Eight
THE FUTURE

Future research

In the preceding chapters I have tried to bring together some of the more elaborate tools that are used during the design of an experimental information retrieval system. Many of the tools themselves are only at the experimental stage and research is still needed, not only to develop a proper understanding of them, but also to work out their implications for IR systems present and future. Perhaps I can briefly indicate some of the topics which invite further research.

1. Automatic classification

Substantial evidence that large document collections can be handled successfully by means of automatic classification will encourage new work into ways of structuring such collections. It could also be expected to boost commercial interest and along with it the support for further development.

It is therefore of some importance that using the kind of data already in existence, that is using document descriptions in terms of keywords, we establish that document clustering on large document collections can be both effective and efficient. This means more research is needed to devise ways of speeding up clustering algorithms without sacrificing too much structure in the data. It may be possible to design probabilistic algorithms for clustering procedures which will compute a classification on the average in less time than it may require for the worst case. For example, it may be possible to cut down the $0(n^2)$ computation time to *expected* $0(n\log n)$, although for some pathological cases it would still require $0(n^2)$. Another way of approaching this problem of speeding up clustering is to look for what one might call almost classifications. It may be possible to compute classification structures which are close to the theoretical structure sought, but are only close approximations which can be computed more efficiently than the ideal.

A big question, that has not yet received much attention, concerns the extent to which retrieval effectiveness is limited by the type of document description used. The use of keywords to describe documents has affected the way in which the design of an automatic classification system has been approached. It is possible that in the future, documents will be represented inside a computer entirely differently. Will grouping of documents still be of interest? I think that it will.

Document classification is a special case of a more general process which would also attempt to exploit relationships between documents. It so happens that dissimilarity coefficients have been used to express a distance-like relationship. Quantifying the relationship in this way has in part been dictated by the nature of the language in which the documents are described. However, were it the case that documents were represented not by keywords but in some other way, perhaps in a more complex language, then relationships between documents would probably best be measured differently as well. Consequently, the structure to represent the relationships might not be a simple hierarchy, except perhaps as a special case. In other words, one should approach document clustering as a process of finding structure in the data which can be exploited to make retrieval both effective and efficient.

An argument parallel to the one in the last paragraph could be given for automatic keyword classification, which in the more general context might be called automatic 'content unit' classification. The methods of handling keywords, which are being and have already been developed, will also address themselves to the automatic construction of classes of

'content units' to be exploited during retrieval. Keyword classification will then remain as a special case.

H. A. Simon in his book The Sciences of the Artificial defined an interesting structure closely related to a classificatory system, namely, that of a *nearly decomposable system*. Such a system is one consisting of subsystems for which the interactions among subsystems is of a different order of magnitude from that of the interactions within subsystems. The analogy with a classification is obvious if one looks upon classes as subsystems. Simon conceived of nearly decomposable systems as ways of describing dynamic systems. The relevant properties are (a) in a nearly decomposable system, the short-run behaviour of each of the component subsystems is approximately independent of the short-run behaviour of the other components; (b) in the long run, the behaviour of any one of the components depends in only an aggregate way on the behaviour of the other components. Now it may be that this is an appropriate analogy for looking at the dynamic behaviour (e.g. updating, change of vocabulary) of document or keyword classifications. Very little is in fact known about the behaviour of classification structures in dynamic environments.

2. File structures

On the file structure chosen and the way it is used depends the efficiency of an information retrieval system.

Inverted files have been rather popular in IR systems. Certainly, in systems based on unweighted keywords especially where queries are formulated in Boolean expressions, an inverted file can give very fast response. Unfortunately, it is not possible to achieve an efficient adaptation of an inverted file to deal with the matching of more elaborate document and query descriptions such as weighted keywords. Research into file structures which could efficiently cope with the more complicated document and query descriptions is still needed. The only way of getting at this may be to start with a document classification and investigate file structures appropriate for it. Along this line it might well prove fruitful to investigate the relationship between document clustering and relational data bases which organise their data according to *n*-ary relations.

There are many more problems in this area which are of interest to IR systems. For example, the physical organisation of large hierarchic structures appropriate to information retrieval is an interesting one. How is one to optimise allocation of storage to a hierarchy if it is to be stored on devices which have different speeds of access?

3. Search strategies

So far fairly simple search strategies have been tried. They have varied between simple serial searches and the cluster-based strategies described in Chapter 5. Tied up with each cluster-based strategy is its method of cluster representation. By changing the cluster representative, the decision and stopping rules of search strategies can usually also be changed. One approach that does not seem to have been tried would involve having a number of cluster representatives each perhaps derived from the data according to different principles.

