14

An experiment: search strategy variations in SDI profiles

Lynn Evans

14.1 Introduction

This research project was one of a number concerned with mechanized information retrieval carried out at INSPEC and supported by grants from, originally, the Department of Education and Science's Office for Scientific and Technical Information (OSTI) and, later, the British Library Research and Development Department (BLRDD).

The first project, in the period 1967–69, had investigated the performance, economics and acceptability to users of a computerized SDI service in electronics[1]. As a result of this study, in 1970 INSPEC offered publicly an SDI service in electronics.

Another research project[2], overlapping the SDI investigation, had established the comparative effectiveness of natural language as a medium for mechanized searching so that, from 1971, all documents input to the INSPEC database were assigned free index terms. These are significant words and phrases from the title, abstract and text of the document, selected by the information scientists as representing the meaningful concepts treated in the document. In April 1971 the cost-recovery service in electronics was replaced by a commercial SDI service covering the whole of the INSPEC subject areas with the important addition that free indexing was now available as a search medium.

In the original INSPEC SDI investigation an EJC-type thesaurus was used for document indexing and profile generation. The complete thesaurus was never produced in printed form and so was not available to users for compiling their own profiles. The introduction of free indexing removed this barrier and prompted another study[3] into the optimum degree of user participation in SDI profile generation. This concluded that users prepared to familiarize themselves with the 'mechanics' of the system compiled their own best profiles as measured by precision (although indirect evidence indicated lower recall values than in the profiles compiled by INSPEC staff). However subscribers were reluctant to become too involved in that 45 per cent of them chose to delegate profile compilation to INSPEC staff and only 14 per cent opted to completely manage their own profiles. The main reason offered in explanation was that users found the whole procedure of getting a

profile started too complicated and having to read a detailed User Manual, including grappling with the intricacies of nested boolean logic, was a positive discouragement.

At that time INSPEC profiles were almost exclusively of the boolean type. The general assumption was that full advantage of the machine facility should be taken and profiles compiled with complex logic to match the user's statement of information requirements. However if the initial user–system interaction was to be eased then simpler search strategies was one area where this might be achieved. There was some evidence of a non-quantitative nature which suggested that, even in profiles incorporating sophisticated logic, many of the relevant matches obtained were derived from comparatively simple parts of the logic statement.

In addition to doubts about the need for sophisticated logic, work at Cranfield in the field of precision engineering[4] showed that the case for any boolean logic was not proved and that a straightforward co-ordination of all the profile terms might be equally or even more efficient.

Quite separately it was also felt that the issue of simpler search strategies was to become important in the development of online retrospective search systems. For these to be operated directly by the people requiring the information rather than by intermediary information scientists and librarians then the man–machine interaction needed to be much simpler than was currently the case. In the event, some 5 years later, online searching of bibliographic databases is still very much the prerogative of intermediaries and is likely to remain so until the development of truly interactive retrieval software.

This then was the rationale leading to the particular experiment described and analysed in this paper, viz. search strategy variations in SDI profiles[5]. The overall aim was to develop the most cost-effective search strategy by studying the cost, retrieval performance, and ease of use of a number of different search strategy types. In the general framework of information retrieval experiment this work was most like a laboratory test but perhaps lacking the rigid control necessary for a 'true' laboratory experiment. Although carried out in the environment of an operational system it was in no way an investigation of the operational system. It is probably most accurately categorized as a developmental project, i.e. one pursued with as much experimental rigour as possible but not at the expense of losing touch with 'real-world' conditions.

14.2 Experiment

Very broadly a conventional evaluation in information retrieval requires: (1) a collection of documents with various attributes (titles, abstracts, assigned index headings, etc.); (2) a set of queries whose subject area is covered by the document collection; and (3) knowledge of which documents in the collection are relevant/non-relevant to particular queries. In addition, in this experiment, the important considerations were the search strategies, profile compilation procedures, and search software.

Documents

The documents used were not special in any way and were obtained on a weekly basis from the INSPEC current file. The most important consideration was that the document collection should be typical of the INSPEC database. Eight weeks' documents were taken in two groups of 4 weeks each separated by a time gap. Four consecutive weeks was considered a reasonably adequate span in that the cycle of most journals would thereby be covered. The reason for having two separate 4-weekly groups was to allow for profile performance analysis and modifications in between the first 4 and last 4 weekly SDI runs. In all over the 8 runs more than 20 000 documents were matched against the profile file, the individual weekly totals being 3095, 2640, 2582, 2194, 2542, 2781, 2123, and 2860.

As with many aspects of this experiment deciding the size of the document collection was empirical. Later, when the pattern of retrieval performance was found to be similar from week to week it was not considered worthwhile completing the performance calculations for all 8 runs so that 20 000 documents would seem to have been more than an adequate number and something like 5000 would have been quite acceptable. Since then of course the question of document collection size has been thoroughly considered as part of the design of an 'ideal' information retrieval test collection[6–7].

Queries

It was considered important that the profiles should be based on real rather than artificial questions and that they should operate in a 'near-real' environment. This was because one of the main objectives was to detect differences between the retrieval performances of the search strategies which showed through despite the hazards associated with real user assessments.

Experience in the INSPEC SDI investigation had highlighted the fact that real users' relevance assessments were often not entirely based on subject content but might be influenced by such factors as the language of the original document, its source, the author's reputation, etc. It was also not unknown for slight changes of interest not to be notified and only become apparent on investigation of poor profile performance.

Despite these known irritations which only tend to cause confusion the feeling was that they are an inevitable feature of operational systems and should not be ignored by the use of artificial queries. On the other hand it was not thought politic to use paying customers of INSPEC's commercial SDI service because of the danger of their alienation when asked to assess documents over and above those retrieved by their optimum profiles.

The compromise decided was to recruit a user group from UK university research workers, the arrangement being that for their agreement to provide subject interest statements and to assess the relevance of document outputs they would gain, at no cost, experience of a mechanized current awareness service. In selecting the university research workers the latest issue of the Department of Education and Science's annual publication *Scientific Research in British Universities and Colleges* was used as a random source of names, addresses and subject interests.

Of the 100 people originally approached (40 Physics, 40

Electrical/Electronics, 20 Computers/Control) the 55 who actually supplied statements of their subject interests comprised

23 from Physics Departments,
27 from Electrical/Electronics Departments, and
5 from Computer/Control Departments.

Given the inevitable delay between their original undertaking to participate in the experiment and their receipt of the first notifications for assessment, the response from users was quite gratifying. Over the 8 runs relevance assessments were received from, on average, 82 per cent of participants, the returns for the individual runs being 84, 89, 86, 87, 76, 78, 78 and 76 per cent.

As with the number of documents the decision to use about 50 users was 'gut-feeling' rather than statistically reasoned. The question of what is an acceptable number of queries cannot of course be considered in isolation and is intrinsically bound up with the number of documents against which the queries are matched. If Q represents the number of queries and D the number of documents it would be tempting to speculate that there is an optimum value for the product QD with minimum acceptable values for Q and D. However it is not only a matter of the numbers of queries and documents but also how many documents are relevant to particular queries. The fewer relevant documents there are the less confidence there can be in the particular recall or precision ratio so there is great doubt associated with, say, a recall figure of 1/1 which could so easily have been 0/1. These are matters which can only be completely controlled in an artificial situation and the statistical bases of relevance assessment have been considered in detail more recently[8].

Search strategies

The strategies to be compared should ideally represent different degrees of intellectual effort when compiling profiles and, also, should require different degrees of sophistication of computer facility. The strategies selected for comparison were all variations of two basic types, viz. those consisting of a single list of terms, and those containing groups of terms, where the groups represent subject concepts in the original query. The list of search strategy types was:

(1) *Co-ordinate matching of terms without weights* (CT)
The output is ranked in order of term co-ordination level, i.e. in order of the number of matching profile and document terms.

(2) *Term-weight cumulation* (TWC)
The profile terms are assigned weights in accordance with their relative importance to the query. The weights of all matching terms are 'summed' to produce a 'document score'. The output is ranked in order of document scores.

(3) *Co-ordinate matching of terms with weights* (CTW)
The profile terms are weighted as in (2). The output is ranked, first in order of term co-ordination level (i.e. number of matching terms), and then by sum of all matching-term weights.

(4) *Co-ordinate matching of groups of terms without weights* (CG)
The profile search terms are divided into groups representing the various concepts in the query. The output is ranked, first in order of group co-ordination level (i.e. number of matching terms where each is from a different profile group), and then in order of total number of matching terms.

(5) *Group-weight cumulation* (GWC)
The profile term-groups of (4) are weighted according to the relative importance of the groups (concepts) to the query. The weights of all matching groups are summed to produce a document score. The output is ranked in order of document scores.

(6) *Group- /term-weight cumulation* (GTWC)
The term-groups of (4) are weighted according to their importance to the query and the individual terms are weighted according to their importance within their group. The output is ranked, first by sum of matching-group weights, second by sum of highest-weighted matching terms from each group, and third by sum of all matching-term weights.

(7) *Co-ordinate matching of groups of terms with weights* (CGW)
The profile term-groups of (4) and the individual terms are weighted. The output is ranked, first in order of group co-ordination level, second in order of sum of matching-group weights, and third in order of sum of matching-term weights.

(8) *Boolean logic* (B)
The profile term-groups of (4) are governed by boolean statements which must be satisfied before any output is obtained. The output is in document number order, i.e. unranked.

(9) *Boolean logic with weights* (BW)
The profile term-groups of (4) are governed by boolean statements which must be satisfied before any output is obtained. After the boolean equations are satisfied the ranking of output may be based on group- and/or term-weight cumulation procedures. In our experiment only term weights were used.

