CHAPTER 10

BASIC PROBLEMS OF INFORMATION RETRIEVAL

In this concluding chapter, an attempt will be made to survey the basic problems of information retrieval. Many of our statements will have been made before, and much of it may appear to be in the nature of a "child's guide" to the subject. Certainly, however, for many of our assertions it would be easier to find contradictory statements rather than supporting statements, but it is hoped that the experimental results obtained in the investigation will provide supporting evidence that has previously been lacking.

The discussion can be grouped into three main aspects. First there are the requisites needed for an I.R. retrieval system. Secondly there are the essential operations to be carried out and thirdly there is the question of system performance. Inevitably, due to the interaction between these various aspects, it will not be possible to maintain these divisions throughout the discussion.

The basic requisites for an I.R. system are that there should be a group of potential users and a collection of documents, (which we will define as being a general term to cover any item bearing intelligible marks) that can potentially meet some or all of the information requirements of the users. There must be a physical environment to store the documents in whatever form they may be, and so far we have come to the stage of describing a 'library' in the classical sense of being an unexploited collection of documents. To turn this into an information retrieval system, we require that the useful information content of the documents should be recorded in a descriptor language and that a store should be available for these records. The descriptor language may be basically simple or have a complex structure and may be operated at a shallow or deep level. The store may be one of many different kinds, such as a printed volume, a conventional card catalogue, a collection of edge- or centre- punched cards or a computer.

The essential operations of the information retrieval system which we wish to consider are limited to those which have been covered in this investigation and which are considered to form the core of the whole problem. There are a number of important actions concerned with acquisition which precede the first operation, but whatever these may be, eventually they lead to the stage where a document has to be indexed. The six operations from thereon consist of:-

(1) concept indexing of the documents,

(2) translating the concept indexing into the descriptor language,

(3) entering coded information into the store,

(4) concept analysis of the question,

(5) translating the concept analysis into descriptor language,

(6) extracting coded information from store.

Later operations cover straightforward clerical matters such as obtaining the original documents, but these are likewise outside the scope of our work. Nor are the methods of assessing relevance of retrieved documents considered here, since it is not essential for this to be carried out within an I.R. system.

The performance of an I.R. system can be considered from a number of aspects. Basically there is the efficiency in regard to recall and relevance. Its performance in these matters will be affected by the indexing, the formulation of the search programme and the specificity of the request. The decision as to the descriptor language and the store will determine the time and cost performance of the system, although this can also be influenced by the efficiency which is demanded.

Having summarised the three groups of problems to be discussed, we will now attempt to show their inter-relationship in the light of the results of the investigation. The most difficult, most onerous task in I.R. is indexing, the plain hard slog of indexing; on how well this is done will depend more than any other factor the efficient performance of an I.R. system. To quote Fairthorne, "Indexing is the basic problem, as well as the costliest bottle-neck of information retrieval".

Indexing, in the sense in which it is conventionally used, is a two-stage process, even though many indexers will merge these two stages into a single process. The first task of indexing is to decide on the subject content of the document which is worthy of inclusion in the index. This we term 'concept indexing' and is in all present conditions an intellectual process. There are attempts to turn it into a pseudo-clerical process, and it might be argued that this has been done with key-word-in-context indexing, and that it is being attempted in a more sophisticated manner by statistical indexing (e.g. Ref. 8). This is to some extent true, but to do this is to make the decision as to what subject concepts should be included in the index according to some pre-arranged plan. A human indexer will also probably try to work to a plan, but is able to alter his requisites in the light of special

circumstances; the penalty for the ability to do this is the probability of human error, as has been demonstrated in the analysis for reasons for failure in the various tests.

However, for the purpose of this discussion, we will assume that the decision concerning the concepts to be included is made by human indexers, and is therefore an intellectual process. The second aspect of indexing is translating the selected concepts into the descriptor language. As has been said, most indexers tend to merge this operation with the previous process of concept indexing, for they will think of their concept in the terms of the descriptor language, but these are two distinct processes, and the second part, the translation into the descriptor language is, or should be, purely a clerical process.

Consider a document which is being indexed. The decision is taken that the document's subject content which is related to the purpose of the index is represented by concepts A, B, C, D and E. This is the basic operation of concept indexing and more than any other matter will determine the resulting efficiency of the system. No document can be retrieved under a given concept unless its concept has been included at this stage. Equally important, if the concept is correctly translated into the descriptor language, it is capable of being retrieved whatever descriptor language is used. To illustrate this latter point, consider four descriptor languages such as were used in the project.

