THE SIGNIFICANCE OF THE CRANFIELD TESTS ON INDEX LANGUAGES

by
Cyril W. Cleverdon

1946 saw the lifting of the security restrictions on large numbers of scientific and technical reports which had been written during World War Two. Pre-war virtually all publication had been in journals, and the report format was strange and unfamiliar, both for the scientific community and for librarians. As such they presented new challenges; the administrative problem of actually being able to obtain copies of the reports was tackled by setting up new government agencies with direct responsibility for collecting and making the reports generally available. The more difficult problem lay in revealing and making accessible the intellectual content of the papers. At that time there were two conventional types of index and two major indexing techniques. An index could be in the form of a card catalogue, as found in most libraries, or alternatively in printed form as, for example, an annual accumulation of an abstract journal. Regarding the techniques of indexing, in Europe there was a tendency to use a classified system, whereas in America the usual practice was to use alphabetical subject headings.

With the deluge of scientific and technical reports, both the physical form of the index and the indexing techniques came under strong attack. While card catalogues and printed indexes still exist, there has been over the past forty years a steady and reasonably placid progress of mechanised systems, culminating now in online systems and CD-ROM, but there was nothing placid about the development of indexing techniques. The early 50s saw many attempts to depart from the conventional systems. In England a small group met regularly to discuss the development of facet classification. This technique breaks away from the conventional enumerative or hierarchical classification, such as the Dewey Decimal Classification, and relies on subject analysis and synthesis by facet principles. However the main thrust of the new methods was in America, from such people as Calvin Mooers with Zatorcording, James Perry with semantic factoring and, in particular, Mortimer Taube. Taube, a government librarian, analysed some 40,000 subject headings used in a major card catalogue and found that the headings were combinations of only some 7,000 different words. He therefore proposed using these individual words as index terms which would be coordinated at the searching stage. This became known as the Uniterm System.

These new techniques generated considerable argument, not only between the proponents of the different systems, but also among the library establish-ment, many of whom saw these new methods as degrading their professional mystiques.

This briefly is the context in which I started my research. In 1946,

after working in public and industrial libraries, I had become Librarian of a small post-graduate College of Aeronautics at Cranfield. In 1952 I became involved in the activities of an aeronautical information panel, which was attached to a group of NATO, and which met every six months in various NATO capital cities. The United States representative on this panel brought to the meetings some of the proponents of the new techniques, and it was from them that I learnt at first hand about the developments. I was immediately attracted by the apparent simplicity of the Uniterm system, so much so that with a colleague, Bob Thorne, we did a small test. For this I indexed 200 papers, and Thorne carried out a number of searches. It was, by any standards, a trivial piece of work, but we were presumptuous enough to issue a short report. (Ref. 1) This was completely ignored in England but aroused some interest in the States, no doubt because of the strong opposition to the Uniterm system by professional librarians.

As an aside, it may be of interest to this Group to note that it was this report which was said to have been the catalyst that popularised the term 'information retrieval'. The original use of this term was by Calvin Mooers in 1951, but as his report was not widely circulated, it made no impact at the time.

I had also been involved in testing a system proposed by the Nationaalluchtvaartlaboratorium of the Netherlands and as a result had come to the conclusion that, for a valid comparison between systems, it would be necessary to control conditions in such a way that performance could be related to economic factors.

While attending a NATO meeting in Ottawa in 1955, I went to Detroit to give a paper at the Special Libraries Association Conference. Controversy over the new methods was still raging, with extravagant claims on one side being countered by absurd arguments on the other side, without any firm data being available to justify either viewpoint. A recent editorial in American Documentation (Ref. 2) had read, "Cautious and searching evaluation of all experimental results is essential in rating the efficiency of documentation systems. May the age-old controversies that arose from the conventional systems not be reborn in the mechanised searching systems of the future". I took this as the theme of my paper, arguing that an independent evaluation of the rival claims was needed, and outlining how this might be done. Mrs. Helen Brownson of the National Science Foundation happened to be in the audience, expressed interest in my suggestion and the outcome was that two years later the Foundation made a grant of $28,000 to cover a two year project on a comparative evaluation of four systems. These were to be a conventional classification, namely the Universal Decimal Classification, a conventional alphabetical subject index, a purposely devised schedule of a facet classification and the Uniterm System of Coordinate Indexing. 18,000 papers in the field of aeronautical engineering were to be indexed by each of these four systems, and 1,200 search questions were obtained for testing.