Basically procedures (1), (2) and (3) involve search profiles comprising a single list of terms (weighted or unweighted) while procedures (4)–(9) inclusive involve profiles comprising groups of terms (in which groups and/or terms may be weighted or unweighted).

In the weighted profile versions two types of weights were used. In procedures (2), (3), (5) and (9) above, the weights were subjectively assigned by the compiler, while in procedures (6) and (7) 'powers of 2' weighting was used[9]. In 'powers of 2' weighting the weights are assigned routinely once the order of importance of individual terms and/or term groups in the search profile has been intellectually decided. This ordering was again decided by the compiler.

Profiles incorporating automatically-assigned weights were not included in the study mainly because the necessary statistics (term frequencies, etc.) were not immediately available. They were to become available subsequently from another INSPEC research project.

In addition to the 9 search strategies listed above, as the project proceeded it was decided to include a further two types:

(10) *Co-ordinate matching of restricted list of terms with weights* (CRTW)
The original list of search terms chosen to represent a particular user's interests in strategies (1)–(9) above was not restricted in any way and comprised, on average, more than 40 terms. It had been felt, mainly as a result of experience with INSPEC's commercial SDI service where similar-sized profiles pertained, that there may be a significant number of unnecessary search terms in profiles, viz. those search terms which 'hit' too infrequently to be useful and those search terms which 'hit' too often to be selective.
A further argument was that although the search profiles used in online retrospective searching contain, in general, far fewer terms than those in current-awareness batch SDI systems, their retrieval performance does not seem to be noticeably inferior.
For these reasons it was thought worthwhile to include for comparison a profile version in which the number of terms was restricted to 20 irrespective of the number in the original list used for search strategies (1)–(9) above. The 20 terms were selected subjectively in order of importance from those used in strategies (1)–(9).

(11) *Controlled-language boolean strategy* (CLB)
In the main experiment the medium used for matching profiles and documents was natural language, viz. the free-index terms assigned to all items added to the INSPEC database. In general the free-index terms are words or phrases occurring in the original document which represent the meaningful concepts treated in the document. In addition to the free indexing, the subject content of all items in the INSPEC database is indicated by two other elements: (i) classification codes, which govern the location of the item in the published abstracts journals, and (ii) controlled subject headings, which appear in the six-monthly indexes to the abstracts and are used mainly for manual retrospective searching. The classification codes and controlled subject headings can of course also be used in machine searching and to this end boolean-type profiles using only classification codes and/or subject headings were prepared.
Originally the main purpose of these controlled-language boolean strategies had been to act as 'back-up' profiles to retrieve relevant documents which the other versions (based on the free-index terms) might miss because of inadequate free indexing or profiling. Knowledge of these additional relevant items would of course mean that the recall performance figures would be that much nearer to being measures of the true rather than the relative recall. However the data available also allowed a direct comparison of the retrieval performance of controlled-language against free-language boolean profiles. Brief details of this secondary experiment are given in section 14.3 (p. 309).

Profile compilation

All the tasks associated with translating the original user statements into profiles incorporating the various search strategies were carried out by one person under controlled conditions. These are described now.

Standard tasks

To measure the intellectual effort involved in compiling profiles incorporating the different search strategies, the profile compilation procedure was divided into a number of standard tasks and times recorded for completion of each task.

It was considered important that, for meaningful comparison of the search strategies, all the profile versions compiled from a particular user statement should use the same basic set of search terms. The first standard task was therefore to produce a list of search terms. The complete list of seven standard tasks, T1–T7 (with corresponding task completion times t_1–t_7) was established as follows:

T1—from user statement prepare list of search terms using various aids such as INSPEC thesaurus, dictionaries, known relevant documents, etc.
T2—arrange the terms of T1 into groups representing the concepts in the original query.
T3—assign boolean equations to govern the groups of T2.
T4—assign weights subjectively to the individual terms of T1.
T5—arrange groups of T2 in order of their importance to the original query.
T6—assign weights subjectively to the groups of T2.
T7—arrange the terms in the groups of T2 in order of their importance within the group.

Figure 14.1 shows the relationship between the various standard tasks and how they lead to the profiles incorporating the different search strategies. The compilation times are then calculated by adding the appropriate task completion times, e.g. t_1 for strategy CT, $t_1 + t_4$ for TWC, and so on.

Strictly there is no compilation time for strategy CRTW since the profile terms were obtained in rather an indirect way. However it seems reasonable to assume that, starting from a user statement of interests, the time taken to produce a list of the 20 most important search terms (effectively strategy CRTW) would not differ very much from that taken to prepare a complete list of all the search terms likely to be useful (strategy CT). On balance the former task could well involve less time. As an approximation we can take both compilation times as being equal to t_1.

The compilation of profiles incorporating strategy CLB was a completely separate and self-contained exercise.

Clearly, with the standard tasks, there is a question as to how interdependent they might be, e.g. in doing task T1 (preparing list of search terms) does one immediately in one's mind start grouping them into concepts, i.e. task T2, and even consider possible boolean equations, i.e. task T3. Also it might reasonably be argued that the 'natural' thing to do starting with the user statement is to isolate the concepts first and then expand them to produce the search terms.

With these considerations in mind the user statements were first divided at random into two groups A and B. In group A all the standard tasks for a particular user statement were completed consecutively 'at one sitting' whereas for group B each task was completed in isolation, i.e. separated in time from the other tasks, so that the memory of doing one task had largely disappeared before the next one was started.

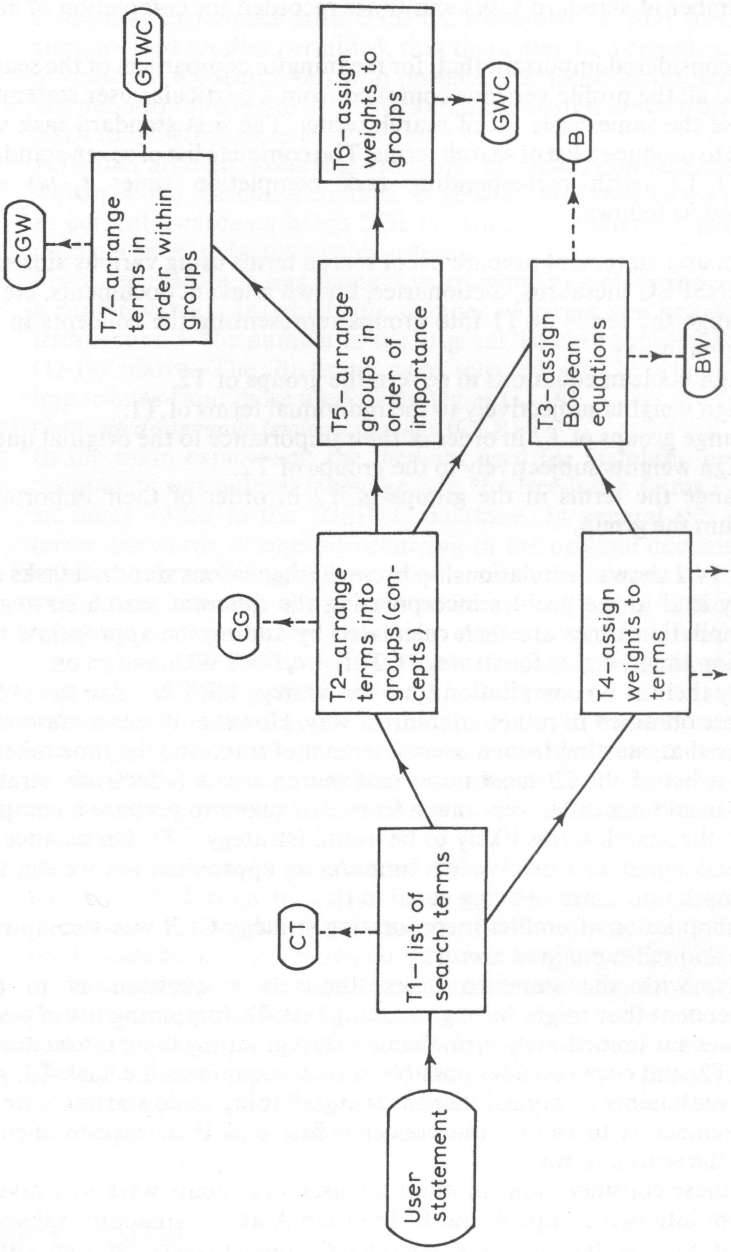

Figure 14.1. Standard tasks in profile compilation procedure

If the tasks are interdependent then, on average, the sum of the task completion times for group A user statements should be less than for group B statements; and, the greater the interdependence the greater the difference in the total task times. It is assumed here that with a total of more than 50 user statements involved there is unlikely to be any bias between groups A and B caused by difference in subject matter or complexity.

The average task completion times for groups A and B were as given in *Table 14.1*.

TABLE 14.1

Users	*Average standard task completion times (min)*							*Average totals*
	t_1	t_2	t_3	t_4	t_5	t_6	t_7	
Group A	34	13	11	11	7	9	3	88
Group B	39	15	15	11	10	9	3	102

Not surprisingly the total of the average task completion times is lower for group A (88 min) than for group B (102 min), thus indicating some degree of interdependence of the tasks. No doubt also part of the difference is due to the effect of having to 're-understand' the user statement each time in group B.

Alternatively it might be argued that groups A and B are not exactly similar or are not quite large enough to discount the effects of individual statement characteristics on the overall average task completion times. There is some support for this in that the average number of search terms is 44 for group A profiles and 50 for group B profiles.