	Descriptor Language			
Concept	I	II	III	IV
A	A_1	A_2	A_3	A_{4a} $_{4b}$
B			B_3	B_4
	$(BC)_1$	$(BC)_2$		
C			C_3	C_4
D	D_1	D_2	D_3	D_{4a} $_{4b}$
E	E_1	E_2	E_3	E_4

TABLE 10.1

TRANSLATION OF CONCEPTS INTO DESCRIPTOR LANGUAGES

Translating the concept indexing into the descriptor language I results in the indexing terms A_1, $(BC)_1$, D_1, and E_1. For descriptor language II, the indexing terms become A_2, $(BC)_2$, D_2 and E_2. Similarly for descriptor languages III and IV, from which it is obvious that the four resulting indexes will contain exactly the same information concerning the document.

This matter obviously raises some further points, and to deal with these it is necessary to consider the nature of descriptor languages in terms of the basic types rather than in terms of examples of types, such as U.D.C. or a particular alphabetical subject heading list. In doing this we would stress the similarity rather than the differences, for it has been the belief that descriptor languages essentially differ from each other which has caused so much confusion.

All practical descriptor languages have common ground in their necessity to have

(a) an alphabetical arrangement,

(b) an arrangement showing relationship between terms.

This is not to say that every example of an actual descriptor language has these two parts, for many simple Uniterm installations or the Keywords index of U.S. Government Technical Reports only have the alphabetical arrangement. It can hardly be denied, however, that they would improve their efficiency if the user had the further assistance of readily being able to locate related terms. Further, assume in an alphabetical subject catalogue, the not uncommon position that a new term is required. Either from memory (if keen enough) or by a visual check, the indexer must go through the other terms already in the system to find whether a new term should be entered as such or whether it should be directly cross-referenced to another term. If "see also" entries are used, then the process must be repeated to find which terms to "refer to" or to "refer from". Failure to do this at the indexing stage would merely transfer the operation to the search stage, where if there were a failure to find required information under a certain term heading, the searcher would have to make his own personal classification or grouping of other terms in the system which might be useful for his purpose.

In practice, certain descriptor languages use the alphabetical arrangement for the entries in the store, while in other cases the alphabetical arrangement is used as an index to some form of classified grouping of entries in the store. To simplify discussion on this point, we would here confine discussion to systems using pre-co-ordination, such as were, in the project, U.D.C., Alphabetical and Facet. Assume that these three systems are represented by descriptor languages I, II,

and III in Table 10.1. It has been shown that into each descriptor language has been fitted the basic concepts A, B, C, D and E. In one case the coding is represented by decimal notation, in the second case by words and in the third case by letter notation, and these will determine the filing orders of the entries in the store. What exactly is the purpose of the different filing orders ? Each system has had exactly the same amount of information put in to it, and therefore each system is potentially capable of retrieving the same amount of information irrespective of the order in which the entries are filed. The only valid reason that can be given for any particular filing order is the convenience of the users of the catalogue. The conventional reason in favour of a classified filing order is that similar subjects are brought in proximity with each other, with the implication that this is not the case with Alphabetical. In favour of an alphabetical filing order is that the searcher can go direct to the store, with the implication that the searcher finds difficulty in using an alphabetical index leading to an unfamiliar notation.

In the project not one of these or of other points was shown to be of any real importance. The technical staff found no difficulty with the complex U.D.C. numbers used in the project, and scored better than with Alphabetical. An analysis of the search programmes showed that U.D.C. searches involved just as many different places in the catalogue to look as with Alphabetical. The analysis showed no significant difference between alphabetical and classified systems for failures which could be in any way put down to the filing order. Convenience for users is, therefore, the predominant consideration.

Whereas we have been considering so far the descriptor languages of a type which are normally used in a pre-co-ordinate manner, the position is, within the context so far considered, exactly the same with post-co-ordinate systems, but before arguing this matter, it is necessary to discuss the final aspect, namely system performance.

