DOCUMENT RESUME

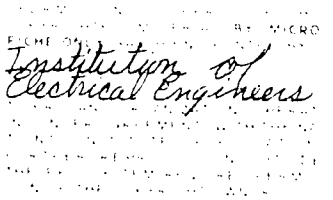
ED 081 447 LI 004 455

AUTHOR Wilman, H.; Hall, Angela M.
TITLE INSPEC: An Experiment in the Batch Processing of Retrospective Searches.
INSTITUTION Institution of Electrical Engineers, London (England).
SPONS AGENCY Office for Scientific and Technical Information, London (England).
REPORT NO INSPEC-R-73-13
PUB DATE Feb 73
NOTE 41p.
AVAILABLE FROM INSPEC, The Institution of Electrical Engineers, Savoy Place, London WC2R OBL, England (pound 1.25)

EDRS PRICE MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS Automation; Comparative Analysis; Computers; Costs; Information Processing; *Information Retrieval; Relevance (Information Retrieval); *Search Strategies
IDENTIFIERS *Manual Searches

ABSTRACT

A series of five batches of twenty searches were carried out in a three year span of the INSPEC data base by the 31 Company's Information Retrieval Service. The same queries were processed manually in the corresponding volumes of Science Abstracts and comparison of the output of each search method exposed reasons for the failures. Emphasis is laid on the fact that a query comprises other factors in addition to its subject matter and means of satisfying these criteria automatically are discussed. The more flexible nature of the manual searches is examined in detail with a view to improving the flexibility of the computer searches. (Author/SJ)



Report No. R73/13
ISBN No. 852964153

INSPEC

An experiment in the batch processing of retrospective searches.

H Wilman

Angela M Hall

February 1973

The Institution of Electrical Engineers
Savoy Place, London WC2R OBL.

SUMMARY

A series of five batches of twenty searches were carried out in a three year span of the INSPEC data base by the Bi's Co Information Retrieval Service. The same queries were processed manually in the corresponding volumes of Science Abstracts and comparison of the output of each search method exposed reasons for the failures. Emphasis is laid on the fact that a query comprises other factors in addition to its subject matter and means of satisfying these criteria automatically are discussed. The more flexible nature of the manual searches is examined in detail with a view to improving the flexibility of the computer searches.

ACKNOWLEDGEMENT

The work reported was supported by a grant from the Office for Scientific and Technical Information of the Department of Education and Science.

CONTENTS

APPENDICES

1. INTRODUCTION

Over the past few years in which INSPEC has been operating a computer-based SDI service, a considerable body of experience has been accumulated concerning the processing of the current-awareness query. Less is known, however of the problems and economics of searching a data base retrospectively.

The 3i Co, through its Information Retrieval Service, provides for the batch processing of profiles against the cumulated files of data bases. The addition of the INSPEC file to the data bases which they currently hold has provided INSPEC with an opportunity to assess, without large capital outlay, the potential of batch-mode processing of retrospective searches. Consequently, an arrangement was made for INSPEC to submit profiles for processing on a regular basis over a period of five months. In this way INSPEC was able to gain experience in the selection of queries suitable for computer searching and in the compilation of profiles.

The performance of the searches was studied in some detail and Section 3 of this report discusses, with the aid of illustrations, the problems which were encountered. Many of these problems were concerned with the compilation of profiles for searching the free-indexed records particular to the INSPEC data base.

In one sense, INSPEC as a data base producer stepped into the shoes of the customers and faced the problems which they experienced when searching the data base. Although INSPEC has a greater familiarity with the content of the data base and indexing procedures, it had no control over the method of processing the data. Hence, it was possible to see how the performance of the searches fell short of the potential of the data base because the processors were not fully aware of the structure and content of records.

Of particular interest to the investigation was the comparison of the computer search with the more conventional search made in a printed subject index. The value assessments made of the computer output and the detailed records of the parallel manual searches which are discussed in Sections 4 and 5 highlight a number of the differences between the two.

As a result of the findings of the retrieval analysis and comparison with the manual search, a number of ways in which the performance of the computer search might be enhanced can be suggested. These suggestions are drawn together in Section 6 and are summarised in Appendices 8 - 11. It is hoped that they will provide useful notes for the guidance of the data base processors and those who compile search profiles.

The costs of the searches although they are unique to the test situation and not representative of an operational service, are presented in Section 7. The report is then completed by the conclusions and recommendations of Sections 8 and 9.