However the average standard task completion times for groups A and B are similar enough to encourage the conclusion that the degree of interdependence of the standard tasks is not so great as to invalidate the original division of the profile compilation process into the 7 standard tasks.

Compilation times

Combining the data for groups A and B the overall average standard task completion times (min) were: $t_1=36$, $t_2=14$, $t_3=13$, $t_4=11$, $t_5=9$, $t_6=9$, and $t_7=3$ min, which gave the average compilation times for the different search strategies as in *Table 14.2*.

TABLE 14.2. Search strategy compilation times (average)

Search strategy	*Average compilation times*	*(min)*
CT	t_1	36
TWC	t_1+t_4	47
CTW	t_1+t_4	47
CG	t_1+t_2	50
GTWC	$t_1+t_2+t_5+t_7$	62
CGW	$t_1+t_2+t_5+t_7$	62
B	$t_1+t_2+t_3$	63
GWC	$t_1+t_2+t_5+t_6$	68
BW	$t_1+t_2+t_3+t_4$	74

Table 14.2 shows that, in terms of information scientist effort, the simplest strategy, CT, takes almost exactly half as long to compile as the most complex strategy, BW.

Modification times

As mentioned on p. 287 above the profiles were analysed and (perhaps) modified just once, viz. after completion of the first group of 4 SDI runs but before starting the second group.

The initial analysis procedure adopted was standard for all profiles irrespective of whether any relevance assessments had been received from the users. Ten minutes were taken for an examination of the profile performance after which time a decision was taken as to whether or not any basic modifications were necessary. Twenty-two users' profiles were in fact amended. It is emphasized that the time taken for any particular modification is assigned in full to all the search strategy variations incorporating that modification, e.g. if 20 min are spent on amending the boolean equations then this time is allocated to both strategies B and BW.

Averaging the modification time data (including all profiles whether modified or not) and adding the 10 min initial analysis time the average strategy modification times obtained were: CT = 13, CG = 13, CTW = 14, CGW = 14, TWC = 15, GWC = 15, GTWC = 15, B = 21, and BW = 22 min.

Discussion

Before leaving the profile compilation procedure it may be useful to discuss the standard tasks in more detail—in particular to consider some of the conflicts that occurred in trying to achieve a balance between experimental rigour and what common sense indicated should be done in a real situation.

It has already been stated that for a valid comparison of search strategies it seemed essential that, for a particular user statement, the same basic set of search terms should be used. In fact occasions arose when this was contrary to the needs of particular strategies, e.g. in the use of negative weights, NOT logic, and WITHIN logic, which facilities do not feature sensibly in the co-ordination strategies CT, CG, CTW, CGW and CRTW. Examples of the use of these facilities are detailed in the original report and the extent to which they were used is indicated by the fact that, of the 55 statements received, negative weights were included for 10 users, NOT logic was included for 8 users, and WITHIN logic was included for 2 users.

Another general problem occurs when the original user statement really covers more than one basic subject interest or question. With boolean strategies, if nesting or sublogic facilities were available, there would be no problem but with co-ordination strategies it seems nonsensical to mix search terms from what are essentially different questions. In those cases where obviously more than one subject interest was involved the user statement was divided and treated as 2 (and once 3) completely separate questions. It is now felt that this should have been done for more of the user statements than was in fact the case, viz. 6 users.

Some specific problems encountered when executing the individual standard tasks were:

(1) *T1—List of search terms*

The initial intention was that, excepting perhaps chemical compounds (e.g. gallium arsenide), the search terms should be strictly singlets. It was felt that any degree of pre-coordination would positively bias against some strategies particularly CT (plain co-ordination of terms). However even with simple CT there are difficulties, e.g. the free-index phrase 'field effect transistor' may also appear as the abbreviation 'FET' and of course both versions have to be catered for in searching. If the singlets-only rule is applied in the search profile then matching on the former produces a co-ordination level of 3 but matching on 'FET' is at a level of 1 only. This anomaly could be avoided by including the term 'FET' in the profile three times but this would be merely simulating weighting techniques; to do so would invalidate the comparison with strategy TWC (weighted list of terms). Other similar examples encountered were STEM/scanning transmission electron microscope, IMPATT/impact avalanche transit time, and LEED/low energy electron diffraction.

Some 65 per cent of the profiles included one or more non-singlet terms although, as a percentage of total search terms, non-singlets amounted to less than 5 per cent.

On the other hand some pre-coordination of terms might positively favour some search strategies. For example, for the CT strategy, the concept 'digital circuit*' would naturally be treated as two terms 'digital' and 'circuit*'. However for a boolean strategy it might be considered safer to search on the term 'digital circuit*' rather than 'digital' AND 'circuit*' since experience has shown that the latter usually throws up a large number of false drops. Of course searching on 'digital circuit*' will fail to match phrases like 'digital logic circuit*'. In an operational system using a boolean strategy the final decision would probably rest entirely on which performance measure the user was more interested—recall or precision.

Another difficulty encountered in preparing the basic list of search terms for a particular user statement was the problem of what to do with nebulous terms like 'measur*', 'propert*', 'design*', 'observ*', etc. It was fairly certain that they could do no harm when used in a weighted list of terms (and might even improve the uniqueness of the ranked output) but their usefulness in a boolean search is not clear and as likely as not to be damaging.

In the interests of retaining the same set of terms in all the search strategy variations for a particular query, some compromise had to be accepted occasionally in the final choice of terms used.

(2) *T2—Arrange terms into groups (concepts)*

In practice it was found that a large degree of latitude was possible in dividing the terms into concepts. At one extreme, for a user interested in 'high power gas and liquid lasers', it might be argued that there are only two basic concepts involved, viz. a device (laser) and a characteristic (power). Alternatively it could be said that there are five separable concepts, viz. high, power, laser, gas, and liquid.

The general policy pursued was to divide into as many concepts as possible. In fact the average number of concepts per user statement was 15, ranging from a minimum of 4 to a maximum of 25.

(3) *T4—Assign weights to terms*
In assigning weights subjectively to individual terms, values in the range 1–10 were used but where a single concept seemed paramount to the query the scale was extended to cover weights of 1–20. The facility for utilizing a wider range of weights was available (−999 to +999) and in an operational system much higher positive and negative weights could be used for ranking highly or excluding altogether items in certain journals or languages, etc.
An interesting characteristic of assigning weights subjectively to individual terms is that almost invariably one finds oneself separating the terms into groups (task T2), mentally ordering the groups into their relative importance (task T5), and setting imaginary threshold weights in a manner very similar to constructing a logical search. It is not suggested that one necessarily has to formally perform tasks T2 and T5 in order to assign individual weights to a single list of terms; only that it is difficult to proceed directly from task T1 to task T4 without thinking in terms of tasks T2 and T5, and even simulating T3 (assigning boolean equations).
A reservation concerning subjective weighting of terms is its possible lack of appeal to individual users interested in compiling their own profiles. No doubt some individuals would delight in the facility; others might be quickly frustrated by the problems of unanticipated homonyms, etc.

(4) *T5—Arrange groups in order of importance*
Normally it is a fairly quick task to arrange subjectively the groups of terms in order of their importance to the query. When 'powers of 2' weighting is being used this task becomes quite crucial to strategy GTWC. The feeling persisted that 'powers of 2' weighting would be better suited to controlled- rather than free-language searching, i.e. it would operate better with greater pre-coordination of terms. If this is so then strategy GTWC may have suffered somewhat from the approach advocated for task T2 which was to divide into as many concepts as possible.

(5) *T6—Assign weights to groups*
The same general procedure was followed in assigning weights subjectively to the groups of terms as was described above in (3) for the individual terms.

Relevance assessments

An important factor when considering the mechanics of the experiment was the role of the user group. It seemed desirable that they should operate in a 'near-real' situation but at the same time it was necessary to have as many documents assessed as possible to ensure the validity of the recall figures.

The users were not, and did not need to be, aware that a number of profile versions (representing the different search strategies) had been compiled from each statement of interests. On a particular search run the separate outputs from all the profile versions prepared from the user's original statement were merged to produce a single set of notifications without duplicates in random (document number) order. This was not only the most convenient procedure for the user but was also methodologically necessary in

that it ensured that the relevance judgements were completely independent of the search strategy and the position of the document in any ranked output. The only question concerned the number of items from the different search strategy outputs that should be merged in the first place.

The profiles in INSPEC's commercial SDI service, operating on subject interests similar to those of our experimental user group and with the same document collections, were at that time producing an average of 12–15 notifications per profile per week. With this figure as a guide it was decided that, for merging, the full output from the (optimum) boolean strategies should be taken with at least the top 25 items from each of the ranked-output strategies. Allowing for duplicates it was anticipated that the merged output would comprise at least 50 notifications per user per run. In those subject areas known to be more productive the full boolean output and the top 30, or even 40, items from the ranked-output strategies were merged. In fact over the total 8 runs the average weekly number of notifications sent to each member of the user group for assessment was 59.

Figure 14.2 illustrates in broad outline the operation to the point where the 'single set of notifications without duplicates' has been produced for despatching to the user for relevance assessments. The actual format of the notifications (6 in × 4 in cards) followed that used in the commercial INSPEC SDI service. They included the main bibliographic information (title, author, affiliation, source reference) plus all the free indexing terms and the main-entry classification codes. The user also received a summary card of the hit document numbers on which he indicated the relevance of each document notified.

