CHAPTER 6

SUPPLEMENTARY TEST PROGRAMMES

The whole emphasis of the basic test programme was concerned with recall, based on the retrieval of source documents. Relevance of the documents retrieved had in other investigations proved to be too difficult a problem to be satisfactorily solved, for relevance is, in the present state of the art, a purely subjective assessment. It will vary with the interests of the different individuals who make the assessment and it can also vary for the same individual at different times. In fact, the only person who can truly assess the relevance of a document to a question is the person who asks the question, and then only at the time when he actively requires the information. Even under these conditions, the high or low relevance of a particular document can obviously be influenced by any other documents retrieved. It was to side-track temporarily these difficulties that the particular test technique of the basic programme was used, and the intention was to make tests of a different nature which would enable more information to be obtained on other matters such as relevance.

A first task was to find exactly what was being measured, exactly what was implied when it was said that Uniterm, for instance, had an efficiency of 85%. At the time of the first public discussion of the preliminary results (Ref. 3), we affirmed that this did, in fact, mean that the searches were retrieving 85% of all the documents in the system which had a degree of relevance which was higher than or at least equal to the document on which the question was based. To this statement there was one qualification, namely that the figure of 85% might be too high due to the fact that there was an unnatural correlation between the document and the question as compared to a true life situation. If it were shown that this correlation resulted in the efficiency being, say, 15% higher than it would otherwise be, then it might be expected that the figure of 70% would represent the real efficiency of recall.

The supporting argument for this view was as follows. It is known that amongst the total collection of documents there is a group of 100 documents which will provide relevant answers to 100 questions, namely the documents which were used by the compilers of the questions. This was the situation in the main test programme, and the result of searches for 100 questions was that with Uniterm

for instance, on an average 85 of this known group of relevant documents was recovered. Assume that there had been another group of 100 equally relevant documents for the same collection of questions; it is not unreasonable to presu me that 85% also of these would have been retrieved. Assume a further 100 relevant documents, with the same search result. Continue this to a stage where there is a collection of 10,000 documents, with 100 documents being relevant to each question. The result of a single search would then be expected to be that 85 of the known 100 documents were retrieved.

The above argument has not been shown to be false or illogical, but further tests were required to show how valid it could be considered. The method of attempting to do this was to select 100 questions. The selection was made in such a way as to ensure that half the questions covered aerodynamic subjects, while the remainder dealt with the other more general subjects in the collection. These were sent in groups to the librárian or information staff at different organisations working in the appropriate subject fields. They were requested (see Appendix 6A) to prepare as complete as possible bibliographies for each question. In particular it was emphasised that a bibliography was not intended to be critical, but that it should include any reference which it appeared might be relevant to the question. We received bibliographies covering 88 questions, and 18 questions had duplicate lists from different organisations. On being received at Cranfield, each bibliography was checked to ascertain which items were included in the documents covered by the index. Eight bibliographies did not include any references to such documents, so there were left bibliographies to 80 search questions. The total number of references in these bibliographies to papers which had been indexed in the project came to 359, varying from one to a maximum of fifteen. The source documents were not to be included in this test, so whenever they appeared in a bibliography, they were crossed out. Each of the 359 documents were then assessed in relation to the appropriate question, with the source document being used as a guide to determine relevance. The assessment rating was '1' for documents as useful as the source document, '2' for documents of some interest and '3' for documents of no interest.

As a result of this assessment, there were 53 documents which had a rating of '1', and 67 documents with a rating of '2', the remainder not being considered of any interest. The number of questions which were covered by documents of top relevance was 35, and 6 other questions had documents of lower relevance.

The breakdown of these questions and documents is shown in the first two columns of Table 6.1. Searches were then made for these 41 questions, but in this case, contrary to the main test programme, the searcher did not know the document numbers of the relevant documents and the search continued until the searcher had covered all reasonable possible programmes. The results of these searches are given in the final column of Table 6.1 and Table 6.2 gives the total figures.

