CHAPTER 4

STATISTICAL ANALYSIS

In addition to the results as given in tables in Chapter 3, there is the possibility of doing a considerable amount of detailed analysis, aimed at finding the effect of various factors. The difficulty in doing this lies in pre-assessing what is likely to be significant, and the danger is that effort will be wasted in carrying such analysis to unnecessary and useless limits. The result is that in this chapter certain matters will be considered, but any conclusions will be based on a varying amount of analysis. Where something of interest appears, to ensure that the results are valid more analysis has been done than in those cases which appear to be producing negative results.

First is considered the question of language and this is done in a simple way by comparing the correlation between the words of the questions and the titles of the documents. Appendix 4A lists 100 questions numbered 20-06 to 20-10 through to 39-06 to 39-10. Appendix 4B lists 100 documents on which these questions are based and which are therefore the objects of the searches. The titles and questions have been compared and the correlation figure varying from 0 to 10 has been given for each question. This has been done by marking with 10 any question where all the key words in the question were found in the title and grading down to 0 where there was no correlation. Table 4.1 shows the 100 questions sorted and grouped into correlation order, with the search results for the four systems.

Correlation Figure	Question Number	SEARCH RESULT			
		UDC	ALPHA.	FACET	UNITERM
0	21-10	f	f	f	f
	29-09	f	f	f	f
	30-06	f	f	f	f
	30-09	s	s	s	s
	30-10	s	s	s	f
	35-06	s 3-3	s 3-3	s 3-3	s 2-4
1	20-09	s	s	s	s
	23-07	f	s	s	f
	26-07	f	s	f	s
	33-10	s 2-2	s 4-0	s 3-1	s 3-1
2	20-06	s	f	s	s
	22-06	s	s	s	s
	22-07	s	s	s	s
	22-10	s	s	s	s
	26-09	s	s	s	s
	26-10	f	s	s	s
	33-09	s	s	f	s
	39-07	f 6-2	f 6-2	f 6-2	s 8-0
3	21-08	s	f	s	s
	23-08	s	s	s	s
	24-10	s	s	s	s
	27-10	f	s	f	s
	29-07	s	f	f	s
	30-07	f	f	f	s
	31-09	f	f	s	f
	32-09	s	s	s	f
	32-10	s	s	s	f
	35-08	s	s	f	s
	36-06	f	s	s	s
	37-08	s	s	f	s
	38-09	s 9-4	s 9-4	s 8-5	s 10-3
4	21-06	s	s	s	s
	21-07	f	s	s	s
	21-09	f	s	f	s
	22-09	s	s	s	s
	23-09	s	s	s	s
	25-09	f	s	s	s
	25-10	s	s	s	s
	27-06	s	f	s	s
	28-07	f	s	s	s
	21-06	s	s	f	s
	38-06	f	f	f	s
	38-10	s 7-5	s 10-2	s 9-3	s 12-0

TABLE 4.1

RESULTS OF SEARCHES GROUPED ACCORDING TO CORRELATION OF LANGUAGE OF TITLE AND QUESTION

s = successful search
f = failure

Continued.....

Correlation Figure	Question Number	SEARCH RESULTS			
		UDC	ALPHA.	FACET	UNITERM
5	20-07	s	s	s	s
	22-08	s	s	s	s
	25-06	s	s	s	s
	25-07	s	s	s	s
	25-08	f	f	f	s
	28-09	s	s	s	s
	29-06	s	s	s	s
	31-10	f	f	f	s
	32-08	s	s	s	s
	34-08	f	s	s	s
	34-10	s	s	s	s
	36-07	s	f	f	s
	38-08	s	s	s	s
	39-09	s 11-3	s 11-3	s 11-3	s 14-0
6	20-10	s	f	f	s
	23-10	s	f	s	s
	24-06	f	s	s	s
	27-08	f	s	f	s
	28-08	s	s	s	s
	28-10	s	s	s	s
	33-06	s	s	s	s
	34-07	s	s	f	s
	34-09	s	s	f	s
	35-07	s	s	s	s
	35-09	f	s	s	s
	36-09	s	s	s	s
	39-06	s	f	f	f
	39-10	f 10-4	s 11-3	s 9-5	s 13-1
7	24-07	s	s	s	s
	24-08	s	s	s	s
	24-09	s	f	f	s
	26-08	s	f	f	f
	27-07	s	f	f	s
	32-07	s	s	s	s
	33-08	s	s	s	s
	34-06	s	s	s	f
	35-10	f	s	f	s
	36-08	s	s	s	s
	36-10	s	s	s	s
	38-07	s	s	s	s
	39-08	s 12-1	f 9-4	s 9-4	s 11-2

TABLE 4.1 (Continued)

		SEARCH RESULT			
Correlation Figure	Question Number	UDC	ALPHA.	FACET	UNITERM
8	20-08	s	f	f	s
	26-06	s	f	s	f
	28-06	s	s	s	s
	29-10	s	s	s	s
	33-07	s	s	f	s
	37-06	s	s	s	s
	37-09	s	s	s	s
	37-10	f 7-1	s 6-2	s 6-2	s 7-1
10	23-06	s	s	s	s
	27-09	s	s	s	s
	29-08	f	s	s	f
	30-08	s	s	f	f
	31-06	s	f	s	f
	31-07	s	s	f	s
	31-08	f	s	f	s
	37-07	s 6-2	s 7-1	s 5-3	s 5-3

TABLE 4.1 (Continued)

This type of analysis was extended to cover a total of 600 searches, and Table 4.2 gives the percentage success rate for each correlation figure together with the accumulated total.