In making his relevance assessments the user was asked to apply a three-category relevance code[1] as follows:

1—highly relevant documents;
2—partly relevant documents;
X—non-relevant documents.

To avoid misleading value judgements, the user was also requested to base his assessment purely on the subject matter and to ignore such things as the language of the original document, the quality of the journal in which it appeared, etc.

This three-category code was deliberately chosen for its relative ease of use by the user. Highly relevant and completely non-relevant items are in general quite quickly assessed, with relevance category 2 providing a useful 'dump' for the difficult or doubtful documents, e.g. those which the user is quite pleased to see but would not be concerned if they had not been retrieved.

Other relevance categories have of course been proposed and used in document retrieval experiments. For example in evaluating operational systems it is useful to distinguish between relevant documents which the user has already seen before being notified via the system from those which are new to him. As a generalization it might be said that too many relevance categories are not advisable with 3 or 4 probably being the optimum number.

A more fundamental issue than relevance categories is the whole concept of relevance. Its nebulous nature has been emphasized increasingly over recent years even to the extent of raising it to the realms of philosophical discourse. Nearly ten years ago Cooper[10] emphasized the distinction between

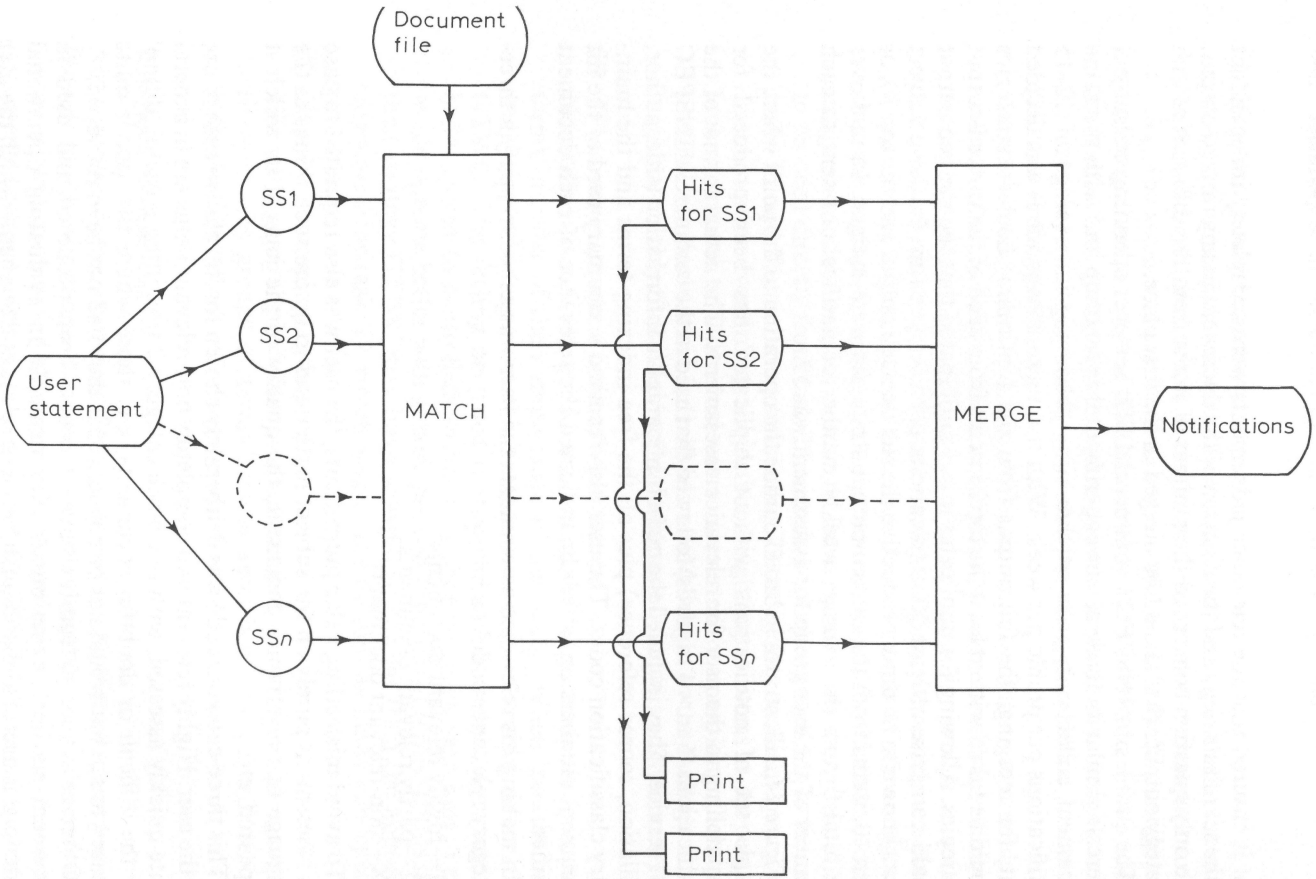

Figure 14.2. Broad outline of experiment. SS1 . . . SS*n* represent the different profile versions (search strategies) of the original query

logical relevance and utility in an information retrieval context before arguing in favour of the latter as the better basis for a measure of retrieval effectiveness[11–12]. The point was succinctly made thus—'the purpose of retrieval systems is (or at least should be) to retrieve documents that are useful, not merely relevant'. Usefulness is defined as the user's subjective evaluation of the personal utility of a retrieval system's output to him. Recognizing the difficulties of such a subjective evaluation it is suggested that more efficient compromise measures may be feasible although no ready-made solutions have been presented.

The subject still features in the literature[13] and on a practical level a real distinction certainly exists between relevance meaning 'aboutness' and relevance meaning 'pertinence'. In the former usage a relevant document is simply one which deals (to a greater or lesser extent) with the same subject matter as that of the query whereas for a document to be pertinent it has to contain information which is new and useful to the originator of the query in the subject area of the query. Obviously knowledge of pertinent documents is more important than knowledge of those which are only about the same subject as the query. However to establish the pertinence of documents requires real users, with real queries, who have the inclination to peruse entire documents. The availability of such a committed user group is rare.

In the experiment reported here relevance was used with the meaning of 'aboutness'. Although the distinction was never spelled out to the users the fact that their assessments were based on less than the whole document ensured this. Also no attempt was made to establish the extent to which users followed up the documents notified to them.

Search software

Software for the project was specially written by the INSPEC Systems Development Department. A generalized search package was developed rather than separate optimum programs specially tailored to the requirements of each search strategy.

The various search facilities available in the package are detailed in the original report[5] and are not of major concern here. Suffice it to say that the following were included: boolean AND, OR, NOT, with (practically) unlimited nesting, quorum logic, contextual logic, positive and negative integer weights (decimal or 'powers of 2'), matching in upper and lower case and in normal, inferior and superior alignment, left and right hand truncation, universal character, etc.

The most important point concerning the software was that with a generalized search package rather than separate optimal programs, the amount of information obtainable on computer costs for the different search strategies was limited. This is discussed further on pp. 306 *et seq*.

14.3 Results

Retrieval performance

The doubts raised in the literature over the years concerning the rather intangible nature of the 'relevance' concept have naturally been extended to

the performance measures based on relevance. In particular the measures first quantified by Cleverdon as recall ratio (proportion retrieved of the total number of relevant documents in the collection), precision or relevance ratio (proportion relevant of the number of retrieved documents), and fallout ratio (proportion retrieved of the total number of non-relevant documents in the collection) have been increasingly questioned.

Despite the reservations that have been expressed about them and despite the fact that theoretically more rigorous alternative measures may have been suggested, it was considered that recall and precision were still the most useful and usable measures of retrieval performance. They are easily understood and do provide answers to two of the most important questions asked of bibliographic retrieval services, viz. 'What proportion of the relevant documents have been retrieved?' and 'What proportion of the documents retrieved are relevant?'.

The recall figures established in this experiment were strictly measures of the relative (or matched) recall rather than the true recall. Matched recall is the percentage retrieved (by a particular strategy) of the total relevant documents found by all the searches for a query. With document collections averaging about 2500 per run and with a 'non-captive' user group it was not considered sensible to try and obtain relevance assessments for all the documents actually searched. However, given the very loose filtering process utilized in the experiment to control the number of notifications sent to the user for assessment (see p. 297 above), it is probable that the recall values obtained were quite close to the true recall.

In the main experiment 10 search strategies were being evaluated. Of these, 8 produced a ranked output of (effectively) unlimited size, thus allowing free choice of cutoff points at which the relative retrieval performances could be compared. The other 2 strategies B and BW, being boolean type, produced a strictly limited output which of course varied from user to user depending on the subject interest covered. The problem was that there did not seem to be any basis for comparing the boolean-type strategies with the rest other than at one point, viz. the number of items retrieved by the boolean.

It was decided to make two types of comparison:

(1) a comparison involving all the non-boolean strategies, based on Salton's rank-order cutoff-point procedure[14], and
(2) a profile-by-profile comparison, involving all the strategies, in which the basis for comparison was the boolean output.

Ranked-output comparison

The raw retrieval data included in the original report need not be reproduced here. Three of the eight search runs were evaluated, viz. runs 1, 5 and 6. The consistency in the relative retrieval performances of the search strategies over these three runs indicated that analysis of the remaining runs was unlikely to yield any different information.

For runs 1 (46 queries), 5 (45 queries) and 6 (46 queries) the cumulative totals of relevant documents retrieved by the different search strategies were aggregated at the following 9 ranked-output positions: 5, 10, 15, 20, 25, 30, 35, 45 and 55 notifications. The corresponding recall and precision figures

(averages of numbers) were also calculated. Perhaps it should be pointed out that the precision values were rather superfluous since, with the ranked-output cutoff procedure, all the available information is contained in the recall figure—at a particular cutoff point if the recall figure for one strategy is better than that for another then the precision figure is automatically similarly so.