Probabilistic search strategies have not been investigated much either* , although such strategies have been tried with some effect in the fields of pattern recognition and automatic medical diagnosis. Of course, in these fields the object descriptions are more detailed than are the document descriptions in IR, which may mean that for these strategies to work in IR we may require the document descriptions to increase in detail.

---

* The work described in Chapter 6 goes some way to remedying this situation.

In Chapter 5 I mentioned that bottom-up search strategies are apparently more successful than the more traditional top-down searches. This leads me to speculate than it may well be that a spanning tree on the documents could be an effective structure for guiding a search for relevant documents. A search strategy based on a spanning tree for the documents may well be able to use the dependence information derived from the spanning tree for the index terms. An interesting research problem would be to see if by allowing some kind of interaction between the two spanning trees one could improve retrieval effectiveness.

4. Simulation

The three areas of research discussed so far could fruitfully be explored through a simulation model. We now have sufficiently detailed knowledge to enable us to specify a reasonable simulation model of an IR system. For example, the shape of the distributions of keywords throughout a document collection is known to influence retrieval effectiveness. By varying these distributions what can one expect to happen to document or keyword classifications? It may be possible to devise more efficient file structures by studying the performance of various file structures while simulating different keyword distributions.

One major open problem is the simulation of relevance. To my knowledge no one has been able to simulate the characteristics of relevant documents successfully. Once this problem has been cracked it opens the way to studying such hypotheses as the Cluster and Association hypothesis by simulation.

5. Evaluation

This has been the most troublesome area in IR. It is now generally agreed that one should be able to do some sort of cost-benefit, or efficiency-effectiveness analysis, of a retrieval system.

In basing a theory of evaluation on the theory of measurement, is it possible to devise a measure of effectiveness not starting with precision and recall but simply with the set of relevant documents and the set of retrieved documents? If so, can we generalise such a measure to take account of degree of relevance? An alternative derivation of an E-type measure could be done in terms of recall and fallout. Is there any advantage to doing this?

Up to now the measurement of effectiveness has proved fairly intractable to statistical analysis. This has been mainly because no reasonable underlying statistical model can be found, however, that is not to say that one does not exist![*]

There may be 'laws' of retrieval such as the well known trade-off between precision and recall that are worth establishing either empirically or by theoretical argument. It has been shown that the trade-off does in fact follow from more basic assumptions about the retrieval model. Similar arguments are needed to establish the upper bounds to retrieval under certain models.

6. Content analysis

There is a need for more intensive research into the problems of what to use to represent the content of documents in a computer.

Information retrieval systems, both operational and experimental, have been keyword based. Some have become quite sophisticated in their use of keywords, for example, they

---

[*] I think the Robertson model described in Chapter 7 goes some way to being considered as a reasonable statistical model.

may include a form of normalisation and some sort of weighting. Some use distributional information to measure the strength of relationships between keywords or between the keyword descriptions of documents. The limit of our ingenuity with keywords seemed to have been reached when a few semantic relationships between words were defined and exploited.

The major reason for this rather simple-minded approach to document retrieval is a very good one. Most of the experimental evidence over the last decade has pointed to the superiority of this approach over the possible alternatives. Nevertheless there is room for more spectacular improvements. It seems that at the root of retrieval effectiveness lies the adequacy (or inadequacy) of the computer representation of documents. No doubt this was recognised to be true in the early days but attempts at that time to move away from keyword representation met with little success. Despite this I would like to see research in IR take another good look at the problem of what should be stored inside the computer.

The time is ripe for another attempt at using natural language to represent documents inside a computer. There is reason for optimism now that a lot more is known about the syntax and semantics of language. We have new sources of ideas in the advances which have been made in other disciplines. In artificial intelligence, work has been directed towards programming a computer to understand natural language. Mechanical procedures for processing (and understanding) natural language are being devised. Similarly, in psycho-linguistics the mechanism by which the human brain understands language is being investigated. Admittedly the way in which developments in these fields can be applied to IR is not immediately obvious, but clearly they are relevant and therefore deserve consideration.

It has never been assumed that a retrieval system should attempt to 'understand' the content of a document. Most IR systems at the moment merely aim at a bibliographic search. Documents are deemed to be relevant on the basis of a superficial description. I do not suggest that it is going to be a simple matter to program a computer to understand documents. What is suggested is that some attempt should be made to construct something like a naïve model, using more than just keywords, of the content of each document in the system. The more sophisticated question-answering systems do something very similar. They have a model of their universe of discourse and can answer questions about it, and can incorporate new facts and rules as they become available.

