VIII EVALUATION EXPERIMENTS

VIII . 1 Review of experiments

As noted earlier, the major evaluation tests of manual indexing systems which have been carried out in the last decade provide the background for the evaluation of automatic indexing. Thus the Comparative Systems Laboratory (Saracevic 1968, 1971), Cranfield (Cleverdon 1966), Inspec (Aitchison 1970) and ISILT (Keen 1972, 1973) tests all found simple manual keyword indexing very competitive with other languages. (Collection details for all the tests referred to in this section are summarised in Table 1.) In particular, really elaborately controlled or structured languages typically performed less well, though there is some evidence that relatively simple controlled languages have a slight performance edge. There is also some evidence that manual thesauri can be improved by taking account of vocabulary distribution properties: see, for example, Salton 1968a.

Unfortunately these evaluation tests were done on different collections, and there was no comparison between collections by individual projects. The experiments were also confined to small collections. Tests of operational systems like that conducted by Lancaster 1968a may well involve large collections, but do not provide the comparative information we need here. Evaluation tests with automatic indexing techniques tend to suffer from the same defects as these manual ones. Many of the Smart experiments, for example, have been carried out on the 35x82 ADI collection. However tests with several different collections have been conducted by the Smart project, and by Sparck Jones, and the rather variable results obtained serve as useful reminders that the relative performance of different devices may vary over collections. Fortunately, the common use of particular collections is increasing, which makes it more rational to attempt to relate different project results to one another. The 42x200 Cranfield collection has been widely used, and there are several others which have been exploited by more than one project. It should, however, be noted that different versions of these collections may be involved. which inhibits detailed comparisons. Special difficulties also arise in making comparisons when different matching functions are used. Strict Boolean searches, simple coordination level matching, and normalising matching coefficients all produce rather different forms of output. Again, averaging over sets of requests may be done differently.

In Section II, I categorised indexing techniques as simple or complex. An example of simple indexing is the use of keywords taken from titles, with word truncation allowed in searching. Complex indexing is illustrated by the use of a controlled language like MeSH. The evaluation of automatic indexing must refer to the comparative performance of simple and complex manual indexing, and itself requires comparisons between simple and complex automatic techniques, and between these and manual ones.

It is a pity that the evidence available for making all these comparisons is so fragmentary. The situation is complicated by the fact that different indexing sources (say title or abstract) may be used, and by the fact that automatic indexing systems may have manual components: an example is Sparck Jones' use of manual keyword lists as input to statistical classification. The discussion which follows is necessarily rather schematic: thus points like the exhaustivity implications of different sources are disregarded, and approaches are defined as manual or automatic according to their predominant characteristics. After all, at this stage what we are interested in is the overall performance of indexing methods, rather than the reasons for it (this is particularly important in relation to differences between projects associated with the use of different matching techniques).

The following attempts to summarise the useful results which have been obtained. I shall use the symbol

$A = B$ to mean that the performance of methods A and B is not noticeably different,

$A > B$ to mean that the performance of method A is noticeably better than that of B,

$A \gg B$ to mean that the performance of method A is materially better than that of B, and

$A \geqslant B$ to mean that the performance of method A ranges from the same as that of B, to noticeably better than that of B.

Manual indexing

The experiments referred to above suggest that where complex indexing methods refer to fairly straightforward subject headings or thesaurus terms, and simple ones to extracted keywords, or phrases, with truncation:

complex $\geqslant$ simple.

Automatic indexing

Relevant comparisons here are those concerned with the use of statistical association techniques versus keywords. (Actual experiments have been based on both automatically and manually extracted keyword lists.) Relevant tests are those by Lesk 1969, Minker 1972, 1973, Sparck Jones 1971a, 1973c and Vaswani 1970. The result, as appeared in Section IV, is

complex $=$ simple.

Allowance should be made for the fact that there is some variation in the results with simple techniques. Subsidiary comparisons are therefore of interest. These concern a) titles versus abstracts and b) weighting.

a) Comparisons between simple automatic indexing from titles and abstracts have been made by Aitchison 1970, Barker 1972a, Cleverdon 1966 and Tell 1970, as well as the Smart project (Salton 1968a,c). There is some variation in the results, probably due to different matching techniques. Aitchison, Barker and Tell found that titles gave better precision than abstracts, but worse recall, on Boolean matching. With coordination levels Aitchison and Cleverdon found titles superior to abstracts (except for the recall ceiling); while with cosine correlation Salton found abstracts better than titles at higher recall, or overall. Assuming some interest in recall, the final result must be

abstract $\geqslant$ title.

b) Experiments with keyword weighting schemes of various types by Salton 1972c, 1973b,c and Sparck Jones 1972, 1973e show that some forms of weighting may be useful, and specifically that collection frequency based weighting can be helpful. These tests involved both automatically and manually obtained keywords, so the results are applicable to simple manual indexing as well as simple automatic indexing. Allowance should therefore be made in the comparisons which follow for the possibility that simple indexing performance can be improved.