In line with my views on economic factors, a number of controls were built in to the indexing process; for example, the three indexers were of different qualifications and experience, and within batches of 100 documents the time allowance for indexing each paper ranged from 2 minutes to 16 minutes. These and other variations were repeated every 6,000 papers so that the learning process, if any, could be evaluated. Card catalogues, prepared for three of the systems, contained some 160,000 cards while the inverted file for the Uniterm system had 3,100 aspect cards.

The search questions had been obtained from several hundred individuals in 58 different organisations, mainly in England and America. Each question was based on a single document in the test collection, and a search was considered successful if

that particular paper was located in the catalogue. The results of the test showed that all four systems were achieving in the region of 74 - 82% efficiency in retrieving the required paper, with the Uniterm system showing a slight advantage, and the facet classification having the lowest rank. The searches were repeated by a group of students, mainly to check whether there was any validity in the argument that classification systems were more difficult for an end-user than an alphabetical subject index, but the results showed no difference from the searches by project staff. Analysis indicated that there was no increase in performance for an indexing time of more than four minutes.

A detailed analysis was made of all cases where there had been a failure to locate the required document. This showed that the large majority of these failures were due to human error, either in the indexing, in the searching, or in the clerical processes of preparing the catalogues. Only one failure in 20 could in any way be tied in to the indexing system. In other words there was a strong presumption that the particular system used appeared to have no significant effect on performance.

The publication of the final report (Ref. 3) attracted wide interest, caused considerable annoyance to the advocates of the different systems, and received some praise but much criticism. Most of this could be ignored, such as the comment, "You had no right to be so intelligent with the Uniterm system; it is meant to be used by persons of low intellect", or the person who wrote, "Subject headings are not meant to be so specific as those you used; that is why it performed so much better than it should have done". The most trenchant criticism came from Professor Swanson of the University of Chicago. (Ref. 4) His major point related to the use of search questions which were based on documents in the test collection. The validity of this point was somewhat lessened in that neither Swanson nor anyone else had been able to propose any other practical technique which would have overcome so effectively the problem that is associated with the determination of relevance, an aspect of evaluation testing which has still not been satisfactorily settled, in spite of being the subject of dozens of papers. The crude method used in Cranfield 1 was a reaction to the debacle of an earlier attempt to compare two systems. In this, two groups each indexed some 15,000 documents and carried out a number of searches in their own system. When they came together to consider the results, there was a total failure to agree on the relevance of the different sets of citations which each group had retrieved, and at the end of the second day of meetings, they were still arguing about the meaning of the first search question.

While Cranfield 1 was under way, two other investigations were made. The first was intended to show to what extent the Cranfield results were influenced by the artificialities of the test design. This was carried out on an operational facet classification system in an industrial organisation; the results and consequent failure analysis closely paralleled the results of Cranfield 1. The second was a test of the Metallurgical Index at Western Reserve University (Ref. 5), notable both for the novel and sophisticated approach to indexing and for the fact that searches were done using a GE225 computer. However, a single search on the 50,000 document collection was reputed to take eight hours, so the test was confined to a subset of 1,000 documents. This meant that it was a relatively trivial task for us at Cranfield to index these documents by a facet classification, and thereby provide the basis for a comparative evaluation. By the time of starting on this project, it was realised that the recall ratio was of limited interest unless the precision ratio was also known, so for the 104 questions used in the test, a check was made to identify all the documents which were relevant

to each question.

Two matters of particular interest came from this test. The semantic factoring technique could be described as a powerful system, making considerable use of roles and links, but it turned out that these devices were so powerful that it was difficult to use them in the search process without eliminating all retrieval.

