CHAPTER 8

Comments

This report has attempted to outline the reasoning and the procedures adopted in the second Cranfield project. It could be argued that comments on these matters should await the publication of the test results, but it is felt more appropriate to conclude this volume by briefly considering some of the short-comings of the design and the techniques used, and showing how the results might possibly be affected.

In Chapter 2, some aspects of the test design were considered from the viewpoint of the decisions which seemed correct in 1961, at the time when the project was prepared. While the test results and the conclusions which can be derived from them will show to what extent the test design is such as to allow the objectives to be achieved, there are certain matters which can be discussed immediately.

The original proposal suggested a collection of 1,200 documents with some 300 questions to be used for searching. For no very good reason, the total of documents in the collection was increased to 1,400; while there would have been no difficulty in finding 300 usable search questions from the 641 that were submitted, only 279 questions were used, and, for most of the tests, this number effectively was reduced to 221. The amount of data which has been obtained from this question-document set is vast, and is more than sufficient for validation of the test results. It can at present only be a matter for discussion as to whether the question-document set was larger than necessary. In many of the tests, sub-sets of the collection were used, sub-sets such as 200 documents and 42 questions. There is a double danger in the use of such comparatively small sets; firstly that they will produce results which are unrepresentative, and secondly that the performance measures will be seriously distorted. To consider the latter point, investigating the effect of generality ratio was a part of the project and although the matter is somewhat complex, it has been possible to work out the relationship between the performance figures for varying generality ratios. This work is reaching the stage where it can be applied in all situations, so this particular problem need no longer create any difficulty in the use of a small collection.

Far more serious is the question of whether the collection size is large enough to give valid results. It has to be remembered that this investigation has been concerned with only one variable, namely index language devices, and this is quite unlike the situation in Cranfield I, in which additional variables were such matters as indexing time, indexers, and type of document. The result is that a much smaller set than the 18,000 documents and 1,200 questions of Cranfield I was required and there does not seem to be any doubt but that the collection of 1,400 documents was large enough for the test. The experience at Cranfield and Harvard of working with a sub-set of 200 documents and 42 questions has produced some useful evidence on the question as to whether the total collection was larger than necessary. With the knowledge that the sub-set produced results very similar to those obtained with the complete collection (when due allowance is made for the generality ratio), it now seems possible that a smaller collection would have served equally well. However, lacking this hind-sight knowledge, it is very likely that the results obtained with a smaller collection would have been subject to criticism which could not have been satisfactorily refuted.

The method of obtaining a document collection and a set of questions turned out to be a perfectly satisfactory way of operating. The response from the authors of research papers was remarkably good, and can be interpreted as showing that the scientific community - at any rate in the field of aeronautical engineering - is interested in documentation problems, and is willing to co-operate in helping to find an answer to these problems. The selection of the comments from the authors (given in Chapter 3) is only a sample, illustrating various points, of the many interesting and encouraging letters which were received.

Tied in to this method of obtaining the set of documents and questions was also the matter of obtaining relevance assessments, and here some reservations have to be admitted concerning the method adopted. This is not to suggest that there is any experimental evidence of there being any better or more satisfactory technique, but rather to say that the matter of relevance assessment is, without any doubt, the most difficult intellectual problem - in fact, one of the very few remaining problems - in the evaluation of information retrieval systems.

In the evaluation of operational systems, there will be many occasions when the only satisfactory technique will be that of using actual questions for test searches, with the questioners assessing the relevance of the documents retrieved at the time when the information is required. Such would be the case if it was desired, for instance, to investigate the effect of different levels of questioner participation in the search programme. As soon as any deviation is made from this technique of operating in a real-life and real-time situation, a less realistic method is being used, although there will frequently be situations where this could be justified for economic or other practical reasons. This latter point is certainly true of an evaluation of an operational system, and it is equally true of the test of an experimental system, where no real user group can be said to exist. A possible weakness of the method adopted in this test lies in the fact that the subjectivity of the relevance assessments might have been such that it will mask the variation in performance of the various devices which were being tested. There is no experimental evidence of any kind at present available that makes it possible to affirm that this is so, but the possibility is such that it requires investigation.

As stated earlier, the problem of relevance decisions is presently the most serious in the field of evaluation, and is attracting the attention of many groups. There is the very interesting work of Katter (ref. 33) in which a large number of people will be asked to make 'distance' judgements between small sets of documents. In this work the important aspect of the test design is to find which type of document surrogates result in distance judgements which match most closely those judgements made by assessing the complete documents. Then there is the work of Cuadra (ref. 34) where up to one hundred individuals will be asked to assess a set of documents in the field of information disemination, storage and retrieval. Here the attempt will be to identify and investigate the variables which influence an individual's response, and a somewhat similar investigation is being directed by Rees at Weston Reserve University (ref. 35). More empirical is an investigation proposed by Cleverdon which is to be undertaken by ASLIB. This is intended to identify the reasons why individuals reject documents which apparently meet their requirements and alternatively why they accept, as relevant, documents, which to a third person seem no more acceptable than those rejected. This investigation will be carried out on some 600 individuals in twelve different organisations and, unlike the other three projects, the relevance assessments will be made in actual operational conditions.

