yet done.

We first decided that it was essential to preserve as units of the system the actual key terms used in any document, so as to retain the advantages of such a system as Uniterm which can extend in any direction with ease. (The one thing which can be confidently foreseen by any librarian is that his library will extend in an unforeseen way).

It would therefore be necessary to make term abstracts of all the documents - since the choice of terms is not limited this presents no great difficulty - and to start from the term vocabulary found in them.

This term vocabulary was then to be arranged so that the property of accommodating near-synonyms held at all levels. It appeared that this could be done by arranging the words under a partial-ordering relation, put informally thus: 'If you ask for A you mustn't complain if you get B' = $A \geq B$. If you ask for something about Russian grammar you can reasonably be given something about Italian nouns. Also, if you ask either for something about mechanical processes, or about translation, you can hope and expect to be given something about Machine Translation. The difference between this arrangement of terms and that described by R.H. Richens in (12), in which the terms were arranged according to the hierarchic classification of the U.D.C., is that terms representing the meets of classes represented by other terms are freely employed, as well as the joins. This eliminates the difficulty referred to above which is inevitably encountered in a hierarchic classification, that it may be difficult to decide into which sub-class a document should go, while at the sane time making it possible to make allowances for structure - as we shall see below.

To make use of the convenient algebraic properties of lattices, the figure this yields can be readily turned into a lattice by including latent elements where needed to satisfy the lattice axioms. (We find that the idea of a term vocabulary as a lattice of this kind occurs both in Fairthorne (3) and Mooers (5).)

Terms are only to be treated as synonymous if it appears that, for any conceivable extension of the library, there will never be any point in distinguishing between them. For example, in a Machine Translation vocabulary, 'multiple meaning problem', 'plurivalance of meaning', and 'ambiguity' (in one sense) would be regarded as synonymous. Since in all other cases the actual terms in the vocabulary are treated as distinct, although of course they may be close to one another in the lattice, it is possible to bring forward a very precise information request - but at the same time some procedure for providing a scale of relevance is necessary. Since considerable accuracy in specifying both documents and requests in the terms of the system is possible, the system cannot be a 'one-shot' one (that is: no initial output, no retrieval) in case the request is worded too exclusively.

As in other systems the documents are represented by holes in punched cards which represent the various terms, and in addition, when a hole is punched in any term card, all the cards representing terms at higher levels of the lattice such that the inclusion relation holds between them and the original term are also punched. This can be easily accomplished if there is a suitable system of cross-references among the term cards themselves. A term abstract is then made of each information request received, the corresponding term cards are removed from the card file and held in register, and the first output (if any) is recorded.

The original course used to produce a scale of relevance (8) would in the present context work as follows: from the terms of the request select all the appropriate term cards, including those to which there are cross-references (i.e. those representing all terms above the original terms in the lattice). Thus for a request for material on 'Mechanical Translation of Russian prepositions', the cards for Machine Translation, Russian, proposition, and also for machines, translation, languages, parts of speech, word classes, grammar, linguistics, and language would be withdrawn from the card file. These would then be superimposed in all possible pairs, and all the documents given would be noted. The documents would then be consulted in order of frequency of appearance in the list of outputs. Clearly this process would be very laborious for requests involving any number of terms; fortunately various simplifications can be made for practical use.

Firstly it can be shown that it is not necessary to compare the cards in pairs; the same scale will result from simply counting the occurrences of the document holes in the same set of cards. This might be useful for a mechanical method, but for hand use another equivalent method is more convenient. Under certain conditions, usually satisfied*, the first two stages of the scale are given by the following procedure. Take as most relevant the set given by superimposing the actual cards representing the terms of the request. Then substituting for each card in turn a card covering it in the lattice, and note the set of outputs. This second set of outputs, having substituted all the covering elements, will constitute the second relevance class. This proves in practice to be far enough along the scale of relevance to give the documents needed. For the request given above it yields the desired result in 5 'peek-a-boo' operations as against 11 counts and an addition.

Structure of information

It has been remarked that, particularly in large libraries, a high proportion of false drops could be caused by failure of the system to take account of the structure of the documents. There will come a time when there will be, for example, so many documents with the notional abstract 'a,b,c,d,e,f,' that it is necessary to distinguish between them when this cannot effectively be done simply by including more terms.