The points which have been shown on the curve in Figure 7.6 (page 72), refer to the performance of the particular system that was being tested, the Facet catalogue of metallurgical literature used in the W.R.U. test. The section shown in solid line is that part of the performance curve for which we have definite information, and this shows how one can obtain, for instance, 90% recall. We also know how to obtain 100% recall, for this can always be done in any system and for any question by looking at every document in the collection. Since it is known

in the W.R.U. test collection that there was for each question an average of 3 documents with some degree of relevance, to take this extreme action of looking at every document would produce the low relevance ratio of $\frac{3 \times 100}{1100} = 0.27$. It would probably be possible to improve this relevance ratio slightly by some form of broad grouping without losing the 100% recall, but inevitably there would be some stage where the improving relevance will result in a breakaway from the 100% recall. This argument will not, of course, hold in an environment where the breadth of the question requirement is equal to the breadth of the collection index. If the requests are for all information on metallurgy, and the document collection consists of only documents on metallurgy, the 100% recall will be achieved with 100% relevance.

In considering recall and relevance it is therefore necessary to consider the environment in which the system is operated and here the most important factor is the type of question which will be put to it. Commonly one finds questions being defined as "specific" or "general", but it is difficult to give any agreed assessment to these terms; a question that might be considered specific in a general reference library would be a general question for a library specialising in the particular subject area of the question. Some suggestions have been proposed for defining specificity or generality by relating it to the number of concepts contained in the question, and this method has been considered. While it may be reasonable to do this in comparing different questions in relation to each other in a given situation, it appears to be of doubtful use when attempting to obtain more precise values. It is therefore proposed that this matter should be controlled by giving question specificity or generality a definite measure which will be obtained by relating the size of the total collection to the number of relevant documents which the collection contains. This would assume that the subject content of the collection was restricted to a single discipline, and the following table is suggested as giving this measure based on a collection of 100,000 documents:-

Number of Relevant Documents	Question Generality Figure
0 - 4	1
5 - 10	2
11 - 20	3
21 - 40	4
41 - 100	5
101 - 500	6
501 - 2000	7
2001 - 10000	8
10001 - 20000	9
More than 20000	10

TABLE 10.2

QUESTION GENERALITY FIGURES

Indications are that the general level of operation is covered by Generality Figures 5 to 6, although individual questions would cover a far wider range.

Since, to return to the previous argument, it is known that without using any system, one can always obtain 100% recall, it therefore appears logical to argue that the indexing is not carried out for the purpose of finding relevant documents so much as the purpose of not having to look at large numbers of irrelevant documents. In stating, therefore, the performance requirements of an information retrieval system, one can either ask that in a given environment, it must give a certain recall ratio with the best possible relevance figure, or say that it must have a given relevance figure with the best possible recall ratio. Add to that the requirement that the cost must not be above a certain figure and one has a complete operational specification.

We can now start working back to the indexing to find how this can be effected. First to the discussion broken off concerning the essential similarity between pre-co-ordinate and post-co-ordinate systems. If the fourth descriptor language in Table 10.1 is taken to represent a post-co-ordinate system such as Uniterm, it is definite that it contains no more information concerning the basic document than do the other three pre-co-ordinate systems. The apparent difference between the two types of systems is that the post-co-ordinate system has the ability to retrieve any combination of the concepts, whereas the pre-co-ordinate system can only do this by having the

requisite entries in the catalogue. It has frequently been implied that this is an outstanding advantage which makes a post-co-ordinate system far more effective in retrieval, but the results of the investigation show that this advantage, though it existed, was not large . In the Facet catalogue of the W.R.U. test, when eight entries were made for each document, it had virtually disappeared. The difference between the two types of system is therefore shown to be not a fundamental difference but merely one of cost or convenience, and it has not been proven as yet on which side the advantage lies.

This brings us back to the indexing decisions, and here the obvious requirement is for more work to be done to help the indexers. For discussion on this point, the following definitions are proposed.

EXHAUSTIVE INDEXING

The indexing of every possible item in a document implies that several different, but quite independent, units of information are given in the document. The criterion is whether such a unit of information is a useful piece of information in its own right and is a truth even when taken out of context. It differs from the bits of subject concepts which go to make up a compound subject and which are represented by the elements in a notation or a compound subject heading or by the descriptors in a post-co-ordinate system.

Definition

The provision in the retrieval system of an entry for each individual unit of information which is capable of standing alone and which comprises a useful and valid piece of knowledge whether considered in the context of the document in which it is contained or in isolation. (Such entries need not necessarily be 'specific' entries).