However, none of these investigations into relevance apply to the problem raised in this test. Here the situation is that a series of tests on various index languages have been carried out, where the scoring for each test is based on the relevance decisions of individuals simulating, as far as possible, a real life situation, with individual variations hopefully evened out by having nearly two hundred different questioners In an investigation that had very similar objectives, Salton went to the other extreme in his original tests. Using only seventeen questions in the general field of the test collection, these questions were specially prepared for the test and did not represent any actual requirements. The set of 400 documents in the collection were then assessed against these prepared questions by a number of students, this assessment being based only on short abstracts. Since the searching was also done on the abstracts, there was obviously the probability of even more distortion than was the case with the source document questions of Cranfield I. The interesting point, however, is that this seemingly crude technique of question preparation and relevance assessment did, in fact, allow a considerable amount of useful data to be obtained concerning the performance characteristics of the various index language options, and this data appears to be sufficiently valid for certain conclusions to be reached. When this evidence is added to that obtained from Cranfield I, there are some grounds for suggesting the possibility that everyone is over-emphasising the importance of relevance assessments in experimental testing, and that, however relatively unscientific the method used, reliable information can be obtained. It is intended to investigate this point in future work at Cranfield by having various people make new relevance assessments of the document-question sets used in the present project. The search results can then be re-scored on the basis of these new - and presumably somewhat different - relevance assessments, and analysis will show whether the comparative performance of various index languages is thereby affected.

In experimental testing, the common practice, not unnaturally, is for the groups to work with document collections with which they have some familiarity, and this project was no exception. The language of aerodynamics might be said to fall somewhat to the left of centre in regard to its precision, it is, in fact, mushy rather than firm. As such it presented a number of difficulties; not only could one find the same notion being expressed in different ways by different authors, one often had the situation where the same notion was expressed in different ways in the same paper. Discussing this point with one of the authors, he said that certain people considered it good style if, after expressing a notion in the title in one way, a new phrase could be used for the abstract and another phrase be found for the actual text. Even without this particular complication, the subject matter was full of semantic problems. An illustration of this is provided by a question (not in any way a-typical) which, as originally received, read

'Has anyone investigated relaxation effects on gaseous heat transfer to a suddenly heated wall'

When asked to suggest alternative search terms, the questioner sent back the following comments.

<u>Relaxation effects.</u> Could be replaced by 'excitation of internal molecular energy modes (or states)'. The excitation could result from collisions between gas molecules alone, or gas molecules and molecules in the solid.
<u>Gaseous heat transfer.</u> 'Gaseous' could be omitted, but does help to limit the field. 'Heat' could be replaced by 'energy', 'transfer' by 'conduction' or 'transmission'.
<u>Suddenly heat ed wall.</u> 'Suddenly' could be replaced by 'rapidly', 'heated' by 'cooled' and 'wall' by 'solid'

Finally, if any of the above permutations are unsuccessful, the question could be rephrased to read 'Has anyone investigated the conditions at the wall behind a plane reflected shock front in a real gas by theoretical analysis'.

The semantic difficulties of papers in aerodynamics provided a very stiff test of the recall devices, and as such it could be considered a suitable subject area for the test. However, the lack of syntactic difficulties caused a change of plan, as considered on pages 56 and 57, in that it was not a practical proposition to use roles. It is an interesting point to consider as to whether another inverse relationship exists, this time between the semantic and syntactic problems involved in the indexing of any particular subject field. Alternatively, and possibly more likely, the position may be that with a mushy subject language, the over-riding necessity of obtaining a reasonable recall ratio inhibits the use of precision devices; in other words, the semantic problems are so difficult to solve that they completely overshadow the syntactic problems. However, in a firm subject language area the semantic problems are more easily solved, so the syntactic problems loom larger, and one can afford to use precision devices, such as roles. If either of these situations exists, it will obviously have consequences in the endeavours to obtain a common sample of documents that can be used to illustrate and evaluate different types of systems, such as the work at Chicago. Here the intention is to have 'an open-ended collection of exemplars of indexing systems applied to a common sample of documents' (ref. 36). The indications are that any given sample of documents would favour certain types of index languages, but handicap other index languages, this being dependent on their strong and weak points in relation to devices intended to overcome the semantic or syntactic problems.

It would seem, that next to the question of relevance assessments, the determination of the effect of subject language precision is the most important problem to be tackled. This is certainly true of experimental situations where it is necessary to compare the performance of tests based on different document collections. For instance, in the comparison of the results obtained by the SMART tests and in Cranfield II, it is now possible (by the methods discussed in a later volume of this report) to normalize the different measures used and the effect of generality ratio. Since it is also theoretically possible to match similar types of index languages and the method of relevance assessment, any remaining difference in performance figures must be due to the firmness level of the language of the two subject areas, namely computers and aerodynamics.

In addition to experimental situations, knowledge of this factor is also important for the design of an operational system which covers a broad subject field, and where there is thereby a wide range in the firmness level. An investigation of this problem could be attempted by a linguistic analysis of the variation of terms in different subject fields - how many different terms or phrases can be used to express the same notion and conversely how many meanings a single term has. The experimental method of investigating the problem to be used at Cranfield will be a procedure that reverses the present project. Instead of testing a large number of index languages against a single document set, it will be necessary to find the different performances achieved when a large number of document sets in different subject fields are tested against a single index language.

No particular fault is at present apparent in regard to the indexing which proceeded according to schedule and was completed during the first year of the project. In the next stage of the work, there was probably an error of judgement in putting so much effort into the preparation of the single-term hierarchies. It is doubtful if anyone has previously attempted to compile classification schedules consisting entirely of single words, and it proved to be a very difficult, but therefore very interesting task. It was right that, in this project, the attempt should have been made, but an earlier realisation of the limited affect which the schedules would have on the performance of the system would have led to a decision that less time should be expended in their preparation.

The main objective of the test is to ascertain the effort of various index language devices on the performance of information retrieval systems. To conclude this volume, it is reasonable to claim that, although there are some operations which might have been done better another way, nothing has happened seriously to militate against the possibility of achieving this test objective.