Automatic versus manual indexing

The number of experiments directly comparing automatic and manual indexing is small.

1. simple automatic v. simple manual

a) titles v. keywords.

These have been compared by Aitchison 1970, Barker 1972a and Hansen 1973 for Boolean search, Aitchison and Cleverdon 1966 for coordination levels, and Salton 1968a,c for correlation matching. Aitchison found titles better on precision and worse on recall, Barker found titles gave slightly better precision and worse recall, and Hansen found the two the same. Aitchison found keywords superior with coordination levels while Cleverdon found them much the same. Salton's one test with the Cranfield collection also shows them much the same.

So we conclude

keywords $\geqslant$ title

b) abstracts v. keywords.

These have also been compared by Aitchison 1970 for Boolean matching, Aitchison, Cleverdon 1966 and Sparck Jones 1973d for coordination levels, and Salton 1968a,c for correlation matching. Aitchison's Boolean matching showed the same performance. With coordination levels all three projects found keywords superior, while Salton (see also Lesk 1968) also shows a slightly better performance with keywords in his comparison for the Cranfield collection. We therefore have

keywords $\geqslant$ abstract.

2. simple automatic v. complex manual

The comparisons here mostly concern titles versus fairly straightforward subject headings or thesaurus terms. Comparisons have been made for Boolean searching by Aitchison 1970, Miller 1971a, Olive 1973, Saracevic 1968, 1971 and Tell 1971; for coordination levels by Aitchison and Cleverdon 1966, and for correlation by Salton 1968a,c. The results tend to show keywords performing better, though there is variation in the test findings. In coordination level matching Aitchison found titles inferior, while Cleverdon found their performance much the same. In the Smart experiments they were inferior. Cleverdon and Salton also compared abstracts and manual indexing, in both cases finding the former inferior. However the recent Smart experiments with weighting (Salton 1972c, 1973b), directly comparing improved simple automatic indexing for abstracts with subject headings, showed the same performance. Nevertheless the overall conclusion must be

complex manual $\geqslant$ simple automatic.

3. complex automatic v. simple manual

Most of the projects investigating statistical association techniques compared performance with simple keywords, but in some cases the keywords were extracted automatically. Comparisons with manual keywords were made for coordination levels by Sparck Jones 1971a, 1973c and for cosine correlation by Lesk 1969. The results were very variable, so it must be concluded that

complex automatic $=$ simple manual.

4. complex automatic v. complex manual

Apparently the only comparison between these two forms of indexing is that made by the Smart project. Lesk 1969 illustrates the results for three collections. The balance of the evidence is in favour of thesauri as opposed to statistical associations, so we have

complex manual $\geqslant$ complex automatic.

These remarks are based on the explicit results given for particular tests in the publications concerned. The fact that different tests and collections tend to be involved under the various headings may mean that the conclusions I have drawn are not necessarily consistent. I shall attempt to pull the threads together in Section IX.

VIII . 2 The SMART Project

The Smart project is the largest and longest term research project in automatic information retrieval. There are few questions connected with automatic indexing on which it has not done some work over the last ten years. Individual publications have been mentioned as appropriate; but it is useful to look briefly at the Smart research as a whole, to see what overall conclusions about automatic indexing can be drawn from it.

Details of the work are given in the projects Information Storage and Retrieval Reports (Salton 1966-). Early research is summarised in Salton 1968a,c. Selected papers are reprinted in Salton 1971a. The investigations can be grouped under four heads. The relevant papers are listed in Table 2. The test collections exploited are listed in Table 3.

1. Experiments on basic automatic indexing.

2. Experiments on automatic index language generation.

3. Experiments on document clustering.

4. Experiments on relevance feedback techniques.

5. Experiments concerned primarily with methodological questions.

1. Early experiments investigated the simple use of non-trivial keyword stems from titles and abstracts, and compared these with manual indexing on the one hand, and manual and automatic methods of organising the extracted vocabulary on the other. The results showed that simple keyword methods, though performing less well than manual thesauri, did not perform conspicuously less well. These tests were also concerned with the effects of within-document frequency weighting, and with different matching coefficients. They showed that alternatives here could affect performance, and that on the whole the best results were obtained with weights, and a normalising matching coefficient like cosine correlation.

2. Initial experiments with index language generation involved statistical association lists for expanded documents and requests. These were not particularly successful, no real improvement over simple keywords being obtained. Recent experiments in controlling an indexing vocabulary either by deleting non-discriminating or common words, or by differential weighting using collection frequencies, show that noticeable performance gains can be obtained, the results being very competitive with independent manual indexing using a controlled vocabulary.

3. Clustering experiments have been based on centroid techniques. The object of clustering is seen as economic; the results show performance losses. This line of work has not been very productive, perhaps because the clustering methods were not really adequate.