2. PROCEDURE

2.1 3i Science Information Center

The investigation was made possible by an arrangement with the 3i Co. for INSPEC to provide a series of enquiries which they would process against the INSPEC data base through their Information Retrieval Service. This service offers to the customer the facility to search for relevant references in a choice of the cumulated files of the major scientific and technical data bases. The searches are carried out in batch-mode and process serial files of the document records.

The search profiles may be formulated either as Boolean search statements or by the use of term weighting, and the common facilities of term truncation and search field specification are available. The Boolean logic statements which may be used are of a restrictive nature. Each profile is partitioned into fields such that terms within a field are related by the operator 'OR' and the fields themselves are related by the operator 'AND'. This format is illustrated in the profile coding sheet shown in Appendix 1.

The profile is matched against selected fields of the data base. In this investigation these contained free-indexing phrases, subject index headings, classification codes and treatment codes. These fields are displayed in the output to show the occurence of term matches. (Appendix 2)

2.2 Questions and Profiles

INSPEC has within its own organisation a library which serves an extensive community of electrical engineers. The library therefore is continually receiving technical enquiries and these provided a sustained source from which a selection could be made for the investigation.

The enquiries, which were recorded by the library information staff, were passed to one of a group of INSPEC staff for the compilation of a profile. The staff who compiled the profiles had a wide knowledge of the content of the data base but with one exception had no experience of profile compilation and so a brief explanation and practical introduction to profile compilation were given at the outset of the investigation.

The profiles were sent in five monthly batches of 20 queries to the 3i Co., where they were key-punched and run against the INSPEC data base for 1969 - 1971.

When each batch of search output was received from the 3i Co, an assessment was made (by the information staff) of the usefulness of each reference in response to the query. This then provided a measure of the effectiveness of the searches.

2.3 Manual Searches in the Printed Subject Indexes

Concurrently with these computer searches, a manual search was carried out by the library information staff by their usual methods. The results of the searches, which were supplied to the enquirer as the normal service, provided a yardstick by which the performance of the computer searches could be assessed.

For 22 queries a detailed record was kept of the steps and decisions taken in the search of the printed indexes in the corresponding volumes of the Abstracts Journals. A more exhaustive analysis could then be made of the reasons for retrieval failures and the ways in which the more flexible nature of the manual search is revealed.

2.4 Review of items retrieved

Whether a search is carried out in a printed subject index or by a computer search the output is usually subject to review usually by recourse to the abstract. The information officer may impose additional constraints in an effort to tailor the output for the enquirer's use, and a number of ways in which this may be accomplished can be suggested. For example, a paper may be rejected because it is in an obscure language or the treatment is too erudite for the enquirer's needs or comprehension. Following the next section, in which the performance of the computer search is discussed, attention is given to the bases on which this review is carried out, and once determined some may suggest alterations to the computer search to improve its performance.

3. RETRIEVAL PERFORMANCE OF THE COMPUTER SEARCHES

3.1 Failure to retrieve

The overall performance of the computer search was disappointing, because a considerable number of searches were completely unsuccessful. Twenty-three of the 100 searches retrieved no references and a further sixteen retrieved only unsuitable items. This failure rate was particularly high in comparison with the manual searches, since in the final 22 searches, 8 computer searches retrieved no references whilst only 1 of the manual searches retrieved no references. Similarly 14 of the computer searches retrieved no suitable items, whilst only 4 of the manual searches provided nothing suitable.

With hindsight, some of the causes of the failures of the computer search can be isolated. Comparisons with the parallel manual searches in particular, serve to expose some areas of failure and suggest means by which these may be overcome.

3.11 The Number of Profile Fields 'ANDED'

The results of the investigation emphasise that, the computer search is, in general, no better able than the manual search to satisfy the query which is expressed by a combination of a large number of concepts. The most commonly occuring reason for the failure of profiles to retrieve any references was the requirement of a match in four or more profile fields. This was instrumental in the complete failure of at least 16 of the 100 searches, and partial failure of 41 searches. It is possible that the data base contained no suitable items, but the fact that the majority of these searches were satisfied by the manual search indicates that the exhaustivity of indexing, although higher than that of the printed index entries, is not high enough to satisfy profiles of this depth.