The single performance figure usually associated with the ranked-output cutoff procedure is normalized recall[15]. Strictly this requires knowledge of the positions of all relevant items in the ranked output. The data used here were the recall figures at 9 cutoff points down to the 55th ranked document. Taking the average of the recall figures at the nine cutoff points gave an 'average' recall value which approximated to the normalized recall in its effect of ranking the search strategies in an order of merit. This order together with the approximated normalized recall values for the eight non-boolean search strategies for runs 1, 5 and 6 are shown in *Table 14.3*. No great weight can be given to the actual values of the approximated normalized recall since they depend on the number and positions of the cutoff points used. However with different cutoff points the relative positions of the search strategies would not be expected to change.

TABLE 14.3. Ranking of search strategies by normalized recall (based on 9 cutoff points, averages of numbers)

Order of merit	*Relevance R1 documents*						*Relevance R1/2 documents*					
	Search strategy (normalized recall)						*Search strategy (normalized recall)*					
	Run 1		*Run 5*		*Run 6*		*Run 1*		*Run 5*		*Run 6*	
1	TWC	(51.5)	TWC	(58.5)	GWC	(56.4)	TWC	(44.5)	GWC	(49.7)	TWC	(47.8)
2	GWC	(51.4)	GWC	(57.3)	TWC	(54.7)	GWC	(44.2)	TWC	(49.2)	GWC	(47.5)
3	GTWC	(49.0)	GTWC	(55.7)	GTWC	(53.4)	CGW	(42.5)	CGW	(46.9)	CTW	(45.9)
4	CRTW	(47.1)	CRTW	(53.0)	CGW	(52.0)	CTW	(42.4)	CTW	(46.3)	CGW	(45.8)
5	CGW	(47.0)	CTW	(52.8)	CG	(50.3)	CRTW	(41.9)	CRTW	(45.6)	CG	(44.3)
6	CTW	(46.1)	CGW	(52.6)	CTW	(50.0)	CG	(40.1)	GTWC	(45.2)	CRTW	(43.9)
7	CG	(43.1)	CG	(48.6)	CRTW	(49.4)	CT	(39.4)	CG	(43.8)	CT	(43.7)
8	CT	(41.8)	CT	(47.3)	CT	(48.5)	GTWC	(39.4)	CT	(42.6)	GTWC	(43.2)

Some points to emerge from *Table 14.3* are:

(1) The two strategies TWC (term-weight cumulation) and GWC (group-weight cumulation) are always the best, occupying first and second positions on all three runs for both relevance R1 and R1/2 documents.
(2) Strategy CT (simple co-ordination of terms) performs the worst, always being in one of the last two positions.
(3) Perhaps surprisingly, strategy CRTW (co-ordinate matching of restricted list of terms) holds its own with the others, mostly taking the middle positions.
(4) Strategy GTWC (group/term 'powers of 2' weight cumulation) is unusual in being always third best for relevance R1 documents but well down the lists when considering R1/2 documents. This suggests that the harsher consequences of 'powers of 2' weighting are not particularly suited to retrieval of 'in-between' partly relevant documents.

In an attempt to obtain more detailed information the raw retrieval data for run 5 were reworked using an additional two cutoff points (positions 1 and 2 in the ranked outputs) and calculating both by averages of numbers and averages of ratios. In the former, the (overall) mean recall ratio is calculated by dividing the total of relevant items retrieved for all the queries by the total of known relevant items for all the queries. In the latter, recall ratios are individually calculated for each query and the mean of these ratios represents the overall system (or, in our case, strategy) recall ratio. It is not obvious which method of calculation is the better; only that if the number of relevant items varies widely from query to query then the two methods may not give the same results. Also, if the two methods do not point to the same conclusions, there must be some dubiety about drawing any conclusions at all.

Table 14.4 shows the results of these additional calculations on the run 5 data. There is agreement that TWC and GWC are the best two strategies but below that the relative positions of the search strategies are not constant. The most extreme difference concerns strategy GTWC for relevance R1 documents—by the 'averages of numbers' calculation GTWC is third best but by the 'averages of ratios' it is in sixth position. Other less drastic differences are also apparent. Incidentally comparison of the 'averages of numbers' results in *Table 14.4* with the run 5 data in *Table 14.3* confirms that the use of the two additional cutoff points in the calculation of normalized recall has not changed the relative positions of any of the search strategies.

TABLE 14.4. Run 5 data—Ranking of search strategies by normalized recall (based on 11 cutoff points, averages of numbers and ratios)

Order of merit	*Relevance R1 documents*		*Relevance R1/2 documents*	
	Search strategy (normalized recall)		*Search strategy (normalized recall)*	
	Average of nos.	*Average of ratios*	*Average of nos.*	*Average of ratios*
1	TWC (49.3)	TWC (54.3)	GWC (41.5)	GWC (47.7)
2	GWC (48.3)	GWC (50.9)	TWC (41.1)	TWC (47.1)
3	GTWC (46.8)	CRTW (50.7)	CGW (39.1)	CGW (46.7)
4	CRTW (44.7)	CTW (49.7)	CTW (38.7)	CTW (45.6)
5	CTW (44.4)	CGW (47.7)	CRTW (38.0)	CG (44.6)
6	CGW (44.4)	GTWC (47.1)	GTWC (37.8)	CRTW (44.1)
7	CG (40.9)	CG (44.4)	CG (36.5)	CT (42.7)
8	CT (39.6)	CT (43.6)	CT (35.5)	GTWC (40.5)

The values given in *Table 14.4* still lack statistical significance so the run 5 data were further analysed by pairing all the search strategies in turn and, using the normalized recall figures for the individual queries, the results tested for significant difference using the sign test. *Table 14.5* shows which search strategies are significantly different at the 0.1, 1, 2, and 5 per cent levels. Where $p > 0.05$ the differences are treated as not significant. Perhaps with more confidence than before it can be concluded that strategies TWC and GWC are the best two (but not distinguishable from each other) and that CT and CG are inferior (particularly when relevance R1 documents only are considered).

TABLE 14.5. Run 5 data—sign test for significant difference

Search strategy	*Significantly better than*		*Not significantly different from*
Relevance R1 documents			
TWC	CG	($\chi^2=16.9, p<0.001$)	GTWC ($\chi^2=3.03$), CRTW ($\chi^2=3.03$)
	CT	($\chi^2=14.4, p<0.001$)	GWC ($\chi^2=0.03$)
	CTW	($\chi^2=10.0, p<0.01$)	
	CGW	($\chi^2=6.4, p<0.02$)	
GWC	CG	($\chi^2=12.1, p<0.001$)	CTW ($\chi^2=3.6$), GTWC ($\chi^2=2.03$)
	CT	($\chi^2=11.0, p<0.001$)	CRTW ($\chi^2=2.03$), TWC
	CGW	($\chi^2=9.0, p<0.01$)	
CRTW	CT	($\chi^2=8.1, p<0.01$)	GTWC ($\chi^2=0.03$), CTW ($\chi^2=0$)
	CG	($\chi^2=4.9, p<0.05$)	CGW ($\chi^2=0$), TWC, GWC
CTW	CT	($\chi^2=18.2, p<0.001$)	CGW ($\chi^2=0.03$), GTWC ($\chi^2=0.03$)
	CG	($\chi^2=11.0, p<0.001$)	GWC, CRTW
CGW	CG	($\chi^2=7.2, p<0.01$)	GTWC ($\chi^2=0.23$), CRTW, CTW
	CT	($\chi^2=6.4, p<0.02$)	
GTWC	CT	($\chi^2=5.6, p<0.02$)	TWC, GWC, CRTW, CTW, CGW
	CG	($\chi^2=5.6, p<0.02$	
CG	—		CT ($\chi^2=1.6$)
CT	—		CG
Relevance R1/2 documents			
TWC	CT	($\chi^2=12.8, p<0.001$)	CRTW ($\chi^2=3.76$), CGW ($\chi^2=2.69$)
	GTWC	($\chi^2=9.8, p<0.01$)	CTW ($\chi^2=1.8$), GWC ($\chi^2=0.02$)
	CG	($\chi^2=7.2, p<0.01$)	
GWC	CRTW	($\chi^2=9.8, p<0.01$)	CTW ($\chi^2=3.76$), CGW ($\chi^2=3.76$)
	GTWC	($\chi^2=8.9, p<0.01$)	TWC
	CT	($\chi^2=8.0, p<0.01$)	
	CG	($\chi^2=6.4, p<0.02$)	
CRTW	—		CT ($\chi^2=1.09$), CGW ($\chi^2=1.09$) GTWC ($\chi^2=0.56$), CG ($\chi^2=0.36$) CTW ($\chi^2=0.36$), TWC
CTW	CT	($\chi^2=21.4, p<0.001$)	CGW ($\chi^2=0.09$), GTWC ($\chi^2=0.02$)
	CG	($\chi^2=6.4, p<0.02$)	TWC, GWC, CRTW
CGW	CT	($\chi^2=11.8, p<0.001$)	GTWC ($\chi^2=1.8$), TWC, GWC
	CG	($\chi^2=5.7, p<0.02$)	CRTW, CTW
GTWC	—		CT ($\chi^2=0.8$), CG ($\chi^2=0$) CRTW, CTW, CGW
CG	CT	($\chi^2=4.4, p<0.05$)	CRTW, GTWC
CT	—		CRTW, GTWC

It might be argued that a more powerful test such as the Wilcoxon matched pairs signed ranks test could have been used since the magnitude of the difference in the normalized recall between pairs of search strategies was known for all the queries. This was not pursued because of unease concerning the validity and overall effect of those queries with very few relevant documents; as mentioned earlier, in the extreme case of a query with only 1 relevant document a recall ratio of 0/1 could so easily be 1/1, and vice versa.