The result was that the WRU system performed significantly less well than our facet catalogue. The second point arose from testing this facet catalogue. Originally an average of twelve entries had been made for each paper. After searching at this level, the number of entries was reduced to eight, then five, and then to three. The results of searches at these four levels showed experimentally for the first time a phenomenon which had originally been theoretically advanced by Robert Fairthorne, namely the inverse relationship of recall and precision. (Fig. 1)

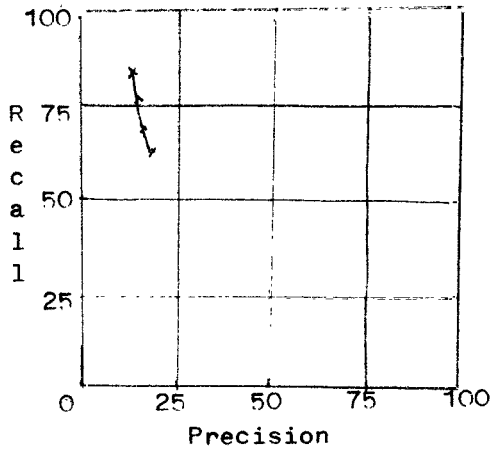

Figure 1

Whereas it had been the general view that there were fundamental differences in the various systems, the experience of working with four systems at the same time convinced us that this was not the case. Each index language was an amalgam of a few different devices, some of which were intended to improve the recall of relevant papers and others were intended to improve precision by preventing the recall of non-relevant papers. We knew that within a single system it was not possible to improve both the recall and the precision ratio simultaneously, but it was hypothesised that there would be some combination of recall and precision devices which would give optimum performance.

Two factors appeared to be critical. The first relates to the level of recognition of the contents of a document. In conventional indexing, it is normally an editorial or management decision as to the number of index terms or entry points which should, on average, be permitted. Within this overall decision, it is an intellectual decision by the indexer as to which aspects of the document should be recognised by assigning index terms. The intellectual aspect of this decision can be bypassed by agreeing to accept, say, all the significant words in the title or all the words in the abstract, in some designated parts of the text, or finally the complete text. Such variations we called 'exhaustivity'. A high level of exhaustivity represents a recall device, for it is obvious that unless a term has been assigned to a given document, the document cannot be retrieved by that term.

The second important factor related to what we called 'specificity'. An index term can be co-extensive with the subject it is describing or it can be, to a varying degree, more general or broader in meaning. For example, a thesaurus is a controlled language, which will probably contain some specific terms, but there will also be terms which are subsumed to a broader, controlled language term. In this respect, a high level of specificity is a precision device in that if, at the input stage, two or more concepts are grouped under a single term, it will be impossible to separate them at the search stage.

A great deal of effort had gone into setting up the four indexes of Cranfield 1. They had fulfilled their

objective, but it was clear that they could not be used for further development. Whereas Cranfield 1 had attempted to simulate an operational situation, with an emphasis on economic factors, we now required a laboratory type situation where, freed as far as possible from the contamination of operational variables, the performance of index languages could be considered in isolation. Taking the view that all index languages were amalgams of recall and precision devices, the objective of Cranfield 2 would be to measure the effect of each of these devices, alone or in any possible combination, on recall and precision. For this, a new collection with rigorous controls was required.

Experience had shown that a large collection was not essential, but it was vital that there should be a complete set of relevance decisions for every question against every document and, for this to be practical, the collection had to be limited in size. To obtain the document set, letters were sent to some 200 authors of recently published research papers, and they were asked to state, in the form of a question, the problem to which their paper was addressed, and to add supplementary questions that arose in the course of their research. They were then requested to indicate, on a scale of 1 to 5, the level of relevance to each question of the references which they had cited. The test collection of 1,400 documents was made up of these references. A group of six students spent three months screening the documents against 279 questions; those which they considered might be relevant were sent to the originator of the question for his final judgement.

The indexing of the documents was a multi-stage process. The indexer first recognised the concepts in the document; sometimes these could be expressed in a single word but more often were two or three words. A weighting in the range of 1 to 3 was assigned to indicate the relative importance of each concept within the document. Each single word occurring in the concepts was then listed and given the appropriate weighting. On average there were 33 single terms, with 14 having the top weighting and an additional 8 having the medium weighting. Finally the concepts were combined into themes.

