CHAPTER 7

Additional Tests

The first year of the project coincided with a time when a number of groups, who had been investigating various methods of statistical association, were becoming interested in the possibility of putting their methods to the test, and we received some enquiries regarding the possibility of the project test collection being used as a 'common sample'. All such groups were, of course, working with computers, so with the agreement of the National Science Foundation, it was arranged that a tape should be prepared of the indexing for the 1400 documents in the test collection. This was done by I.B.M. (U.K.) Ltd., and an example of the printout for document 1420 is given in Appendix 6.1.

In the end, for various reasons, none of the groups in America was able to make use of these tapes. However, in England, Drs. Roger and Karen Needham decided to use the Cranfield collection for a test of the 'clumping' technique developed at the Cambridge Language Research Unit. (ref. 29). Since the computer to be used was the Atlas, it was necessary to prepare a set of paper tapes from the punched cards. The problems involved in this are not for us to relate, but the indexing has now been completed, and a copy of the printout for document 1420 is also given in Appendix 6.1.

At a later stage in the project, when the results were coming through, a meeting with Professor Salton made it clear that the research which he had been undertaking at Harvard was basically along similar lines to the work at Cranfield, in that both groups were concerned with comparing the performance of various index language devices, The difference lay in the methods adopted for the clerical processes of the testing, and the SMART programme (ref. 30) gave the flexibility of rapid testing of any set of documents for which the necessary relevance assessments had been made in relation to a set of questions, so long as these were in a subject field for which suitable vocabularies had been prepared. The original testing of the SMART programme had been carried out on a collection of abstracts dealing with computers, and for both groups the prospect of using the programme to test the subset of documents taken from the Cranfield project was very attractive. For Professor Salton, it gave the opportunity of testing his programme in a different subject area; for us it opened up a completely new field. There would be the opportunity for directly comparing the results of the devices being investigated at Cranfield with the similar, but more complexly calculated, devices used at Harvard. Secondly, there was the possibility that it would assist in solving some of the interesting problems involved in the presentation of results. The recall-precision curves, based on a series of cutoffs, were producing at Cranfield quite different figures from the normalised recall and normalised precision based on the ranked output at Harvard. This was only to be expected, since the method of calculation was so different, but it was important to be able to find how to equate the different sets of figures. The final point of interest was that though the Harvard searching was normally done on document abstracts, the flexibility of the SMART programme made it practical for the searches to be carried out on both the abstracts and the indexing which had been done at Cranfield, thus providing for the first time a comparison between searches based on abstracts and on indexing.

A member of the Cranfield group spent a week at Harvard, and as a result of the visit, it was arranged that a subset of the collection, consisting of 200 documents and 42 questions, should be processed at Harvard, and that searches should be made

on both the indexing of these documents and also their abstracts. Subsequently the decision was taken to extend this work so as to cover the whole of the Cranfield collection.

Citation indexing

It was in 1961 that the first major grant was given for a citation index (ref. 32), and the following year we were asked for our views on how a citation index could be evaluated. Citation indexing is basically a method of forming classes of documents which are all related through a common reference to a base document. There will, of course, be many occasions when the class consists of only a single entry; however, in the more numerous cases where the class consists of two or more documents, then citation indexing can be considered equivalent to bibliographic coupling at a strength of one. Bibliographic Coupling has been developed by Dr. M. Kessler at the Massachusetts Institute of Technology (ref. 20), and in relation to citation indexing, can be considered as a precision device, since it progressively narrows the class of documents as the demand for common references increases in number. Citation indexing and bibliographic coupling could therefore be tested in the same way as any other device; it was, however, necessary to prepare an index for this purpose.