One way of doing this would be to take account of the frequency of occurrence of the different terms. The matching operation would then be between vectors the elements of which are frequencies, and the matching relation would be the vector distance relation. The trouble here would be that distance between long and short books on the same subject would be large.

(*) The condition is that if S is the set of terms of the request and $J(\Sigma)$ is the set of elements of the lattice which are a, some $a \epsilon \Sigma$, and $J'(\Sigma)$ the set obtained by removing from $J(\Sigma)$ the elements substituted for, then $J(J'(\Sigma)) = J(\Sigma)$. (See (13)).

This could perhaps be overcome if the frequency vectors were normalised, or logarithms of frequencies were used. This method of giving both a scale of relevance and a reference to structure has not been tested because it is not physically easy to do; it does however appear to be a most complete treatment and will be tried if the more tractable methods prove insufficient.

Another way is to treat a set of terms as a single term whose place in the lattice is the lattice meet of the terms of the set. An example is provided by the terms 'Machine' and 'Translation'. These have a meet 'MT'. Thus a document about MT will be coded as MT, and a document which has references both to M *&* T but not to MT will be coded under M & T separately. A request involving MT will then not yield the latter at the first stages of relevance. This process may be elaborated to any extent that is found necessary. It is structurally similar to that proposed by Mooers (5), involving interlocking descriptor sets, and the coding of n-tuples taken from these as if they were themselves descriptors, and may be illustrated by the same kind of example. The use of groups of terms or descriptors in this way, which may be made a fairly elegant process although there is an essential arbitrariness in deciding when to apply it, seems to be the best that can be done without coding the documents in a way that will be to some extent message-preserving. The difficulties in the way of message-preserving abstracting and coding are well known; the one that is most important in this context is perhaps that the operations may no longer bo carried out independently of the order of the cards and terms, and some asymmetrical operations will be necessary. All this detracts very much from the manipulative simplicity of the system.

Conclusion

The system given above has been and is being tested on the offprint library of the Cambridge Language Research Unit. Investigations are continuing into the use of compressed coding to lessen the physical work. A description of the technique and apparatus will form another communication.

The tests in progress, together with comparison with the published details of other systems, lead *us* to claim the following advantages for our system:

1) It has the advantages of descriptor or head systems in its treatment of related terms without
   a) the difficulty of abstracting for them;
   b) leaving the advantages of their hierarchies unrecognised;
   c) their inflexibility in an expanding library.
2) It uses a scale of relevance and does not therefore fail if there is no initial output.
3) It retains the advantage of 'Uniterm' in that it cannot be caught out by an unforeseen change in the structure of the library.
4) It can deal with a request in general terms without producing at once all the most detailed work on the subject. This is illustrated by the classification we have set up, which will be discussed in detail in the following paper.

Cambridge Language Research Unit.
October 1957-

References.

(1) Bush, V. "As we may think", Atlantic Monthly, 176 (1945), 101-108.

(2) Perry, J.W., Kent, A. & Berry, M. "Machine Literature Searching" New York, 1956.

(3) Fairthorne, R.A. "The patterns of retrieval", Amer.Doc. 7, 1956, 65-70.

(4) Mooers, C.N. "Zatocoding and developments in Information Retrieval", Aslib Proceedings 8, 1956, 1-20.

(5) Mooers, C.N. "Information Retrieval on Structured Content", 3rd London Symposium on Information Theory, '55, 121-134.

(6) Bernier, G.L. "Correlative Indexes II: Correlative trope indexes", Amer.Doc. 8, 1957, 103-122.

(7) Luhn, H.P. "A statistical approach to mechanised literature searching", I.B.M. Research Center, N.Y. 1957.

(8) Masterman, M. "Potentialities of a Mechanical Thesaurus", M.I.T. Conference on MT, 1956.

(9) Halliday, M.A.K. "The thesaurus type mechanical dictionary and application to English Preposition Classification", M.I.T. Conference on MT, 1956.

(10) Parker-Rhodes, A.F. "An Algebraic Thesaurus" M.I.T. Conference on MT, 1956.

(11) C.L.R.U. series on Mechanical study of context, privately circulated, 1957.

(12) Richens, R.E. "An Abstracting and Information Service for Plant Breeding and Genetics", in Casey & Perry, "Punched cards and their application in Science and Industry", N.Y. 1951.

(13) Needham, R.M. "A property of finite lattices", C.L.R.U. research note, 1957.