Starting from the absolute basic level of single terms in the natural language of the documents, the intention of Cranfield 2 was to progress to more complex index languages by introducing various recall and precision devices. An example of how these languages were obtained is given in Figure 2.

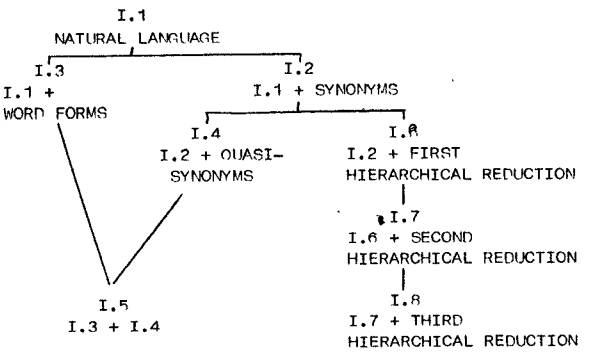

Figure 2

Starting from I.1, single terms in natural language, I.2 is I.1 plus synonyms, while I.3 is I.1 plus word forms. I.4 takes I.2 and includes quasi-synonyms (that is words which are sometimes but not always synonyms) while I.5 combines I.3 and I.4. Index languages I.5, I.6 and I.7 are formed by successive hierarchical reductions of the natural language terms, so that the 3,150 terms of language I.1 were reduced to 310 for language I.7. Similar principles were applied to obtain fifteen index languages based on concepts and six based on controlled language terms taken from the Thesaurus of the Engineers Joint Council.

A sample test had shown that there

would be serious clerical problems in carrying out coordinate level searches for 29 index languages, and for some weeks we flirted with the idea of using a computer. At that time there was no program which was remotely capable of doing what was required but fortunately a member of my staff, Michael Keen, came up with an ingenious idea which allowed us to simulate computer searching, albeit with considerable clerical effort. Essentially this involved making a separate index for each question by the preparation of a set of search sheets, from which clerical staff, in a single pass, could ascertain which documents would be retrieved at all levels of coordination, at the three levels of exhaustivity, at four levels of relevance and by four search rules, for each of the 29 index languages. When the totals of documents retrieved at the various coordination levels had been summed, it was possible to calculate the recall and precision figures (Fig.3)

SINGLE TERMS. NATURAL LANGUAGE .

Coordin-ation level	Documents Retrieved Rel.	Non-rel.	Recall Ratio	Precision Ratio
1	1,510	159,122	95.0%	0.9%
2	1,203	58,122	80.7%	2.2%
3	940	21,933	59.5%	4.1%
4	606	7359	38.1%	7.6%
5	314	2380	19.7%	11.6%
6	154	699	9.7%	18.0%
7	74	216	4.7%	25.5%
8	22	43	1.4%	33.6%
9	8	5	0.5%	61.5%
10	1	0	0.1%	100.0.

Figure 3

and thence to prepare performance plots as in Figure 4 which relates to three of the single term languages.

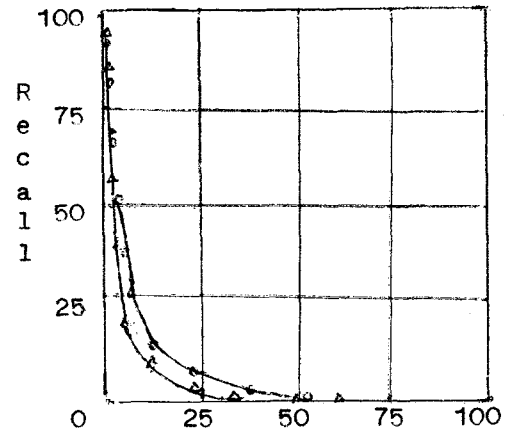

Figure 4

Such plots are not easy to interpret, so recourse was had to an amended version of a single measure, known as normalised recall, which had first been used by Professor Salton with the SMART system.