The first stage was to prepare xerox copies of the citations in the 1400 documents in the test collection; against each citation was put the code number for the citing documents, after which each citation was cut up so as to appear on a separate slip of paper. This resulted in some 20,000 slips of many various sizes, which had to be sorted into author alphabetical order. This being done, the slips were pasted onto sheets of paper; where two or more slips related to the same cited document, only one example was pasted in; the references to the additional citing documents were entered alongside. This can be seen in Fig. 7.1, which covers a series of references to papers by H.J. Allen, in particular a paper written with A.J. Eggers entitled 'A study of the motion and aerodynamic heating of missiles entering the earth's atmosphere at high supersonic speeds.' (NACA TN 4047). This is shown to have been cited by thirteen papers in the test collection.

This procedure resulted in a normal citation index; to obtain the index for bibliographic coupling required three further stages. First, each cited reference having two or more citations was given a code number, the paper by Allen and Eggers being A25, and a separate card was prepared for each cited reference. On this reference card was written the code for the cited document (i.e. A25) and then, in numerical order the codes for the citing documents. Fig. 7.2 illustrates the reference card prepared in connection with the paper by Allen and Eggers shown in Fig. 7.1.

The reference cards were sorted into numerical order depending on the lowest number on each card. Since these numbers represented the codes for the citing documents, they ranged from 1001-2400. Each card was then taken in turn, and the information from all reference cards having the same starting number was transferred to a master card. As an example, the master card shown in Fig. 7.3 illustrates the position with regard to document 1067, this number being posted in the top left hand corner. In the column headings are entered the code numbers for the documents which have been cited by document 1067, this information being obtained from the reference cards such as Fig. 7.2, and these being, in this particular case, A25, W32, A23, E23, F90, and O24. In the first column of the master card are entered the document numbers of all other citing papers, this information being again obtained from the reference cards. A tick is put against each number in the column under the appropriate heading

Allen, D. J.: The Application of Multhopp's Subsonic Lifting Surface Theory to the Calculation of the Aerodynamic Forces Acting on a Wing of Finite Aspect Ratio Oscillating in Arbitrary Elastic Modes With Control Surface Freedom. Design Dept. Rep. No. 1191, Hawker Aircraft, Ltd. [Kingston-on-Thames, England], June 1953. 1704

D. N. de G. Allen and R. V. Southwell — Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. *Quart. J. Mech. App. Math.* III. Part 2. June, 1955. 2078 2080 (A21)

[13] Allen, F. J., *An Elastic-Plastic Theory of the Response of Cantilevers to Air Blast Loading*, BRL MR No. 880, April 1955.

Allen, H. J.: General Theory of Airfoil Sections having arbitrary Shape or Pressure Distribution. NACA Tech. Rept. No. 833, 1947. 1266 1624 (A22)

[12] Allen, H. Julian, *Pressure Distribution and Some Effects of Viscosity on Slender Inclined Bodies of Revolution*, N.A.C.A. T.N. No. 2044, 1950. 1921 1197

Allen, H. Julian: Motion of a Ballistic Missile Angularly Misaligned With the Flight Path Upon Entering the Atmosphere and Its Effect Upon Aerodynamic Heating, Aerodynamic Loads, and Miss Distance. NACA TN 4048, 1957. 2001 1067 1816 1639 2379 1716 (A23)

4. Allen, H. Julian: A Simplified Method for the Calculation of Airfoil Pressure Distribution. NACA TN No. 708, 1939. 1194

3. Allen, H. Julian: Estimation of the Forces and Moments Acting on Inclined Bodies of Revolution of High Fineness Ratio. NACA RM A9I26, 1949. 1197

8. Allen, H. Julian: Calculation of the Chordwise Load Distribution over Airfoil Sections with Plain, Split, or Partially Hinged Trailing-Edge Flaps. NACA Rep. No. 634, 1938.