SPECIFIC INDEXING

An entry may be made under a heading which can be very broad and which includes the subject being indexed, or very narrow and covering only part of the subject, or can reach a degree of 'specificity' such that the heading represents the total subject content of the topic being indexed and no more. (i.e. the parts of the heading together are necessary and sufficient). In other words the ultimate in 'specificity' is the situation wherein the heading is co-extensive with the subject it represents.

Definition

The provision of a heading in a pre-co-ordinate system, which by definition or usage, is co-extensive with the subject of the unit of information being indexed, whether the unit of information is part of a document or comprises the entire document. In post-co-ordinate indexing, specific indexing consists of the posting of the unit of information to all the terms which taken together form a description of the unit which is co-extensive with its subject content. 'Specific' used in this sense is absolute, but can also be used relatively to indicate different levels of specificity in indexing.

SYNTACTIC INDEXING

Most alphabetical subject headings which consist of several elements are syntactic in that they display adjectival and adverbal relationships, and show the objects of processes, the subjects of problems, etc., but this type of heading in the present state of the art usually shows a low level of specificity. Such headings as coloned U.D.C. numbers are not syntactic in that there can be ambiguity regarding the relationship of one element to another. The use of 'role indicators' is syntactic in principle but in the case of post-co-ordinate indexing 'syntactic headings' as such do not exist until they are formulated as a search programme.

Definition

The use of headings which display the relationship between the various elements, as distinct from those which merely show the existence of several attributes relevant to the subject indexed.

WEIGHTED INDEXING

Conventional indexing involves a 'Yes' or 'No' assessment; a concept either is or is not considered worthy of inclusion in the index. It is possible for the indexer to indicate the relative importance of different concepts in a document.

Definition

The provision with the indexing term of an indication of the relative importance of the term in the context of the document indexed.

Comparatively little is known of the cost and the effect on efficiency of exhaustive indexing, apart from the certainty that it will result in higher recall but lower relevance. Specific indexing on the other hand, will result in lower recall but higher relevance. This will always be true whatever descriptor languages are

used, if, as was implied with the systems set out in Table 10, they are operated at the same level.

Considerable emphasis is being placed on techniques designed to show the association between terms, but none of these techniques appear to have any chance of by-passing this fundamental problem of recall and relevance. As Fairthorne puts it one can have "all but not only" or "only but not all". It is important to realise that it is the indexing and the descriptor language which determines the highest relevance ratio that a system can attain, whereas recall can always be improved by adjusting the search programme, even, if necessary, up to the absurd limit where the whole collection is retrieved. If a classification scheme such as U.D.C. is never used beyond the first four figures, it will be useless to attempt in retrieval to distinguish between, say, 'altitude performance of piston engined cargo aircraft' (629.138.4:621.431.75:533.6.015.5) and 'range performance of jet-propelled private aircraft' (629.138.56:621.45:533.6.015.74) for both would be shown as 629.1:621.4:533.6.

The same is true in a post-co-ordinate system if the descriptor language only includes 300 descriptors as against 4,000 or so uniterms. Such indexing will give good recall, but the best possible relevance will be lower than it would be with the more specific indexing which can be done with the full U.D.C. schedules or by using precise uniterms. Yet with the greater specificity and therefore better relevance, it is always possible to broaden a search so as to achieve higher recall, the penalty being that search procedures are more complex than when the indexing is less specific.

It now becomes necessary to amplify the argument based on Table 10.1. It remains true that given the same concept indexing, any two descriptor languages will have the same information content, and therefore the same potentiality for retrieval. This might be modified by stating that their performance will be similar if they are operated at the same level of indexing specificity. If this level varies, then for a single level of search programme, it might be that the performance of four systems could be shown as in Table 10.3.