The order of merit by this measure for the 29 languages is shown in Figure 5. The original hypothesis had been that, starting from single natural language terms, the addition of recall and precision devices would inevitably improve the performance. It was the same hypothesis which has resulted in the production over the past thirty years of hundreds, if not thousands of thesauri on every possible subject, an activity which shows few signs of abating in spite of the fact that there is now abundant proof that the hypothesis was wrong. Neither we nor anybody else had considered it as remotely possible that an index language based on single terms in the natural language of the documents would be so effective that the performance could only be improved by confounding word forms or true synonyms. In Figure 5 it will be noted that, in the main, single term languages occupy the top positions, the controlled term

Norm. Recall	INDEX LANGUAGE
65.82	S.T. Word forms
65.23	Synonyms
65.00	Natural language
64.47	S.T. Synonyms. Word forms. Quasi-synonyms
64.41	S.T. Hierarchy 1st stage
64.05	S.T. Hierarchy 2nd Stage
63.05	S.T. Synonyms. Quasi-Synonyms
63.05	Concepts. Hierarchy and alphabetical selection
62.88	Concepts. Alphabetical 2nd stage selection
61.76	C.T. Basic terms
61.76	C.T. Narrower terms
61.17	S.T. Hierarchy 3rd stage
60.11	C.T. Broader terms
59.70	C.T. Related terms
59.58	C.T. Narrower and broader terms
59.17	C.T. Narrower, broader and Related terms
57.41	Concepts. Complete combination
57.11	Concepts. 1st stage selection
55.88	Concepts. Complete species and superordinate
55.76	Concepts. Hierarchical Selection
55.41	Concepts. Complete species
55.05	Concepts. Selected species and superordinate
53.88	Concepts. Selected coordinate and collateral
53.52	Concepts. Selected species
52.47	Concepts. Complete collateral
52.05	Concepts. superordinate
51.82	Concepts. Selected coordinate
47.41	Concepts. Synonyms
44.64	Concepts. Natural language

Figure 5

languages occupy the middle ground while the concept languages generally take the lowest positions.
While this table demonstrates the effect of specificity, the effect of exhaustivity required further analysis.

Level of exhaustivity	Average No. of terms	Norm. Recall
Titles	7	59.76
Indexing 1	14	62.88
Indexing 2	22	63.57
Indexing 3	33	65.00
Abstracts	60	60.94

Figure 6

Figure 6 shows the results when the language is held constant but exhaustivity is increased from an average of seven terms, as in a title, to the average of sixty terms occurring in an abstract. There is a steady improvement to the optimum indexing level of 33 terms, followed by a sharp decline, implying that an abstract is over-exhaustive, and suggesting that full-text searching would further degrade performance. Such a statement has to be qualified in that this optimum level of exhaustivity applies only to the environment of the test. Subsequent work has shown that it is probably representative of databases in science and technology; it would not necessarily apply, for example, to legal databases.

For myself, the results of Cranfield 2 appeared far more significant than the earlier test, but they did not arouse so much general comment, partly, I think, because the results were so unexpected as to appear unbelievable and also because the nature of the test removed it so far from normal experience. Swanson returned to the attack, arguing again that the method of obtaining the relevance decision had influenced the results. In doing so, he ignored two separate investigations, by Professor Salton (Ref. 7) and at Cranfield (Ref. 8) where completely new sets of relevance decisions were used without making any change in the comparative results.

The Cranfield 2 test collection has been used by several other research groups, in particular Karen Sparck-Jones at Cambridge (Ref. 9) and Salton at Cornell (Ref. 10). Salton, in his early studies reported, in line with Cranfield, "that phrase languages are not superior to single terms as indexing devices, that synonym dictionaries improve performance but that other dictionary types, such as hierarchies, are not as effective as expected".

Whereas at the start of Cranfield 2 it was thought that there were a number of potentially useful devices to improve either recall or precision, at the end we came to the view that exhaustivity and specificity are, to quote Keen, "the only general principles that exist to understand the fundamentals of retrieval performance". It has been amply demonstrated that in any operational system there is an inverse relationship between recall and precision. Similarly with exhaustivity and specificity. As specificity increases, one moves from low precision and high recall to high precision and low recall. Conversely, an increase in exhaustivity results in a move from low recall and high precision to high recall and low precision.