[10] Allen, H. J., "Hypersonic flight and the re-entry problem," J. Aerospace Sci. 25, 217-227 (1958) 2188 1344 (A24)

Allen, H. Julian, and Eggers, A. J., Jr.: A Study of the Motion and Aerodynamic Heating of Missiles Entering the Earth's Atmosphere at High Supersonic Speeds. NACA TN 4047, 1957. 1163 1077, 1715, 1815, 1719 1164, 2346, 1982 2379 1067, 1816, 1978, 2002 (A25)

[24] Allen, Harrison, Jr., and Fletcher, E. A., *Combustion of Various Highly Reactive Fuels in a 3.84-by-10 inch Mach 2 Wind Tunnel*, NASA MEMO 1-15-59E. 2269

H. J. Allen, M. A. Heaslet and G. E. Nitzberg — The interaction of boundary layer and compression shock and its effect upon airfoil pressure distributions. N.A.C.A. RM A7A02 (1947). 1798 1070 (A26)

Allen, H. Julian, and Nitzberg, Gerald E.: The Effect of Compressibility on the Growth of the Laminar Boundary Layer on Low-Drag Wings and Bodies. NACA ACR, Jan. 1943 1073

8. Allen, H. Julian, and Perkins, Edward W.: A Study of the Effects of Viscosity on the Flow Over Slender Inclined Bodies of Revolution. NACA Rep. 1048, 1951. (Supersedes NACA TN 2044.) 1189 1927 1433 1373 1466 1464 1225 (A27)

Allen, H. Julian, and Perkins, Edward W.: Characteristics of Flow Over Inclined Bodies of Revolution. NACA RM A50L07, 1951. 1712 1197 1924 1923 2213 (A28)

12. H. J. Allen and W. G. Vincenti; "Wall Interference in a Two-Dimensional Flow Wind Tunnel, with Consideration of the Effect of Compressibility," *N.A.C.A., T.R.* 782 (1944). 1672 1714 1203 (A29)

14 Allen, R.A., Keck, J.C. and Camm, J.C., "The Recombination of Nitrogen at 6400 K, "Avco-Everett Research Lab., Research Note 243, June 1961. 1552

FIGURE 7.1 CITATION INDEX SHEET

A25			
1067	1077	1163	1164
1715	1719	1815	1816
1978	1982	2002	2346
2379			

FIGURE 7.2 CITATION INDEX REFERENCE CARD

1067

	A25	W32	A-23	E23	F90	0-24			
1077	✓								
1163	✓			✓					
1164	✓			✓					
1715	✓	✓			✓	✓			
1719	✓								
1815	✓								
1816	✓								
1978	✓								
1982	✓								
2002	✓								
2346	✓								
2379	✓		✓		✓				
1116		✓							
1150		✓							
1428		✓							
1448		✓							
1518		✓							
1531		✓							
1660		✓							
1704		✓							
1767		✓							
2137		✓							
1639			✓		✓				
1716			✓		✓				
1816			✓						
2001			✓						
1225				✓					
2329				✓					
2345				✓					
1499					✓				
1759					✓				
2000					✓				

FIGURE 7.3 CITATION INDEX MASTER CARD

to indicate that it cited this particular reference. However, when a document number appears in connection with another cited reference, the number is not repeated, but a tick is put in the appropriate column. For instance, in Fig. 7.3, it can be seen that this document has two references in common with document 1163, 1164, 1639 and 1716, three references in common with document 2379, and four references in common with document 1715. To return to Fig. 7.2, when document 1067 had been entered, then this number was crossed off and the reference card was re-sorted in the pack under the next number, namely 1077; again this number was crossed off when the master card had been entered for document 1077, and the reference card re-filed on the next number and so on until all the document numbers had been entered.

The final stage was to go through the master cards and prepare the bibliographic coupling card (Fig. 7.4). This showed the master document and all the other documents with which it had two or more references in common.

It is clearly a matter for argument as to how a citation index should be tested operationally, but within the context of these experimental investigations, it was relatively simple to decide on the method to be used. Our concern was how a citation index operated in regard to recall and precision and the procedure adopted was as follows. For a certain question, the relevant documents were known as well as their relevance level. The numbers of the relevant documents were written across the score sheet as shown in Fig. 7.5, referring to question 34. In order to avoid complexity, a fairly simple example has been taken, where there were six documents all of relevance 3.

