INFORMATION RETRIEVAL AND THE THESAURUS

R.M. Needham T. Joyce
Cambridge Language Research Unit.

Introduction

This paper describes various developments of the retrieval system devised in Cambridge last year, which we described in a paper 'The Thesaurus Approach to Information Retrieval' (Amer. Doc. 1958). These are in part concerned with mechanically setting up the system, and in part with the interesting possibilities of achieving better retrieval by having more flexible search procedures rather than by more elaborate indexing. To make this paper comprehensible on its own we have included here a brief summary of the theoretical part of the earlier description. The reader is, however, referred to that paper for the details and background of the method.

Nature and purpose of the CLRU system

We set out to provide a method of indexing documents which would avoid the main difficulties of 'multiple aspect indexing':

1) If the indexing is done by 'terms' or 'descriptors' it may be difficult in encoding a document or request to decide precisely which terms should be used. It is desirable to use a method which is not very sensitive to subjective variations in the coding - that is one where these variations (and also clerical blunders) do not produce losses or false drops. Allied to this is the requirement that the enquirer should not obtain unsatisfactory results because he can only specify in rather a tentative way what he wants, or because he uses an unfamiliar terminology. For example, workers on MT use a variety of terms for a kind of basic sentence structure - sentence core, kernel, and so on. An enquirer for work on this subject should be able to get all the relevant documents while only using in his request the term with which he is familiar.

2) If, in order to provide the flexibility just mentioned, terms of an inclusive application are used (Mooers (1), Whelan (2)), it is desirable to avoid a difficulty which can arise when the library expands. It may be that some descriptor has been used with a large range of application in some subject in which the library was not very interested when the scheme was set up. If the library then expands so that much more detailed retrieval is needed in that subject, the librarian does not want to have to re-read and re-classify a large number of earlier documents.

Previously described systems have tended to avoid one or the other of these troubles but not both. For example 'Uniterm' avoids 2) and the 'Zator' system avoids 1). We shall in describing our work refer to these difficulties simply as (1)& (2).

We first decided that it was necessary to retain in the system the actual key terms used in the documents - thus dealing with (2), Therefore *term abstracts* of the documents were made which were simply lists of the significant terms used. From this was prepared the term vocabulary of the library. The term vocabulary was then 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 $B' \equiv A \geq B$. If you ask for something on Russian grammar you can reasonably be given something about Russian nouns. Also, if you ask *either* for something on mechanical processes, *or* for something on translation, you can hope and expect to be given something on Machine Translation. The partial-ordering achieved by this means puts near-synonyms immediately below the same covering term, and if the system is given additional elements so as to make it strictly a lattice they turn out to have a very small lattice distance between them. Now in order to deal with (1) it is necessary to have a 'scale of relevance' procedure for going from a particular request to closely allied ones, and so on.

For details of the process the reader is referred to our earlier paper, from which the following is an extract:

> 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 terms 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 easily be 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 then removed from the card file and held in register, and the output (if any) recorded........Take as most relevant the set (of outputs) given by superimposing the actual cards representing the terms of the request. Then substitute for each card in turn a card covering it in the lattice, and note the set of outputs. The second sets of outputs, having substituted all the covering elements in turn, will constitute the second relevance class.

The method of dealing with what Mooers describes as 'structured content' will not be described again here.

Tests of the system

A small-scale test of the system has been carried out on the CLRU offprint library; work is in progress on increasing the size of the sample. Several interesting points came up in the testing (3), which was mainly designed to detect and remove clerical errors. Firstly, again a contribution to the avoidance of (1), the system is very insensitive to mistakes. For the kind of mistakes that happen are:

a. Errors of term-abstracting. These are not infrequent, as great speed in abstracting is desirable. We made the abstracts of offprints in about 4 min. each, and this is too fast for great accuracy.
b. Errors of punching. In their effects on the system these are equivalent to errors of abstracting.
c. Errors of term-abstracting of requests. These are fairly rare.