In experimental testing, much emphasis is placed on the performance measures of recall and precision but I would advocate extreme caution in their interpretation. In operational situations there is strong evidence to support the view that in the majority of searches a high level of recall is not required. In forty years experience as an industrial or academic librarian, I can recollect only four occasions when the end-user was trying to obtain 100% recall; the vast majority of users required a few relevant papers. This is supported by an investigation by Lantz (Ref. 11) with some 2,000 scientists and engineers. He found that the number of relevant citations retrieved in online searches far outnumbered those which were subsequently used. This is shown in Figure 7 where it can be seen that even though there were a hundred relevant citations, engineers never used more than eight papers.

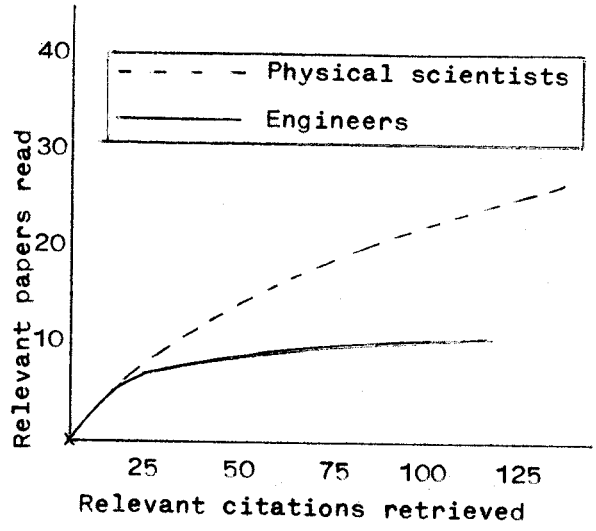

Figure 7

In regard to precision, experimentally it would be considered a significant improvement if, at a given recall ratio, the precision ratio could be raised from 30% to 40%, yet for an end-user in an actual search the improvement would hardly be noticeable. As I argued earlier, the ultimate measure in an operational system must include cost. Assuming, as appears reasonable, that the objective of an IR system is to retrieve relevant citations without retrieving non-relevant citations, and at the lowest possible cost, a measure taking this into account can be expressed as

$$C_r = \frac{C_s + (F \times D_n)}{D_r}$$

where C_s is the cost of a search, D_n and D_r the number of non-relevant and relevant citations retrieved and F is a fine. Applying this measure to the evaluation of MEDLARS in 1968 and putting the fine at 20 cents, the cost of retrieving a relevant citation was shown to be $1.58.

Five centres took part in the evaluation and their performance in terms of recall and precision varied as shown:

	Recall Ratio	Precision Ratio
Centre A	69.2%	40.7%
Centre B	64.6%	43.2%
Centre C	57.9%	50.9%
Centre D	55.5%	55.6%
Centre E	43.3%	57.2%

With the inverse relationship of recall and precision, one cannot say that the performance of one Centre is better than another, but applying the cost measure shows a significant difference in C_r.

	C_r
Centre A	$1.42
Centre B	$1.50
Centre C	$1.50
Centre D	$1.64
Centre C	$2.13

Even this leaves the matter open to question. As previously argued, most users are satisfied with far less than total recall. Assume that the users of Medlars considered thirty relevant papers to be sufficient, and therefore all relevant papers above this number would be unwanted. Applying a fine of 20 cents for all such unwanted and non-relevant retrievals, the cost of a relevant and wanted citation shows a complete reversal from the figures above.

	C_r
Centre A	$5.30
Centre B	$5.07
Centre C	$4.60
Centre D	$4.53
Centre E	$4.13

These figures have no real meaning but I produce them to illustrate the extreme care necessary in interpreting test results.