Boolean comparison

The method used for this individual profile-by-profile comparison of all the strategies was as follows:

(1) The output for the boolean strategy B (the only strategy producing an unranked output) was counted (say x items), as were the number of relevance R1 (say y) and relevance R1/2 (say z) documents it contained.
(2) For the remaining strategies in turn, starting from the top of the ranked output, the numbers of items taken to retrieve y relevance R1 and z relevance R1/2 documents were noted—say x_1 and x_2 items respectively.

If both x_1 and x_2 for a particular strategy are less than the boolean output x, then that strategy is performing better than the boolean strategy B for both R1 and R1/2 documents. If $x_1 < x < x_2$, then that strategy is performing better than strategy B for R1 documents but not for R1/2 documents. And so on.

This procedure gave a best strategy (or joint best strategies) for each query on each run. The main disadvantage of the method is that it is entirely dependent on the boolean output; if there is no boolean output or if the boolean output contains no relevance R1 or R1/2 documents, there is no basis for comparison. Also, under the conditions of this experiment, a comparison is not valid if the boolean output is very large.

Aggregating, from runs 1, 5 and 6, the results for the top 3 search strategies only for each query a ranking of search strategies was obtained (*Table 14.6*). This shows the percentage of times each search strategy occupied one of the top 3 positions. Incidentally, in the calculations, where, for example, two (or more) strategies were equal best they were both (all) ranked first but the next best was ranked third (or fourth, etc., as appropriate). That is, a jointly held first position was considered of equal merit to a uniquely held first position.

It is seen from *Table 14.6* that, by this boolean comparison method, clearly the two best strategies are respectively BW and GWC, with strategy TWC, not so decisively, third best. The least promising strategy according to this method is CG. It is interesting that in the ranked-output comparison based on normalized recall (p. 300), strategies GWC and TWC are practically indistinguishable whereas in the boolean comparison strategy GWC comes out better. A possible explanation for this difference was thought to lie in the methods of comparison—the ranked-output method is essentially 'neutral' but the boolean comparison, being based on the boolean output, may be more oriented towards those strategies which comprise term groups (e.g. GWC) rather than a single list of terms (e.g. TWC). However this is not confirmed by the relative positions of strategies CT (single list of terms) and CG (term groups) in the two comparisons.

In addition to the disadvantages of the boolean comparison method mentioned above it is now thought that there may be more fundamental objections. Not only are strategies B and BW considered to be misleadingly rated relative to the other strategies (B too low, BW too high) but the whole concept of evaluating an optimum boolean (yes/no) output against a ranked output may be questionable—it is not comparing like with like. On the other hand one of the principal criticisms of the experiment might be said to have been the failure to develop a valid method for comparing optimum boolean outputs with ranked outputs. To use a less-than-optimum boolean statement as a weak filter in a first pass of the document collection and then ranking the resulting output by some weighting scheme is a useful experimental convenience but it does not correspond to the strategy BW in this experiment.

TABLE 14.6. Ranking of search strategies by boolean comparison

Order of merit	*Relevance R1 documents*						*Relevance R1/2 documents*					
	Ranked 1 (%)		*Ranked 1 or 2 (%)*		*Ranked 1, 2 or 3 (%)*		*Ranked 1 (%)*		*Ranked 1 or 2 (%)*		*Ranked 1, 2 or 3 (%)*	
1	BW	(52.1)	BW	(65.8)	BW	(74.0)	BW	(45.6)	BW	(55.6)	BW	(65.6)
2	GWC	(23.3)	GWC	(35.6)	GWC	(46.6)	GWC	(33.3)	GWC	(42.2)	GWC	(54.4)
3	GTWC	(23.3)	TWC	(31.5)	TWC	(38.4)	TWC	(23.3)	TWC	(28.9)	CTW	(40.0)
4	TWC	(17.8)	GTWC	(27.4)	CTW	(34.2)	CGW	(23.3)	CRTW	(28.9)	CRTW	(40.0)
5	CTW	(17.8)	CTW	(23.3)	B	(34.2)	CRTW	(23.3)	CGW	(27.8)	TWC	(36.7)
6	CRTW	(17.8)	CRTW	(23.3)	GTWC	(32.9)	CT	(21.1)	CT	(26.7)	GTWC	(35.6)
7	CT	(9.6)	B	(23.3)	CRTW	(30.1)	GTWC	(16.7)	B	(26.7)	CGW	(35.6)
8	CGW	(9.6)	CT	(16.4)	CGW	(21.9)	B	(16.7)	CTW	(24.4)	B	(35.6)
9	CG	(8.2)	CGW	(15.1)	CG	(20.5)	CTW	(15.6)	GTWC	(24.4)	CT	(33.3)
10	B	(8.2)	CG	(12.3)	CT	(19.2)	CG	(14.4)	CG	(21.1.)	CG	(32.2)

Our boolean statements were meant to be optimum ones but to have optimum boolean profiles producing ranked outputs suggests a conceptual contradiction. To have used less-than-optimum boolean statements would be no real test of boolean logic.

Costs

In a computerized search system the costs of the system may be allocated to a number of factors—original database production or purchase; computer operations such as searching, output preparation and printing, profile maintenance, etc.; and non-computer operations such as the time of information scientists, keyboarding, general clerical, etc. Two of these factors were considered in this investigation, viz. information scientist time and computer search costs. Now, nearly five years later, the values used for salaries and computer processing costs are dated. Rather than simply update them this section considers in more general terms what was learned in the experiment about the various cost factors and how they related to conditions in an operational system.

Intellectual (information scientist) effort

Although an accurate record of events in the experiment, the times obtained for profile compilation and modification (pp. 293 and 294 above) are somewhat artificial in the context of a real system in that: (1) only one analysis/modification stage was undertaken; (2) the times for trivial (but necessary) profile modifications were not recorded; and (3) only the actual compilation/modification times were included, i.e. there is no allowance for 'dead-time' (tea breaks, etc.) which in real life might involve an additional 20 per cent or so.

To translate the figures on pp. 293 and 294 to a more real situation we could: (1) assume an average of 3 significant profile modifications per year; and (2) add 20 per cent dead-time to the above compilation/modification times. If that is done comparative estimates are obtained for the information scientist effort required per query per annum for the different search strategies (*Table 14.7*).

TABLE 14.7. Estimated annual information scientist effort per query

Search strategy	*Information scientist time* (min)
CT	90
CTW	107
CG	107
TWC	110
CGW	125
GTWC	128
GWC	136
B	151
BW	168

Assuming a 35-hour week and 47 weeks worked per year, the number of profiles handled per year by an information scientist working full time would

range from about 1000 for CT-type profiles to about 550 for BW-type profiles.

These figures still tend to the optimistic in that they account only for the purely 'intellectual' aspects of profile handling. Normally time will also be spent liaising with users (existing and potential), checking subscriptions, etc. This variable, however, related more to job responsibility than any basic profile compilation or modification procedure, would not be expected to change by much the relative intellectual effort required by the different search strategies.

Computer search costs

With a generalized search package rather than separate optimal programs, the amount of information obtainable on the actual computer costs for the different search strategies was limited. Tests with the software showed that it is the matching of profile and document terms rather than profile logic evaluation which controls the search rate; in other words, the number of profile search terms is paramount and the profile logic evaluation represents a minor part of the total computer processing cost.

In the experiment, for each user statement, all search strategies except CRTW used the same set of search terms, and it may be assumed that the computer search costs for the different search strategies except CRTW would not differ by more than 20 per cent.

Financial considerations dictated that the different search strategy types were not run separately against the document collections and that the computer matching runs were conducted as a background job timeshared with a variety of other tasks which differed from run to run.

In these less-than-ideal conditions the computer search costs per query (search strategy) per year were some 3–5 times the cost of information scientist time. The other costs mentioned earlier but not investigated would diminish further the contribution of information scientist time to perhaps 10–15 per cent of the total overall costs. The changing balance in the man–machine equation probably means that the information scientist time would now contribute a greater share to the total costs than was the case 5 years ago.

Clearly there are two main approaches to effecting savings in the computer search costs, viz. reduction in the number of search terms and simplification of the term-matching procedure. The latter might be achieved by forgoing some of the more sophisticated matching facilities, e.g. simultaneous left- and right-hand truncation, ability to distinguish upper and lower case characters, and the universal character. Individual profiles can be found where one or more of these facilities is very useful and convenient but in most instances their absence can be surmounted by the use of additional search terms at a lesser overall cost. The value of these facilities varies with subject area and in the chemical field the truncation facility might be considered vital. A valid evaluation of their importance would probably require a large number of search profiles for any significant effect to be apparent.

Cost-effectiveness

It became apparent early in the project that insufficient data would be obtained to enable any absolute conclusions to be drawn concerning the

overall most cost-effective search strategy. The discussion on p. 307 suggests that the computer search costs for all search strategies except CRTW would not differ by more than 20 per cent because they all use the same basic set of search terms. However strategy CRTW, which by definition always contains 20 search terms compared with an average of 46 terms for the other strategies, is thereby much more economic in terms of computer search costs. If its retrieval performance is on a par with the other strategies then strategy CRTW must be the most cost-effective of all the 10 strategies evaluated in the main experiment. The results presented on pp. 299 et seq. show that this is in fact the case.