	No. of relevant documents		Number of relevant documents retrieved			
			UDC	ALPHA	FACET	UNITERM
	1	2	1 - 2	1 - 2	1 - 2	1 - 2
PQ1	1	2	1 - 1	1 - 2	1 - 2	1 - 2
PQ3	1	1	1 - 1	1 - 1	1 - 1	1 - 1
PQ6	1	1	0 - 1	1 - 1	1 - 1	1 - 1
PQ9	2	2	1 - 2	1 - 2	0 - 1	1 - 2
PQ10	1	2	1 - 1	1 - 1	1 - 1	1 - 1
PQ13	0	1	0 - 1	0 - 1	0 - 1	0 - 1
PQ14	2	1	2 - 1	2 - 0	1 - 0	2 - 0
PQ15	1	0	0 - 0	0 - 0	0 - 0	0 - 0
PQ18	3	3	3 - 2	2 - 2	2 - 2	3 - 2
PQ21	1	1	1 - 1	1 - 1	1 - 1	1 - 1
PQ26	1	2	1 - 1	0 - 1	0 - 1	1 - 1
PQ27	1	1	0 - 1	0 - 1	0 - 1	0 - 1
PQ28	1	0	1 - 0	1 - 0	1 - 0	1 - 0
PQ33	0	1	0 - 1	0 - 1	0 - 1	0 - 1
PQ35	3	5	2 - 3	2 - 4	2 - 3	3 - 4
PQ39	1	1	1 - 1	1 - 1	1 - 1	1 - 1
PQ41	1	3	1 - 2	1 - 2	1 - 2	1 - 3
PQ45	1	2	0 - 2	0 - 1	0 - 1	0 - 2
PQ48	1	2	1 - 1	1 - 2	1 - 1	1 - 1
PQ50	2	2	2 - 1	2 - 1	1 - 1	2 - 1
PQ53	1	1	1 - 1	1 - 1	1 - 1	1 - 1
PQ56	1	1	0 - 1	1 - 0	1 - 0	0 - 1
PQ59	0	2	0 - 2	0 - 2	0 - 1	0 - 2
PQ61	6	5	4 - 4	4 - 3	3 - 3	4 - 3
PQ64	1	1	1 - 1	1 - 0	1 - 0	1 - 1
PQ65	1	2	0 - 1	0 - 1	0 - 1	0 - 1
PQ68	2	3	1 - 2	1 - 2	1 - 1	2 - 1
PQ70	1	1	1 - 0	1 - 0	1 - 0	1 - 0
PQ72	1	0	0 - 0	0 - 0	0 - 0	0 - 0
PQ79	4	3	4 - 2	4 - 2	3 - 2	3 - 2
PQ80	2	3	2 - 1	2 - 2	1 - 2	2 - 2
PQ82	1	0	1 - 0	0 - 0	0 - 0	0 - 0
PQ84	0	1	0 - 1	0 - 1	0 - 1	0 - 1
PQ87	0	1	0 - 1	0 - 1	0 - 1	0 - 1
PQ88	2	3	1 - 3	1 - 2	1 - 2	1 - 3
PQ90	1	1	1 - 0	1 - 0	1 - 0	1 - 0
PQ91	1	1	1 - 1	1 - 0	1 - 0	1 - 1
PQ92	1	2	1 - 2	1 - 2	1 - 1	1 - 2
PQ94	1	1	0 - 1	0 - 1	0 - 1	0 - 1
PQ96	1	1	1 - 1	1 - 1	1 - 1	1 - 1
PQ99	0	1	0 - 0	1 - 1	0 - 1	0 - 1
			39 - 49	40 - 47	32 - 41	40 - 51
			74% - 73%	75% - 70%	60% - 61%	75% - 76%

TABLE 6.1

RESULTS OF SEARCHES FOR RELEVANT DOCUMENTS FROM BIBLIOGRAPHIES

	RELEVANCE	
	1	2
U.D.C.	39 (74%)	49 (73%)
ALPHA	40 (75%)	47 (70%)
FACET	32 (60%)	41 (61%)
UNITERM	40 (75%)	51 (76%)

TABLE 6.2

SUMMARY OF RESULTS OF SEARCHES FOR RELEVANT DOCUMENTS FROM BIBLIOGRAPHIES

While it is the case that, because of the relatively small number of documents concerned with this test, the standard error is high, it would seem probable that there has been a slight but definite reduction in the efficiency as compared with the main test. Possibly significant is that Uniterm shows the largest drop for the figures in the basic test.

This test involved a great deal of effort, not so much for the project staff as for those who voluntarily co-operated by compiling the bibliographies and it is doubtful if the more valid information that could be obtained by a larger programme would be commensurate with the effort involved. It was a disappointment to find so few relevant documents amongst those which had been indexed in the project. Unfortunately this test was under way before the relevance assessments (discussed later in this chapter) had been carried out; had the latter been completed first, it would have been no surprise to find that the large majority of documents listed in the bibliographies were of no interest and it would certainly be unreasonable to criticise those who gave generously of their time.