Correlation	UDC	ALPHA.	FACET	UNITERM
0	48 (48)	45 (45)	52 (52)	44 (44)
1	60 (53)	74 (53)	65 (58)	70 (52)
2	72 (60)	70 (58)	63 (59)	85 (63)
3	69 (64)	73 (63)	74 (62)	79 (65)
4	66 (65)	84 (69)	67 (63)	89 (72)
5	77 (68)	74 (70)	71 (65)	83 (74)
6	72 (67)	79 (72)	68 (65)	91 (77)
7	88 (72)	85 (74)	75 (67)	92 (80)
8	84 (74)	90 (77)	80 (70)	87 (82)
9	81 (76)	89 (79)	78 (72)	90 (84)
10	82 (77)	92 (82)	77 (73)	89 (86)

TABLE 4.2

PERCENTAGE SUCCESS RATE FOR SEARCHES WITH VARYING DEGREES OF CORRELATION BETWEEN TITLE AND QUESTION

(Figures in brackets represent accumulated percentage figure)

Leading on from this, a test was made independently, of the retrieval figure which would be obtained by using a form of key-word-in-context title indexing, but as will be appreciated this in effect duplicated the work of the term correlation test, for it was found that 97% of the source documents could be retrieved by this technique. This figure corresponds with the 3% of source documents not located when there was zero correlation, and also corresponds to the figure of 3% of the questions which (as considered in Chapter 5) were considered to be misleading or bad. Not quite so explicable is that whereas the maximum recall efficiency by any single system was 85%, yet by one system or another the source document was obtained in 97 out of every 100 cases. The fact that 97% retrieval was obtained by KWIC indexing is not so much an argument in favour of this technique as an interesting commentary on the effectiveness of the titles of the reports and papers used in the project. To achieve the figure of 97% would have been equally possible by Uniterm if the search rules had been relaxed in such a way as to accept as a success any search where a single correlating term was found. The penalty in this case, as with KWIC indexing, would have been the increased number of irrelevant documents that were also retrieved. This aspect is considered in more detail in Chapters 9 and 10.

It can be argued that some questions were too easy or obvious, and that a more realistic result might be given by eliminating such questions from the analysis. It is difficult to decide exactly which question is easy, but an arbitrary choice would be to consider as such all questions where the source document was retrieved by the four systems. This was done for 300 searches and Tables 4.3. and 4.4. show the general figures and the figures when broken down by indexing times.

UDC	54%
ALPHA.	58%
FACET	43%
UNITERM	63%

TABLE 4.3

PERCENTAGE RETRIEVAL EFFICIENCY FOR SEARCHES BY PROJECT STAFF AFTER ELIMINATING QUESTIONS WHERE SOURCE DOCUMENTS WERE RETRIEVED BY ALL SYSTEMS

	16 mins	12 mins	8 mins	4 mins	2 mins
UDC	49	60	54	58	54
ALPHA.	66	67	58	67	52
FACET	38	52	39	44	46
UNITERM	70	69	62	76	45

TABLE 4.4

PERCENTAGE RETRIEVAL EFFICIENCY BY INDEXING TIMES FOR SEARCHES BY PROJECT STAFF AFTER ELIMINATING QUESTIONS WHERE SOURCE DOCUMENTS WERE RETRIEVED BY ALL SYSTEMS

As would be expected, the figures represent a departure from those in the main test, but it is noteworthy how they repeat the general relationship between the systems and, in particular, the way in which the efficiency of 4-minute indexing time is emphasised.

In the first stage of the work, the indexing decisions were partly controlled by time, and the effect that this had on the number of entries for each document is shown in Table 1.2. (page 4). This variation might be expected to appear in the results when broken down by time (Table 3.2, page 22), but from this it would appear that many of the additional entries made when indexing at 16 minutes were redundant, since there was such a small increase in efficiency of recall. This problem of redundant indexing was considered in the W.R.U. Project (Chapter 7), but a comparison of the effect of number of postings on success or failure is given in Table 4.5.

Appendix 4C gives the results of 100 searches shown against the number of entries made in indexing the source document, and the analysis has been carried to a further 200 searches within this document group. The resulting figures, broken down by indexing time are given in Table 4.5, but would appear to show little of significance.