4. Tests with relevance feedback have examined the use of information for relevant and non-relevant documents retrieved in an initial search in modifying requests for new searches: relevant terms in the request may be upgraded, and non-relevant ones downgraded. The tests generally show that noticeable performance improvements can be obtained by

techniques which do not involve the user in very much effort, and which represent automatic revision of indexing. The correct characterisation of performance in this type of experiment is not readily determined, and care is needed in interpreting the ordinary recall/precision graphs given. Tests have also been carried out showing that permanent changes to document descriptions may be useful.

5. The range of experiments done has raised a number of methodological and related questions, and some experiments have been carried out to examine points involved in feedback evaluation (1970c), relevance judgement variation (1968d), generality (1972b) and the use of mixed language data bases (1970b). It cannot be claimed that the problems involved are all resolved, but more obvious criticisms may be overcome.

It should be emphasised that direct comparisons with manual indexing have been few: for the 42x200 Cranfield collection, and, recently, for the 29x450 Medlars collection. The latest results are promising for automatic indexing, but it would be nice to have more comparative evidence of this sort.

The main weakness of the Smart experiments has been the small scale of the collections mainly used, apparent in Table 3. A further difficulty is that different tests may have been carried out with different collections, making detailed cross checking rather difficult. Further complication appears where different versions (stem, thesaurus) of particular collections are involved. It is to be hoped that some larger scale experiments will be carried out with the more successful techniques, and that fairly rigorous cross comparisons will be made. A point which should perhaps also be made is that where statistical significance tests may justify the assertion that method A is better than method B, in many cases the real difference in performance is not large.

Table 1 Evaluation test collections

Aitchison 1970	97	x 542	Inspec	Electrical engineering
Barker 1972a	193	nK	CAC	Chemistry
	48	nK	CBAC/POST	Chemistry
Cleverdon 1966	221	1400	Cranfield	Aeronautics
	42	200	Cranfield	Aeronautics
Hansen 1973	20	nK	CAC	Chemistry
Keen 1972,1973	63	800	ISILT	Documentation
Lancaster 1968a	302	nK	Medlars	Medicine
Lesk 1968,1969	42	200	Cranfield	Aeronautics
	35	82	ADI	Documentation
	34	780	IRE	Computing
Miller 1971a	25	nK	Medlars	Medicine
Minker 1972	34	780	IRE	Computing
	18	273	Medlars	Medicine
1973	34	780	IRE	Computing
	18	273	Medlars	Medicine
	35	82	ADI	Documentation
Olive 1973	60	12765	NSA	Nuclear science
Salton 1968a,c	42	200	Cranfield	Aeronautics
	35	82	ADI	Documentation
	34	780	IRE	Computing
1972c,1973b	29	450	Medlars	Medicine
1973c	24	450	Medlars	Medicine
	24	424	Cranfield	Aeronautics
	24	425	Time	World affairs
Saracevic 1968,1971	24	600		Tropical diseases
Sparck Jones 1971a	42	200	Cranfield	Aeronautics
1972,1973c,e	42	200	Cranfield	Aeronautics
	63	797	Keen (ISILT)	Documentation
	97	541	Inspec	Electrical engineering
1973d	42	200	Cranfield	Aeronautics
	47	407	Keen	Documentation
Tell 1971	7	nK	Inspec	Electrical engineering
	53	nK	POST	Polymer science
Vaswani 1970	93	11571		mixed

Table 2 Smart publications; Salton unless otherwise indicated

1. basic	1968a,c, 1970a, Keen 1967, Lesk 1968
2. index language	1968a,c, 1969a, 1972a,c, 1973b,c, Lesk 1968,1969
3. document clustering	1968b, 1972d, Rocchio 1966, Dattola 1969, Kerchner 1971, Murray 1972
4. relevance feedback	1968b, 1969b,c,d, 1972d, Brauen 1969, Ide 1969, Kerchner 1971
5. methodology	1968d, 1970b,c, 1972b

		ADI	IRE			Cranfield								Ispra		Medlars				Time	
	Reqs	35	17	24	34	42	36	22	22	24	155	36	225	48	48	18	29	24	29	83	24
	Docs	82	375	375	780	200	200	200	424	424	424	1400	1400	1268	468 1095	273	450	450	852	425	425
Salton	1968a	x			x	x															
	1968b	x		x		x															
	1968c	x			x	x															
	1968d													x							
	1969a															x					
	1969b					x															
	1969c					x															
	1969d					x															
	1970a -																				
	1970b														x						
	1970c -																				
	1971a	x	x	x	x	x					x			x	x	x					
	1971b					x															
	1972a	x				x															
	1972b						x	x	x			x		x							
	1972c																x				
	1972d																x				
	1973a																			x	
	1973b																x		x		
	1973c									x								x			x
Keen	1967	x			x	x															
Lesk	1968	x	x			x															
	1969	x			x	x															
Brauen	1969										x										
Dattola	1969	x				x															
Ide	1969					x															
Kerchner	1971												x		x						
Murray	1972												x								

<u>Table 3</u> Smart test collections