If it is agreed that 85% efficiency in the main test is equivalent to retrieving 85% of the relevant documents, it is still necessary to make some qualifying statement concerning the operating conditions. The 85% efficiency of Uniterm was achieved using the search rules as considered in Chapter 2. These permitted the searcher to drop one, but not more than one, of the basic concepts originally considered necessary. If the search programme requiring ABCD proved unsuccessful it was permissible to search ABC, ABD, ACD or BCD but not permissible to search AB or any other two-concept term, much less A or B or C or D on their own. As stated earlier, with KWIC indexing and also with Uniterm, if a single concept had been accepted, the efficiency would be 97%. Obviously, if it had not been possible to accept something less precise than the original require-

ment, it is certain that the recall efficiency would have been lower. The result is that there is the possibility of quoting three different performance figures, those with Uniterm as an example being:

65% when all concepts are required

85% when one less concept than the required is accepted

97% when a single Uniterm is accepted.

The only practical method of showing these varying points is by plotting them against relevance ratio, that is the percentage of the retrieved documents which have an agreed relevance. This matter is considered in more detail in Chapters 9 and 10, and mention is only made here in connection with the analysis that was made. Sufficient now to make the point that as the recall figure (i.e. the percentage of potentially relevant documents in the collection) rises, the relevance ratio (i.e. the percentage of relevant documents amongst the total of those retrieved) must fall and conversely as the recall figure drops, so the relevance ratio will improve.

In order to find the relevance ratio, an assessment was made of the number of documents which had been retrieved in the course of the searches. For this purpose a random sample was taken of 79 questions spread over all indexing variables. From the master search cards there was obtained a list of all the references found in the course of the searches. This showed that the total number of documents obtained in the searches was as in Table 6.3.

	Total	Average per search
U.D.C.	3171	40
ALPHABETICAL	2122	27
FACET	1910	24
UNITERM	1527	19

TABLE 6.3

TOTAL DOCUMENTS RETRIEVED IN SAMPLE OF 79 SEARCHES

These numbers may seem large, but in every case were swollen by very heavy retrieval in certain searches and if the twelve searches with the highest retrieval figures were deducted in each case, the figures for the remaining 67 searches would read as in Table 6.4.

	Total	Average
U.D.C.	1346	20
ALPHABETICAL	940	14
FACET	1060	16
UNITERM	895	13

TABLE 6.4

DOCUMENTS RETRIEVED IN SAMPLE OF 67 SEARCHES

A sample was taken of the documents retrieved by each system. The intention was that an assessment analysis of approximately 400 documents should be made for each system. Care had to be taken not to bias the sample by using documents which had been retrieved by all systems, as this would have tended to include those which had a strong probability of being relevant. Naturally, however, there was some duplication and the sample finally involved a total of 759 documents, which were assessed in the same manner as described for the previous test. From this random sample, it was found that a total of 59 documents with the top relevance rating had been retrieved, in addition of course, to the source documents. These were in many cases retrieved by more than one system, but they turned up in the analysis of the different systems as set our in the first column of Table 6.5.

	a	b	b + source	Relevance ratio
U.D.C.	19	150	229	7%
ALPHA.	32	165	244	12.5%
FACET	16	85	164	7.5%
UNITERM	28	101	180	12%

TABLE 6.5

RELEVANCE RATIO OF DOCUMENTS

RETRIEVED IN 79 SEARCHES

a. Relevant documents in sample assessed
b. Assumed relevant documents in total retrieved

To find the total number of relevant documents retrieved, the total in this sample was multiplied by the appropriate factor depending on the total number of documents retrieved by each system (as shown in Table 6.3). This figure is shown in the second column, while the third column gives the total when the number of source documents have been added to the figures in column 2. Finally the relevance is obtained by finding the percentage of relevant documents against the total number of documents retrieved, (as given in Table 6.3).

As will be seen, the results showed a suspiciously large variation regarding relevance ratio, so a further check was made. This consisted of finding exactly how many of the 59 top-relevance documents had in fact been retrieved by each system as against those which had happened to be included in the assessed sample. The result of this check is given in the second column of Table 6.6 and shows a major change, for Alphabetical, which had originally disclosed the most non-source relevant documents, now dropped to the bottom, while the other three systems were all very level. The third column shows the percentage retrieval from the known collection of 59 non-source relevant documents, and this should be compared with the final column of Table 6.2. It does not, however, tell the full story, for the non-source relevant documents included in this analysis were only those retrieved by the successful programme for the particular system. To explain this point, it will be recollected that in the main test, the search was only carried to the stage where the source document was located. This might, in some cases, have involved many different programmes with one system, but only a single programme with another system. Normally the more searches, the more documents retrieved, and this would usually result in more relevant documents being retrieved, with the penalty that more irrelevant documents will also be brought out. Sinc e the non-source relevant documents were only those that had been retrieved in the course of possibly limited searches, a check on this was made by going back to the master search card and indexing card for all the failures to retrieve these non-source relevant documents, and attempting to assess whether an extended and complete search programme would have retrieved them. The result of this analysis was that the figures in the second and third columns should be improved to those given in columns 4 and 5 of Table 6.6, and represent a considerable increase on the figures for Table 6.2 and even an increase on the searches for source documents in the main test programme.