		General	Search Average	
	minutes	Average	Success	Failure
U.D.C.	16	5.8	6.0	4.3
	12	4.0	4.9	2.9
	8	4.2	4.5	3.8
	4	3.7	4.7	2.8
	2	3.5	3.4	3.8
ALPHA.	16	3.8	4.1	2.7
	12	3.5	3.8	3.1
	8	3.2	3.5	2.7
	4	2.2	2.3	2.0
	2	2.4	2.5	2.2
FACET	16	6.8	6.8	6.8
	12	6.8	6.7	6.9
	8	6.8	6.0	7.7
	4	4.7	4.8	4.3
	2	5.0	5.2	4.3
UNITERM	16	13.0	13.0	13.0
	12	9.2	9.9	7.8
	8	10.6	12.3	7.4
	4	8.5	8.8	7.7
	2	8.5	8.6	8.2

TABLE 4.5

COMPARISON OF AVERAGE POSTINGS BY INDEXING TIMES FOR DOCUMENTS P12001 to P14000 FOR SUCCESSFUL AND FAILED SEARCHES BY PROJECT STAFF

A more interesting set of figures is produced by the analysis of the terms used in indexing source documents and the terms used in searching for these documents. The complete lists of terms in the four systems are given in Appendix 4D, and Table 4.6 shows the comparative count for the use of the terms.

	UDC	ALPHA.		FACET	UNITERM
		Main Headings	Sub-Headings		
Terms used in indexing Total	381	272	155	396	582
Indexing and searching	252	141	86	240	347
Indexing only	129	131	69	156	235
Terms used in searching Total	408	259	107	343	458
Indexing and searching	252	141	86	240	347
Searching only	156	118	21	103	111

TABLE 4.6

USE OF TERMS IN INDEXING 200 SOURCE DOCUMENTS AND MAKING 200 SEARCHES

A general comment on these figures is the high number of non-indexed terms used in searching and particularly that with U.D.C. the total of terms used only in searching is greater than that used only in indexing. It is, however, in the more detailed analysis of the figures presented in Appendix 4D that useful work could be done, although no attempt has been made to exploit this possibility within the present project. As an example, one can consider the matter of redundant terms, of which some are set out in Table 4.7.

	UDC	ALPHA.	FACET	UNITERM
Tests	14-3	59-5	46-4	48-17
Wind tunnel tests		42-1	24-2	
Calculations		62-11	52-1	36-14
Analysis		15-2	10-0	16-0
Design	9-1	18-8	12-3	22-16

TABLE 4.7

POSSIBLY REDUNDANT TERMS SHOWING NUMBER OF TIMES USED IN INDEXING AND SEARCHING

It is practically certain that none of these terms assisted in locating a source document which would otherwise not have been retrieved, and all that these terms could do would be to limit the number of irrelevant documents. An analysis of the point could be made by checking the catalogues, and would probably confirm the view that such terms were, in the context of the work, redundant.

As another example, there was a category of terms (notation 'I') in the Facet schedules covering Spatial Properties. Most of these terms are not available in U.D.C. 6 such terms were used in the indexing of the sample 200 documents, with a total of 65 uses. 20 of these terms were also used in searching on 37 occasions. This shows a slightly lower average use than for the terms throughout the schedules, but is high enough to justify the inclusion of this category. The category General Properties (Z) had 16 terms which were used on 34 occasions in indexing, but only 5 terms were used in searching on 6 occasions, and this category appears of more doubtful value. Analysis of this nature would appear to have possibilities of providing information for the design of systems, and it is hoped that it will form the subject for an independent thesis.

As was discussed in Chapter 2, note was taken of the number of separate programmes which were necessary to retrieve the source document. Appendix 4E shows the results for 100 searches and indicates how many search programmes were involved by each system. This analysis has been extended to cover the 400 searches in the second round of testing, and the results are given in Table 4.8.

No. of searches	UDC	ALPHA	FACET	UNITERM
1	36%	40%	51%	54%
2	39%	33%	36%	25%
3	8%	13%	9%	16%
4	10%	5%	1%	2%
5	6%	4%	1%	3%
6 or more	1%	5%	2%	0

TABLE 4.8

PERCENTAGE OF NUMBER OF SEARCHES REQUIRED IN SECOND ROUND OF TESTING

These cover the second round of testing, when the searches generated their own programmes. In the third round of testing, as has been explained, the search programmes were standard for all systems, so a further check was made of the searches in this round. As will be seen from Table 4.9, this did not show any significant difference, apart from a slight general improvement.

The number of searches which retrieved the source document with the first or second programmes indicates that, with the techniques and levels of indexing used in the project, the formulation of a successful programme is not particularly difficult, but other aspects of this point will be considered in the discussion on the failures in Chapter 5.

No. of programmes	UDC	ALPHA.	FACET	UNITERM
1	42	42	50	57
2	39	36	39	26
3	10	14	8	15
4	7	4	1	1
5	2	2	1	1
6 or more	0	2	0	0

TABLE 4.9

PERCENTAGE OF NUMBER OF SEARCHES REQUIRED
IN THIRD ROUND OF TESTING