A more modest possibility with the available information is a comparison of the cost-effectiveness of all the search strategies in terms of information scientist cost only. A suitable measure of effectiveness is 'relevant documents retrieved' and, certainly in the SDI situation, the number of relevant documents retrieved at cutoff points of, say, 15 or 25 documents would seem to be the most appropriate.

The basic 'effectiveness' data included in the original report show, for example, that at a cutoff point of 15 items, strategy CT retrieved 164, 127 and 120 relevance R1 documents, respectively, on runs 1 (46 queries), 5 (45 queries) and 6 (46 queries). This gives an average figure of $(164+127+120)/(46+45+46)$, i.e. 3.0 relevance R1 documents retrieved per query per run by strategy CT at a cutoff point of 15 items. Assuming 50 runs per year (weekly SDI service) this figure becomes 150 relevance R1 documents retrieved per query per year. Dividing this number by the one given for information scientist effort on strategy CT in *Table 14.7* gives a figure for the cost-effectiveness of strategy CT at a cutoff point of 15 items, i.e. 150/1.5 or 100 relevance R1 documents retrieved per year per hour of information scientist effort. Repeating the calculation for all the strategies at cutoff points of 15 and 25 items gives the comparative cost-effectiveness figures shown in *Table 14.8*.

TABLE 14.8. Cost-effectiveness of search strategies (information scientist effort only)

Order of merit	*Strategy cost-effectiveness (relevant documents retrieved/year/hour of information scientist time)*			
	Relevance R1 documents		*Relevance R1/2 documents*	
	Cutoff 15	*Cutoff 25*	*Cutoff 15*	*Cutoff 25*
1	CRTW (108.8)	CRTW (150.9)	CRTW (251.1)	CRTW (355.0)
2	CT (100.0)	CT (139.4)	CT (233.6)	CT (347.0)
3	TWC (98.2)	TWC (136.0)	TWC (211.6)	TWC (311.4)
4	CTW (91.9)	CTW (124.7)	CTW (204.5)	CTW (303.3)
5	CG (89.0)	CG (121.4)	CG (202.8)	CG (294.1)
6	GTWC (83.3)	GTWC (114.8)	CGW (180.8)	CGW (264.9)
7	CGW (81.8)	GWC (109.5)	GWC (175.5)	GWC (251.9)
8	GWC (81.3)	CGW (108.3)	GTWC (169.9)	GTWC (245.5)
9	BW (62.8)	BW (77.8)	BW (128.0)	BW (167.9)

Notes: (i) The information scientist effort for strategy CRTW was assumed equal to that for strategy CT (see p. 291)
(ii) The figures for boolean strategy BW are not strictly comparable with the others since some of the boolean outputs were less than 15/25 items. Strategy B was omitted.

Some points of interest to emerge from *Table 14.8* are:

(1) Strategy CRTW is confirmed, without qualification, as the most cost-effective search strategy.
(2) Compared with the ranking of strategies based on retrieval performance only (p. 301), strategies which have strikingly changed their relative positions are CT (upwards), GWC (downwards), and BW (downwards).
(3) The 4 strategies comprising basically a single list of search terms, viz. CRTW, CT, TWC and CTW, occupy the top four positions.

Free- and controlled-language comparison

Although not envisaged as part of the original project the data that became available during the experiment was considered suitable for a direct comparison of free-language and controlled-language boolean profiles in the INSPEC environment. The data covered profile compilation times, number of search terms, and recall/precision performance figures. These are all detailed in the original report and are only summarized here.

The average compilation time for the controlled-language boolean profiles (31 min) was just less than half that for the free-language boolean profiles (65 min). The times recorded were for a compiler who was already familiar with the controlled language concerned—INSPEC's thesaurus and unified classification scheme. These compilation times may be slightly biassed in favour of the controlled-language profiles because invariably the free-language versions were compiled first. When the controlled-language version came to be compiled there would probably be some memory of the original user statement even though the free-language version may have been compiled some weeks earlier.

As well as having shorter compilation times, the controlled-language boolean profiles were smaller than the free-language boolean profiles by a factor of $2\frac{1}{2}$, averaging 19 terms and 47 terms respectively. It should be pointed out that one reason for the smaller number of search terms in the controlled-language profiles is that they were used in searching in a 'free-text' way, i.e. extensive use was made of truncation in the search terms. Assuming an approximately linear relationship between the number of profile search terms and computer search time, this factor of $2\frac{1}{2}$ would be largely reflected in the computer search costs in favour of the controlled-language profiles.

There is a further saving of search time for controlled-language profiles because the controlled-language searchable fields are smaller than the free-language searchable field in the INSPEC database. Statistics then current indicated that the relative sizes of the three fields were in the ratios:

Free indexing	10
Subject headings (thesaurus terms)	5
Unified classification codes	1

Because compilation of the controlled-language boolean profiles was started after the main experiment had got under way the quantity of experimental data for the first few SDI runs was limited. Only the recall/precision figures for the last three of the eight runs were analysed, i.e. those for runs 6, 7 and 8. Overall values for recall and precision calculated by the usual two averaging methods (average of numbers and average of ratios)

are given in *Table 14.9*. Although they seem to show a consistently better overall retrieval performance for the controlled-language boolean profiles further analysis using the sign test for significant difference did not support this. *Table 14.10* records the number of times controlled-language (CL) or free-language (FL) profiles showed superior retrieval performance on runs 6, 7 and 8. The highest χ^2 value for this data is 2.0 so no significant difference is indicated even at the 10 per cent level.

TABLE 14.9. Retrieval performance of controlled-language and free-language boolean profiles

Run no.	*No. of queries*	*Profile type*	*Averaging method*	*Retrieval performance*			
				Recall (%)		*Precision* (%)	
				R1	*R1/2*	*R1*	*R1/2*
6	24	Controlled-language	Av. of nos.	75.2	60.3	29.1	61.9
			Av. of ratios	64.6	50.6	25.6	59.6
		Free-language	Av. of nos.	56.1	43.8	23.1	48.0
			Av. of ratios	60.3	44.8	27.4	53.0
7	34	Controlled-language	Av. of nos.	58.0	49.4	20.8	57.6
			Av. of ratios	57.6	46.5	25.6	58.6
		Free-language	Av. of nos	51.4	39.7	17.6	44.4
			Av. of ratios	57.2	42.8	21.4	49.8
8	32	Controlled-language	Av. of nos.	63.7	53.0	19.5	50.4
			Av. of ratios	53.7	45.7	18.1	43.9
		Free-language	Av. of nos.	57.9	45.9	15.8	39.0
			Av. of ratios	53.6	42.8	14.3	40.7

TABLE 14.10. Retrieval performance of controlled-language (CL) and free-language (FL) boolean profiles

Run no.	*Recall*						*Precision*					
	R1 documents			*R1/2 documents*			*R1 documents*			*R1/2 documents*		
	CL better	*FL better*	*Same*	*CL better*	*FL better*	*Same*	*CL better*	*FL better*	*Same*	*CL better*	*FL better*	*Same*
6	8	4	9	12	9	3	8	10	6	14	10	0
7	10	8	11	15	12	7	13	15	6	20	12	2
8	8	9	12	15	12	5	15	9	8	19	10	3

14.4 Conclusions

As is often the case with experiments in information retrieval where conditions are peculiar to one situation or organization, the results obtained in the major project may be valid only for the INSPEC database. In particular a factor that might be expected to influence the experiment would be the medium used for matching profiles and documents; in this case the free-index terms assigned to all items in the database. INSPEC's operational statistics at the time indicated that the free-index field contained on average

some 7 phrases/item, or, about 15 singlet terms/item. It might be argued that this number may be small enough to be prejudicial against certain of the search strategies. No particular view is offered on this point other than that it was not felt to be the case during the experiment and does not seem to be obviously so. With the above qualification the following conclusions may be drawn from the experimental results:

(1) As measured by information scientist effort expended on the purely intellectual aspects of profile compilation and modification, the simplest search strategy, CT (co-ordinate matching of terms without weights), occupied almost exactly half as much time as the most complex strategy, BW (boolean logic with weights).
(2) The search strategies exhibiting the best retrieval performance were GWC (group-weight cumulation) and TWC (term-weight cumulation). In the boolean comparison of retrieval performance, strategy BW appeared to do very well but it is now considered that the method of evaluation was faulty and no conclusions are drawn concerning either of the boolean search strategies. The worst performer was strategy CT, always being in one of the last two positions.
(3) Although the best retrieval performances were produced by strategies using weighting techniques, experience gained during the project in subjectively assigning weights to terms suggested that the majority of SDI users would not be particularly attracted to doing this task for themselves.
(4) The most cost-effective strategy overall was CRTW (co-ordinate matching of restricted list of terms with weights). In terms of information scientist effort only, the most cost-effective strategies were CRTW, CT and TWC, and, although not strictly comparable, the least cost-effective was BW.

In the secondary experiment comparing controlled-language and free-language boolean profiles, the former: (1) were compiled more quickly (given pre-knowledge of the controlled language); (2) comprised fewer search terms; and (3) showed comparable overall retrieval performance. Their main drawback is that the use of controlled language is not likely to appeal to those non-information workers who wish to prepare their own profiles. Although not evident in this study another factor which can work against controlled-language profiles is that in subject areas where new terminology is being introduced rapidly the controlled language may lag behind and be inadequate until updated.