In the ease of errors of term-abstracting if a term is inserted which ought not to be there, an extra hole will be punched on the cards for that term and for all terms above it. A false drop will only result from this if the absence of this term is the only reason for not wanting a document; this may be better put by saying that any particular document is usually over-specified by the terms of the request. If, however, a term is omitted which ought to be included in an abstract, a failure of retrieval at the first stage is likely to result. However, there is unlikely to be a permanent loss, for the following reason. Any paper is on some particular subject; its terms tend to lie mainly in the same region of the lattice. Thus as we go up the lattice from the term in the request which is causing the trouble (without of course knowing which it is) we shall come to a card which has the required document punched on it because of another low-level term. Thus we shall retrieve the required document later than we otherwise should. It seems empirically to be the case that if the scale of relevance is pursued until obviously irrelevant material is given, all the relevant material will have emerged. Notice that there can strictly speaking be no losses in the system for the scale of relevance if taken all the way eventually retrieves the whole library.

For these reasons the system is not very sensitive to blunders. It is of course desirable to be as accurate as possible, but it is an asset that the effects of a small error are unlikely to be disastrous.

It also emerged from the tests that the scheme is very effective at dealing with requests for which the library has no answer. It produces at first nothing, and later papers on related topics. An example is a request for 'Stochastic methods of information retrieval'. There is nothing on this in the library; no one has tried searching a library in a random way until he finds what he wants; so the immediate output is null. At the later stage, however, we get papers on stochastic methods of MT, on statistics and literature searching, and on retrieval with random superimposed coding. These are the best the library can do for the enquirer.

Treatment of general terms

A problem arises from the lattice structure when very general terms are used in a request. A term such as 'logic' which comes high up in the lattice has punched on its card holes representing documents which are not directly about logic at all; for instance a document including 'sentence core' finishes up punched on 'logic' via 'operational syntax', 'logical operations' and 'formal logic'. But it is fairly certain that an enquirer will not be in any sense thinking of 'sentence cores' when he uses the word 'logic' in his request. We have successfully dealt with this by a very simple device. The term which was 'logic' is turned into a latent element 'L', and a new element is put immediately below it called 'logic'. A document will only have a hole on this new card if it actually has the term 'logic' in its term abstract. So if a request has the word 'logic' in it, the card which is extracted first is the new 'logic' card, and the all-inclusive 'L' is only brought in at the second stage of relevance. In the cases where we have applied this device it has been very successful. What it achieves *may* perhaps be more elegantly achieved by means of lattice-operations in a computer, but this simple expedient seems adequate for hand operation.

Treatment of general documents

There is a class of documents which give great trouble to retrieval systems because they are about a very large variety of topics, and so tend to be retrieved often, although they are rarely what the enquirer wants. The typical specimen of this type is the annual progress report of a research establishment. The trouble does not arise so much with books, for they may be encoded chapter by chapter as a series of separate documents, nor does it arise with documents whose terms can be grouped as described in the section on structure in our earlier paper (4).

The difficulty may be averted if, through mechanising the system, different types of search procedure can be specified for different types of request. This possibility is explored below.

Further mechanisation

1) Aids to setting up the system

A drawback of the system we propose is the amount of work involved in setting it up. When it is compared with 'Uniterm', which in this respect it resembles, the additional work is seen to lie in the preparation of the lattice and the transferring of the punchings from one card to another. The latter may be ignored, as it can be done by a standard punched-card copier. The former, however, is more serious. In the pilot project it was achieved by considering the terms, and noticing relations between them from our knowledge of the uses of the words. This suggests that it might be possible to prepare the lattice from a suitable dictionary, for this is where one looks to find the uses of words. Now it is desirable to see whether it is possible to obtain suitable information from a general purpose dictionary, and in order to do this we are testing a programme on Edsac 2 to analyse data obtained from general linguistic knowledge. We assume that the kind of data that we are interested in can be represented as a set for ordering relations between pairs of terms, and we wish to find out whether consistent and correct sets of relations can be obtained in this way. This assumption seems justified on empirical grounds; it is much easier to set up ordered pairs of terms, such as 'nouns ≤ grammar', 'grammar ≤ linguistics', considered independently of all other terms, than it is to try to set up a system as a whole. Accordingly we take a list of terms, attach arbitrary numbers to them, and prepare a list of ordered pairs representing the partial-ordering relations between them. The programme then tests this set of relations to see whether it is a genuine partial-ordering, that is to see whether the relation obeys the transitivity axiom.