Such an approach would make 'feedback' a major tool. Feedback, as used currently, is based on the assumption that a user will be able to establish the relevance of a document on the basis of data, like its title, its abstract, and/or the list of terms by which it has been indexed. This works to an extent but is inadequate. If the content of the document were understood by the machine, its relevance could easily be discovered by the user. When he retrieved a document, he could ask some simple questions about it and thus establish its relevance and importance with confidence.

Future developments

Much of the work in IR has suffered from the difficulty of comparing retrieval results. Experiments have been done with a large variety of document collections, and rarely has the same document collection been used in quite the same form in more than one piece of research. Therefore one is always left with the suspicion that worker A's results may be data specific and that were he to test them on worker B's date, they would not hold.

The lesson that is to be learnt is that should new research get underway it will be very important to have a suitable data-base ready. I have in mind a natural-language document

collection, probably using the full test of each document. It should be constructed with many applications in mind and then be made universally available.*

Information retrieval systems are likely to play an every increasing part in the community. They are likely to be on-line and interactive. The hardware to accomplish this is already available but its universal implementation will only follow after it has been made commercially viable.

One major recent development is that computers and data-bases are becoming linked into networks. It is foreseeable that individuals will have access to these networks through their private telephones and use normal television sets as output devices. The main impact of this for IR systems will be that they will have to be simple to communicate with, which means they will have to use ordinary language, and they will have to be competent in their ability to provide relevant information. The VIEWDATA system provided by the British Post Office is a good example of a system that will need to satisfy these demands.

By extending the user population to include the non-specialist, it is likely that an IR system will be expected to provide not just a citation, but a display of the text, or part of it, and perhaps answer simple questions about the retrieved documents. Even specialists may well desire of an IR system that it do more than just retrieve citations.

To bring all this about the document retrieval system will have to be interfaced and integrated with data retrieval systems, to give access to facts related to those in the documents. An obvious application lies in a chemical or medical retrieval system. Suppose a person has retrieved a set of documents about a specific chemical compound, and that perhaps some spectral data was given. He may like to consult a data retrieval system giving him details about related compounds. Or he may want to go on-line to, say, DENDRAL which will give him a list of possible compounds consistent with the spectral data. Finally, he may wish to do some statistical analysis of the data contained in the documents. For this he will need access to a set of statistical programs.

Another example can be found in the context of computer-aided instruction, where it is clearly a good idea to give a student access to a document retrieval system which will provide him with further reading on a topic of his immediate interest. The main thrust of these examples is that an important consideration in the design of a retrieval system should be the manner in which it can be interfaced with other systems.

Although the networking of medium sized computers has made headline news, and individuals and institutions have been urged to buy into a network as a way of achieving access to a number of computers, it is by no means clear that this will always be the best strategy. Quite recently a revolution has taken place in the mini-computer market. It is now possible to buy a moderately powerful computer for a relatively small outlay. Since information channels are likely to be routed through libraries for some time to come, it is interesting to think about the way in which the cheaper hardware may affect their future role. Libraries have been keen to provide users with access to large data-bases, stored and controlled some where else often situated at a great distance, possibly even in another country. One option libraries have is the one I have just mentioned, that is, they could connect a console into a large network. An alternative, and more flexible approach, would be for them to have a mini-computer maintaining access to a small, recently published chunk of the document collection. They would be able to change it periodically. The mini would be part of the network but the user would have the option of invoking the local or global system. The local system could then be tailored to local needs which would give it an important advantage. Such things as personal files, containing say user profiles

---

* A study recommending the provision of such an experimental test bed has recently been completed, see Sparck Jones and van Rijsbergen, 'Information retrieval test collections', *Journal of Documentation*, **32**, 59-75 (1976).

could be maintained on the mini. In addition, if the local library's catalogue and subject index were available on-line, it would prove very useful in conjunction with the document retrieval system. A user could quickly check whether the library had copies of the documents retrieved as well as any related books.

Another hardware development likely to influence the development of IR systems is the marketing of cheap micro-processors. Because these cost so little now, many people have been thinking of designing 'intelligent' terminals to IR systems, that is, ones which are able to do some of the processing instead of leaving it all the main computer. One effect of this may well be that some of the so-called more expensive operations can now be carried out at the terminal, whereas previously they would have been prohibited.

As automation advances, much lip service is paid to the likely benefit to society. It is an unfortunate fact that so much modern technology is established before we can actually assess whether or not we want it. In the case of information retrieval systems, there is still time to predict and investigate their impact. If we think that IR systems will make an important contribution, we ought to be clear about what it is we are going to provide and why it will be an improvement on the conventional methods of retrieving information.