The numbers in the left hand column indicate the coupling strength going from a maximum of 6+ down to 1, which latter represents citation indexing. The appropriate bibliographic coupling cards were then taken from the pack, the first of these relating to document 1067. As can be seen in Fig. 7.4, document 1715 had a match of 4 with document 1067; there were no other documents at this level of match, and since document 1715 is also a relevant document to question 34 (see Fig. 7.5), this is counted as a success and the score is entered appropriately. The document which matches at a level of 3 is not relevant, so this now makes the score one relevant and one non-relevant. At a match of 2, three of the documents are relevant (1164, 1639 and 1716), so the score here becomes four relevant and two non-relevant. By referring to the cards shown in Fig. 7.3, we can calculate the number of documents involved with a single match. There are no other relevant documents in this set, but many non-relevant, so the score for this is shown as four relevant and thirty-two non-relevant.

This process is repeated for all the other relevant documents, as shown in Fig. 7.5. When this has been done, the scores can be totalled to give a set of figures where obviously the maximum recall and the lowest precision will be obtained at a match of 1, and maximum precision with lowest recall obtained at a match of 6+. However, there are various approaches that can be taken in compiling the score, and these will be considered in the volume of test results. Such analysis was done for documents of all degrees of relevance.

In bibliographic coupling as discussed by Kessler, account is only taken of the actual match rather than what might be called the proportional match. For instance, two review-type articles may each have fifty references, as against two other papers which have only three references. If the former pair of papers have five references in common, this would be considered a stronger coupling strength than the latter pair

1067						
2	3	4				
1163 1164 1639 1716	2379	1715				

FIGURE 7.4 BIBLIOGRAPHIC COUPLING CARD

RELEVANT DOCUMENTS

COUPLING STRENGTH	1067		1164		1639		1715		1716		1717	
	R	N-R	R	N-R	R	N-R	R	N-R	R	N-R	R	N-R
6+												
5+												
4+	1	0	0	1			1	0				
3+	1	1	0	2	1	0	1	0	1	0		
2+	4	2	1	6	2	3	1	5	2	2	0	3
1+	4	32	1	49	2	63	1	47	2	33	0	18

SCORE

	RELEVANT	NON-RELEVANT	RECALL RATIO	PRECISION RATIO
6+	-	-	-	-
5+	-	-	-	-
4+	2	1	33%	66%
3+	4	3	66%	57%
2+	5	19	83%	21%
1+	5	201	83%	2%

FIGURE 7.5 SCORE SHEET FOR BIBLIOGRAPHIC COUPLING FOR QUESTION 34.

1067 (6)	
3.	1715 (9)
7.	1716 (7)
11.	2379 (17)
12.	1163 (9)
16.	1639 (11)
25.	1164 (17)

FIGURE 7.6 RECALCULATED BIBLIOGRAPHIC COUPLING CARD
Figures in brackets represent number of references in documents. Figures in first column give the calculated weighting.

which have, for example, two references in common. However, proportionately, it could be argued that the latter represents a stronger match than the former. To test this, the number of references in each document of a matching pair were multiplied, the resultant figure was then divided by the square of the number of matches and the final figure was considered as the level of coupling strength. For example, document 1067 (see Fig. 7.4) had six references, and document 1163 had eight references, giving a multipland of 48. These had a match of 2, so dividing by 2^2 gives a final weighted figure of 12. Document 1715, however, had nine references which, combined with document 1067, gives a multipland of 54. In this case, since there is a match of four, this figure has to be divided by 4^2, giving a final weighted figure of 3. When the matches for document 1067 had all been worked out, the weighting becomes as in Fig. 7.6. In many cases, the result of this exercise showed significant changes in coupling strength, and therefore the collection was re-tested in the manner described earlier, only this time the scoring was based on these new coupling levels.