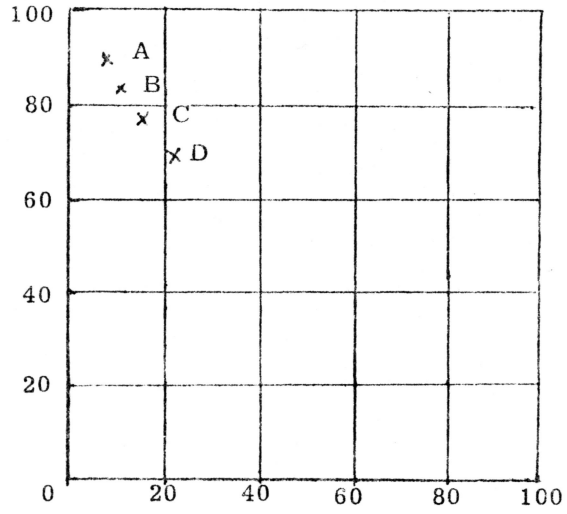

TABLE 10.3

HYPOTHETICAL PERFORMANCE OF FOUR DESCRIPTOR LANGUAGES OPERATING AT DIFFERENT LEVELS

It is suggested, however, that system D, which can be seen to be utilising more specific indexing, could, with broader levels of searching, achieve the same performance figure as system A. On the other hand, it would not be possible for system A to match the performance figure of D, since it lacks the power to improve the relevancy figure. All this leads to the conclusion that it is not the alternatives of classified or alphabetical arrangement, of post-co-ordinate or pre-co-ordinate indexing (much less the alternatives of manual or mechanical searching) which *make any real difference in performance but the power of the descriptor language*, allied to the standard of the indexing. The 'power' of a descriptor language is in its ability to eliminate irrelevant references, and in addition to a hospitality for specific indexing, there are at least two other devices which can be used, namely, "syntactic indexing" and "weighted indexing". As far as is known, no valid evidence exists as to what improvement these techniques can bring about, but logically it would appear that they must assist in eliminating irrelevant references. Any such improvement would probably have to be paid for by higher input costs.

To summarise the argument, it is maintained that the most important factor in the efficiency of an I. R. system is the concept indexing. More work is required to assist the indexer in this stage of the work, particularly in relation to the influence of exhaustive indexing on performance. There is the possibility that computers will be able to do concept indexing; if the infallible machine can be given a programme which will allow it to operate at 90% of perfect indexing, it will be near the level reached by the fallible human indexer. The translation of the concept indexing (whether done by humans or machines) into the descriptor language is a clerical process which can, in itself, not affect the efficiency of the system. The main difference of importance between types of descriptor languages is in their power, which includes their receptivity for specific indexing and other techniques such as syntactic indexing and weighted indexing. These improve the ability of the system to eliminate irrelevant references and obtain a good relevance ratio. The relevance ratio in combination with the recall ratio is the measure of performance of an I. R. system, and can be expressed in a series of curves indicating the performance according to environment, indexing level and preciseness of search.

In an earlier paper we quoted a remark made by Mr. M.J. Lighthill, F.R.S., Director of the Royal Aircraft Establishment, Farnborough. It was taken from a paper in which he was discussing the advances made in aeronautical engineering over the last fifty years. Mr. Lighthill wrote, "Countless ingenious experiments on models lay at the back of every advance and brilliant theories have been devised to make sense of the experiments". The present investigation has been one such experiment, (but we would not claim that it was particularly "ingenious") and we hope that others will find in this report the data to help devise theories that make sense of the experiment. More than anything this investigation has highlighted the major basic problem in information retrieval. As they appear to us, we have attempted to set them out in this chapter; the continuation of the project will attempt to find some of the solutions.

REFERENCES

1. Cleverdon, C.W. — Final Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, September 1960.

2. Gull, D. — Seven years of work on the organisation of materials in the special library. American Documentation, Vol.7, 1956, pp 320-329.

3. Cleverdon, C.W. — Aslib Cranfield Research Project on the comparative efficiency of indexing systems. Aslib Proceedings, Vol.12, 1960, pp 421-431.

4. Hyslop, M. — Machine literature searching - from experiment to experience. American Documentation, Vol.12, 1961, pp 49-52.

5. Kent, A. — Exploitation of recorded information. Development of an operational machine searching service for the literature of metallurgy and allied subjects. American Documentation, Vol.11, 1960, pp 173-186.

6. Melton, J., Buscher, W. — The Cleverdon-Western Reserve University Experiment - Search Strategies. Proceedings of the Conference on Information Retrieval in Action. Cleveland, 1962.

7. Rees, A. — The Cleverdon-Western Reserve University Experiment - Search Results. Proceedings of the Conference on Information Retrieval in Action. Cleveland, 1962.

8. Maron, M.E., Kuhns, J.L. — On relevance, probabilistic indexing and information retrieval. Journal of the Association for Computing Machinery, Vol.7, 1960, pp 216-244.