Recently I read an editorial eulogising the brave new world of information retrieval which the writer feels is opening before us. All that is needed, he wrote, is that research should go back to basics, and that successful practices, such as classification "should be generalised and improved". I would like to echo this facile optimism, but experience of the past thirty years leads me to the conclusion that there is, in information retrieval, a barrier which it is very difficult, and perhaps impossible, to breach. A senior researcher recently asked me, "How can one explain the apparent contradiction between the fact that language understanding in human beings depends on having the terms available in context, whereas in a retrieval setting the extra context appears to hurt more often than not?" It is not easy to find an explanation, but possibly it may be related to the seemingly random nature of information retrieval. For example, if two competent individuals prepare a thesaurus on a given subject, only 60% of the terms may be in common. If they index a document, only 30% of the terms may be assigned by both indexers. If they carry out a given search, only 40% of the citations will be retrieved in common, and if they decide the relevance of a given set of documents to a given question, there may be only 60% agreement (Refs. 8, 12,13,14). If there is any hope of significant improvement, it lies, I believe, in doing away with Boolean searching. This technique was propagated by Taube some 45 years ago to ameliorate the problem of post-coordinate searching, which was then done by the visual comparison of lists of numbers, a very slow process and prone to error. Developments in this activity came slowly, with punched cards, sorted at first by hand and then by machine, followed by trials with early computers, and gradually the high-speed computers of today. At no stage in this development was the improvement of sufficient significance to justify a rethink of what was being done, but it is ironic that we continue to use such a search technique when

computers are now doing in a fraction of a second what would have taken hours to do when the Boolean search technique was devised.

Practically all researchers in the field have investigated systems giving a ranked output, yet all major publicly available databases require Boolean searching. There is strong evidence to suggest that, compared to a ranked output, Boolean searching degrades performance, but more importantly, it is user-hostile, and as such it has been the major source of discouragement for end-users to carry out their own searches. Summit (Ref. 15) reported that, in 1989, 88% of searches on Dialog were carried out by intermediaries; this is not surprising when Miller (Ref. 16) can report that of searches by end-users, 46% were complete failures in that no citations were retrieved.

There is no valid argument which can be advanced in favour of Boolean searching, and in my view the adoption of any form of ranked output would be beneficial. However, there is need for further investigation to optimise such a system in the important areas of performance, of cost and, above all, of convenience to the users. It is still an open question as to whether, compared to a simple coordination level search, the use of a complex system using such devices, for example, as term position, inverse collection frequency or iteration, will give sufficient improvement in performance to compensate for possible additional costs and inconvenience to the user. It appears unlikely that such work will be carried out by the commercial operators of databases and the field is open to those who can experiment within their own organisation.

* * * * * * * * * * *

REFERENCES

1. CLEVERDON, C.W. and THORNE, R.G.
   'An Experiment with the Uniterm System'
   RAE Library Memo 7, 1954
2. 'The Truth, The Whole Truth . . .'
   American Documentation. 6, p.58
3. CLEVERDON, C.W.
   'Comparative Efficiency of Indexing Systems', 2 vols.
   Cranfield. 1960 - 62
4. SWANSON, D.R.
   'Evidence Underlying the Cranfield Results'
   Library Quarterly, 35, pp 1 - 20
5. AITCHISON, J. and CLEVERDON, C.W.
   'Test of the Index of Metallurgical Literature'
   Cranfield, 1963
6. CLEVERDON, C.W. et al
   'Factors Determining Performance of Index Systems'. 2 vols.
   Cranfield, 1968
7. LESK, M. and SALTON, G.
   'Relevance Assessments and Retrieval System Evaluation'
   I.R.S. Cornell. 1968
8. CLEVERDON, C.W.
   'Effects of Variations in Relevance Assessments'
   Cranfield Library Report 3, 1968
9. SPARCK-JONES, K. and BATES, R.G.
   'Research in Automatic Indexing'
   Computer Lab., Cambridge, 1977
10. SALTON, G.
    'The SMART Retrieval System : Experiments in Automatic Document Processing'
    Prentice-Hall, 1971.
11. LANTZ, B.E.
    'The Relationship Between Documents and Relevant References'
    Jnl. Documentation 37, pp.134,145
12. Private Communication
13. BORKO, H.
    'Inter-Indexer Consistency'
    Cranfield Conf. 1979
14. CLEVERDON, C.W.
    'Comparative Evaluation of Searching by Controlled and Natural Language in a NASA Database'
    European Space Agency Report 1/432. 1977
15. SUMMIT, R.K.
    'In Search of the Elusive End-User'
    Online Review. 13, pp. 185 - 91
16. MITCHELL, N. et al.
    'End-User Searching in a Medical School Library
    Med.Ref.Services Quart. 7, pp. 1-33