	a	b	c	d	e
U.D.C.	19	43	71%	46	78%
ALPHABETICAL	32	36	61%	48	81%
FACET	16	41	68%	50	85%
UNITERM	28	42	70%	53	90%

TABLE 6.6

NON-SOURCE RELEVANT DOCUMENTS RETRIEVED IN 79 SEARCHES

a. Relevant documents in sample assessed
b. Known relevant documents retrieved
c. Percentage of known relevant documents retrieved
d. Known relevant documents which could have been retrieved by improved searches
e. Percentage of known relevant documents which could have been retrieved by improved searches

This somewhat tortuous analysis serves to emphasise nothing more than the extreme danger of placing too much credence on any of the figures which are not otherwise corroborated. To recap, there is the known figure of recall as given in the main test. Of this figure the claim has been made, earlier in this chapter, that it represents not only the recall figure for source documents, but also the recall figure for all relevant documents in the collection. The two supplementary tests have shown that this statement is probably true with one proviso. It appears doubtful if an equal percentage of the total of non-source relevant documents were actually retrieved, but it does appear from the final column of Table 6.6 that an equal figure could have been obtained with extra reasonable search programmes. On the other hand, the final column of Table 6.2 indicates a general lowering of the main test figures for recall and this could be taken to indicate that the questions were slanted towards the source documents.

In point of fact, the documents being assessed in these two further tests numbered only 53 and 59 respectively, and this is too small a figure to have any real validity. To have increased this figure materially would have involved a large amount of extra work and would still only produce figures whose basis would be the unprovable assumption that the relevance assessment was correct. It was quite impossible to go back to the originators of the questions for them to determine the relevance, and even if this had been possible, it would, in view of the lapse of time, have been impractical to do so.

The present investigation was not geared to carry out this refinement of operation, and it was decided that it was unwise to spend much further time in trying to make fine measurements with a crude instrument. However, the ratio of relevant documents to irrelevant documents in the operation of this test can be said with some certainty to lie between 6% and 14%. Whereas it is, at this level, a very wide range, it does indicate, possibly for the first time, the region in which information retrieval systems are conventionally working. Crude as these further tests were, they did give pointers to the more valuable analysis described in Chapter 7 and the further programme considered in Chapter 10.

A further test was made in an attempt to find what improvement could be obtained by combining the specialised knowledge of the technical staff and the project staff in searching. For this purpose, in each system 60 searches which had been unsuccessful both by technical and project staff were repeated. These failures were all in the first two rounds of testing, when the search programmes had not been co-ordinated between the four systems, and the test was done prior to the analysis of failures described in Chapter 5. The result of this collaboration between technical and project staff was that a further five source documents were located in U.D.C. and Facet, six in Alphabetical and four in Uniterm.

Another test, which falls into a different category, was made with Facet. The results for Facet had been disappointing, for they were markedly lower than the other systems. Personal observations of the searchers, reinforced by the analysis of failures, was that the main weakness of Facet lay in the fixed order and chain index. This difficulty was forecast in the section written by Mr. J. Sharp in Chapter 4 of Ref. 1, and it was contemplated that it might be worth testing Facet by using it in a post-co-ordinate manner. It was, however, first decided to try the effect of using it as one does with U.D.C., ignoring the fixed order andpermitting the free co-ordination of terms in any order which the indexer considered reasonable and possibly useful for retrieval.

2,000 master indexing cards were taken, and from these Miss Warburton regrouped the notation as made in the original indexing. A typical example of this was document P14287 which was originally given the single entry: Cd(Zqv)Juy Ncd Nfj Nfk Of Yas, with the chain index entry "Solution: Stagnation: Boundary layer: Compressible flow: Laminar flow: Angle of Yaw: Infinite: Wings". When re-indexed, this involved entries which were as follows:

Juy Cd(Zqv)
Of Nfk Nfj Ncd
Cd(Zqv)Ncd Nfk
Cd(Zqv)Nfk Ncd
Cd(Zqv)Juy

The instructions were that an average of 4 - 5 entries should be made for each document, so as to maintain the level used in the U.D.C. and Alphabetical catalogues. On completion of this task, 400 searches were made again using the elements requested in the original search programmes. The successful searches using this method came to 332, an average of 83%. This was 8% higher than Facet had achieved by the use of chain index and fixed order, and was higher than either U.D.C., Alphabetical or Uniterm in the main test.