14.5 Retrospect

Looking back after some five years the experiment is seen to have been in the mainstream of information retrieval research at the time. On the whole its methodology was based on established procedures and it also reflected the changing emphasis in retrieval experiments, viz. whereas in the 1960s the main interest had been in indexing languages, by the early 1970s the concentration was on search techniques. With the growing interest in

automatic indexing it is now being seen more clearly how interdependent are indexing and searching methods.

At various points in the above description some shortcomings of the original experiment have been mentioned. It may be useful to conclude by gathering together and discussing these defects and also those questions which were raised but remained unresolved. It is hoped that some activities were performed adequately but inevitably they are of less interest and will only be mentioned briefly.

Those parts of the investigation which are considered to have been sound include: a very adequate document collection; a meaningful range of search strategies; a realistic profile compilation method involving standard tasks which allowed an accurate measure of the effort required from the information scientist on the different search strategies; a valid procedure for collecting relevance assessments; and the recruitment of the user group and the mechanics of the experiment in general.

Less satisfactory areas include: the rather low number of queries; retrieval performance evaluation by the boolean comparison method; the absence of automatic term-weighting; the lightweight nature of the cost data; and the significance of the experimental results.

Concerning the number of queries it is now considered (although nowhere proved) that perhaps twice the number of queries would have been more convincing; or, at least a number sufficient enough that the results of a few individual queries do not obtrude on the overall results. In our experiment this effect was exemplified by the differences observed when calculating by the two averaging methods, numbers and ratios. With a greater number of queries it would also have been possible to ignore those queries for which there were too few or, less importantly, too many relevant items in the collection. It is not clear what the implications of such a practice are but certainly the results would thereby be more reproducible. As has already been mentioned too many recall/precision ratios of the order 0/1, 1/1, etc., are not really acceptable. The problem could have been eased indirectly if a more drastic approach had been taken originally with some of the user interest statements. Those that clearly comprised more than one question could have been treated separately. This would have resulted in 'cleaner' profiles of which fewer were overlong, some profile performances would probably have been subject to less extraneous influences, and the number of queries would have been larger. Although a token number of the user statements were in fact split up more could have been and the experiment would have been better for it. At the time the view taken was that as little as possible should be done to change the conditions from that of 'real life' and, since these were statements very like those received from users of an operational system, the less tampering the better. This is now deemed to have been misguided and to have done what is now suggested would not have affected the validity of the test in any way.

The most disappointing outcome of the whole experiment was the failure to develop an acceptable method for comparing an optimum boolean strategy with any strategy producing a ranked output. A few simple examples quickly show the inappropriateness of using the boolean output itself as the basis for comparison. Very little can be offered in the way of a solution even now and

maybe it is not sensible to attempt such a comparison in that like is not being compared with like.

An often-voiced criticism of the original report was that too little attention had been paid to the statistical significance of the experimental results. An attempt has been made to rectify that in this paper and undoubtedly it has helped to clarify the picture. However establishing statistical significance when differences are not obvious is still a limited achievement and there is much appeal in the philosophy of what might be termed a Cleverdonian maxim—if one needs to resort to statistical techniques to establish performance differences in information retrieval experiments then the differences are not worth knowing about.

It was mentioned almost in passing that assigning weights subjectively to search terms and term-groups was unlikely to appeal greatly to users attempting their own profiling. For this reason alone it is a pity that automatic weighting of terms was not possible in the evaluation. Also since all profiling was carried out by one experienced compiler there was no impression gained as to the likely ease-of-use, and acceptance, by end-users (as opposed to professional information staff) of the different search strategies. This failing was realized from the start but it was thought to be too difficult to surmount easily. The decision to use only one compiler at least ensured the control of this variable.

It has been thought that perhaps not enough was attempted at the time to establish the reasons why the strategies performed as they did. The question of how much failure analysis should be done was considered at some length. Where strategies performed as might have been anticipated (e.g. it was not surprising to find that the best retrieval performance was produced by strategies using weighting techniques) there seemed little purpose in detailed analysis. In the evaluation of an operational system it is clearly important to obtain a measure of which activities are responsible for retrieval failures—in particular what proportion can be allocated to poor indexing or to profile compilation errors. In this experiment since, for each query, all strategies shared the same basic list of search terms, such failures would be common to all strategies. Thus the main interest was in distinguishing any differences due to the characteristics of the strategies themselves. The most promising procedure seemed to be to examine those queries for which the strategies, which performed best overall, did unusually badly. Pursuing this method showed up one clear link between search strategy performance and type of user statement. In those strategies comprising a single list of terms (CT, TWC, CTW) as opposed to those including term groups, there is the possible deleterious effect of having one concept 'swamping' all others in the list of profile search terms. The damage occurs when the document free-indexing is similarly 'unbalanced'; this of course can quite legitimately be inevitable. For example, the concept 'metals' comprises more than 60 individual metal elements and the literature is such that often a paper on some aspect of metal behaviour deals with a number of different metals all of which are properly included among the index terms. The result is that the outputs from single-list strategies are top-heavy with the individual 'metal' terms which, in the term-group strategies, are controlled by virtue of being in a term group which contributes only once to the total weight irrespective of the number of

matching terms in that group. Although the effect was quite considerable with the 'metals' concept it did not occur so obviously with other concepts, because most documents are such that the same concept is not required to be represented a large number of times in the free-index terms of a particular document.

Perhaps the one surprising experimental result which did merit consideration was the comparatively good retrieval performance of strategy CRTW. Of the other strategies CRTW was most similar in type to CTW. The fact that a reduction in the number of search terms by more than half (CRTW always contained 20 terms and CTW contained, on average, 46 terms) had produced no significant difference in retrieval performance was unexpected and rather deflating. It suggests that the number of search terms in profiles should be optimized rather than compiled exhaustively. The decision that strategy CRTW should contain a maximum of 20 terms for each query was arbitrary and not much more than a reasonable assumption, and the optimum number of terms would be expected to vary from query to query depending on their subject matter. One explanation for the result, offered originally by Cleverdon, is that the arithmetic product—'Number of profile search terms × Number of document index terms'—has a critical value which, if exceeded, results in a deteriorating retrieval performance. This view is supported by the results from another INSPEC project[16] in which profile search terms were expanded automatically by reference to a thesaurus as a source of, successively, synonyms, narrower terms and 'see also' terms. It was found, against expectation, that although the retrieval performances 'were not usefully different . . ., the general trend was for poorer recall with each expansion'. The important factor was that the base profile version contained 41 terms on average and the final expanded version contained 122 terms on average, figures which were probably well above the optimum. This effect of profile length can of course be counteracted by the use of weights and grouping of terms but it remains an interesting point when searching by the simplest strategy, straight co-ordination of unweighted terms.

References

1. CLAGUE, P. *SDI Investigation 1967–1969*, 5 Vols, Report R71/6, INSPEC, Institution of Electrical Engineers, London (1971)
2. AITCHISON, T. M. *et al. Comparative Evaluation of Index Languages: Part 1–Design; Part 2–Results*, Reports R70/1 and R70/2, INSPEC, Institution of Electrical Engineers, London (1969, 1970)
3. EVANS, L. *Optimum Degree of User Participation in SDI Profile Generation*, Report R73/12, INSPEC, Institution of Electrical Engineers, London (1973)
4. CLEVERDON, C. W. and HARDING, P. *Report on an Investigation into a Mechanised Information Retrieval Service in a Specialised Subject Area* (CRISPE Project), Cranfield Institute of Technology (1970)
5. EVANS, L. *Search Strategy Variations in SDI Profiles*, Report R75/21, INSPEC, Institution of Electrical Engineers, London (1975)
6. SPARCK JONES, K. and VAN RIJSBERGEN, C. J. *Report on the Need for and Provision of an 'Ideal' Information Retrieval Test Collection*, British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge (1975)
7. SPARCK JONES, K. and BATES, R. G. *Report on a Design Study for the 'Ideal' Information Retrieval Test Collection*, British Library Research and Development Report 5428, Computer Laboratory, University of Cambridge (1977)

8. GILBERT, H. and SPARCK JONES, K. *Statistical Bases of Relevance Assessment for the 'Ideal' Information Retrieval Test Collection*, British Library Research and Development Report 5481, Computer Laboratory, University of Cambridge (1979)
9. MATTHEWS, F. W. and THOMSON, L. Weighted term search: a computer program for an inverted coordinate index on magnetic tape, *Journal of Chemical Documentation* **7**, 49–56 (1967)
10. COOPER, W. S. A definition of relevance for information retrieval, *Information Storage and Retrieval* **7**, 19–37 (1971)
11. COOPER, W. S. On selecting a measure of retrieval effectiveness, *Journal of the American Society for Information Science* **24**, 87–100 (1973)
12. COOPER, W. S. On selecting a measure of retrieval effectiveness: Part II—Implementation of the philosophy, *Journal of the American Society for Information Science* **24**, 413–424 (1973)
13. SWANSON, D. R. Information retrieval as a trial-and-error process, *Library Quarterly* **47**, 128–148 (1977)
14. SALTON, G. *Automatic Information Organisation and Retrieval*, McGraw-Hill, New York (1968)
15. CLEVERDON, C. and KEEN, M. *Factors Determining the Performance of Indexing Systems: Volume 2—Test Results*, ASLIB Cranfield Research Project, College of Aeronautics, Cranfield (1966)
16. EVANS, L. and GOULD, A. M. *Automatic Aids to Profile Construction: Expansion of Search Terms by Thesaurus*, Report R76/25, INSPEC, Institution of Electrical Engineers, London (1976)