We shall be by November in a position to give the results of tests of this kind, the relations being derived both from a dictionary and from general knowledge. It is doubtful whether we shall by then have been able to test the process using a completely mechanical dictionary, for such a thing does not yet exist.

As far as hand operation is concerned, there is no need to go any further in testing the data; for any mechanisation of the operation of the system it will probably be necessary to ensure that the set of terms forms a lattice.

2) Mechanisation of the operation of the system

The problem of mechanising further the operation of the system is in effect the problem of performing the operations of Non-Boolean lattice algebras in a cheap and efficient way. This is not the place to give a detailed description of the processes involved in general lattice algebra; the Unit's work on lattice encoding and efficient computing is described in a workpaper (5). We may however discuss here the uses that could be made of a system mechanised in that way to obtain better results by means of more elaborate searching procedures.

To make clear the advantages that may be gained by a more inclusive armoury of lattice operations, we first consider the search process in lattice terms. Any set of terms may in theory correspond to a document. So the potential documents may be represented as the points of the Boolean lattice of meets of sets of terms, and a subset of these points will correspond to the actual documents of the library.

Now any request, in its given form, specifies a point of this lattice, which will not in general correspond to a document: our problem is to extract the documents which are in some useful sense 'nearest' to the point specified. We may direct a search for near documents in three ways:- upwards, that is through points including the 'request point', which corresponds to leaving terms out of the request; downwards, that is through points included by the request point, which corresponds to including more terms; and sideways, which corresponds to replacing terms of the request by related terms as in the scale of relevance procedure described above.

Now the punched card procedure is Incapable of performing the second mode of search; indeed it always produces at once all documents which would be given by all the stages of that mode. If by means of more elaborate machinery all three modes are possible, it becomes practicable to use different searching processes to suit the requests. Two examples are:-

1) Requests for works of a general nature, elementary introductions, etc»

Retrieve documents having all terms of the request, and as many others as possible. Order the output so that the documents with most additional terms come first.

If necessary repeat using the standard scale of relevance procedure.

2) Requests for specialised work.

Retrieve documents having all terms of the request and as few others as possible: order the output so that the documents with least additional terms come first. Again in the standard scale of relevance procedure may be superimposed on this.

Both these revised searching systems are very simply expressed in terms of lattice operations; they amount to searching for elements below a given element (here the meet of terms of the request), starting in 1) at the 0-element of the lattice and in 2) at the given element. For the details of the computing procedure the reader is referred to the relevant workpaper ( ).

Conclusion

Since the process of mechanised retrieval is becoming better understood, we think emphasis should be laid on developing more flexible retrieval processes, so that various of the finer grades of retrieval may be achieved by that means rather than by more complicated indexing of documents. We have tried to show how such questions as the retrieval of general or specialist documents may be dealt with by altering the retrieval strategy, instead of such inelegant devices as having a class called 'general works'. However, much remains to be done, particularly in connection with setting up the system economically for an existing library.

REFERENCES

GENERAL. M. MASTERMAN Potentialities of a Mechanical Thesaurus.
A.F. PARKER-RHODES An Algebraic Thesaurus.
(Both read at 2nd International Conference on Machine Translation, M.I.T. 1956)
The C.L.R.U. Workpapers on Mechanical Study of Context.

CITED IN TEXT.

(1) MOOERS, C.N. "Zatocoding and Information Retrieval" Aslib Proc. 8, 1956.

(2) WHELAN, S. Paper for this conference,

(3) MILLER, A.H.J. Tests of C.L.R.U. Retrieval System. C.L.R.U. Workpaper 1957.

(4) JOYCE, T. & NEEDHAM, R.M., Thesaurus Approach to Information Retrieval. Amer. Doc. 1958.

(5) PARKER-RHODES, A.F. & NEEDHAM, R.M. Methods of Performing Non-Boolean Lattice Algebra on a Digital Computer. C.L.R.U. Workpaper 1958.