The exhaustive profiles occur most commonly because the inexperienced profile compilers expressed every non-trivial term of the query in the profile. This pressure to satisfy every concept in the query is weaker in a manual search, because the searcher knows from experience that the printed index entries are not sufficiently exhaustive for this approach, and he accepts a lower level of match. This relaxation of requirements does not apparently lead to the retrieval of much noise. For example, 5 of the final 22 profiles comprised four or more profile fields. Only one of these profiles, matched automatically, retrieved

a document and this was not relevant, but all the queries were satisfied by the manual search to a precision of 62%, i.e. 62% of the index entries selected were judged to be suitable when their abstracts were consulted.

The most successful computer searches were those which demanded a single concept, e.g. 'auto-closing circuit-breakers'. Queries of this nature may normally be expressed in two or three search fields. This was the situation with 27 of the 35 successful searches. It is also interesting to note that eight of these comprised one of the printed subject index headings qualified by a single adjective and would, therefore, have been quickly satisfied in a manual search.

The frequency of occurrence of the many-termed query is one of the characteristics which distinguishes the queries in the present investigation from the current-awareness profile. Although a current-awareness profile may include a large number of concepts these are often alternatives and not required simultaneously. It is suggested by the results that it is not natural for a profile compiler to write a profile which is less specific than the query, but that he must be encouraged to do so in the interests of retrieval. A number of considerations for minimising the number of neccessary search fields are listed in Appendix 6.

3.12 Exhaustivity Within Profile Fields

Whilst increasing the number of 'ANDED' fields in a profile reduces the probability of retrieval, increasing the number of alternative terms within a field increases the probability of success. These alternative terms would commonly be synonyms, or related terms. For example, a profile of a query about 'metal foils' has improved probability of success if a list of named metals is added to the field 'metal', e.g.:

Field 1	Field 2
Metal	Foil
Aluminium	Film
Silver	
Gold	
Copper	
Iron	
Nickel	

This example is straight-forward but there are some terms for which synonyms or specific examples are less easily defined. For instance, it is difficult for the profile compiler to construct a comprehensive list of the

'Electric hand tools' which cause interference. The profile compiler must, however, attempt to do so because the absence of control of document indexing at input permits the listing of any tool by name.

The investigation supplied examples of other unsatisfactory search terms, e.g. 'equipment' and 'device'. The indexer customarily describes a piece of equipment by name and not in these general terms. An even greater problem is the use of subjective terms such as 'development' and 'advances' in a profile, e.g. 'developments in oscilloscopes'. There is little possibility of enhancing a profile with these terms. If they are included in the profile then the items retrieved will be of the review type and particular developments will not be retrieved since they are unlikely to contain these terms. If on the other hand, the 'development' apsect of the query is omitted and only 'oscilloscope' used in the profile, the retrieval will be large and will contain a lot of unsuitable items. It is probable that an enquirer who couches his query in terms such as 'advances' or 'development' will be best pleased with a small number of review papers. If, however, he requires a comprehensive search, he would prefer the large and crudely selected output in order that he might make the final subjective judgements himself. It is very important to the profile compiler that he should know what purpose the enquirer will use the reference and have an approximate idea of how many papers he is able to read.

This information about the enquirer's requirements is more important for the retrospective search than for the current-awareness service for which the purpose is normally a watching brief with a practical limit to the output size dependent on the effort the user is willing to spend in following up the items. This effort generally remains constant from week to week.

3.13 Bipartite Queries

Another type of query which made profile compilation difficult was that in which the enquirer asked initially for papers with a braod subject coverage and then expressed interest in a few specific examples, e.g. 'Doppler effect - acoustical, optical, radar - especially re-unscrambling the return signal from various types of interference'.

This dual appraoch basically constitutes two profiles, and again if the enquirer's purpose is known, a more useful profile can be compiled.

3.14 Comparative Terms and Value Ranges

Eight of the queries included terms such as 'high frequency' and 'small scale'. These are terms which are used in the indexing but whose use in a profile is more difficult.

The indexer will not always consider this aspect of a paper sufficiently important for inclusion or he may express it as a numerical range e.g. 3,000 - 30,000KHz. The search system is not equipped to compare the magnitude of numbers. Unless the range in the profile is exactly the same and is expressed in identical units no match is possible. Even should numerical values be converted to the more general terms e.g. 'high frequency' this also can be confusing since high frequency means something different to the audio specialist and radio engineer.

3.2 Retrieval of Inaccurate Items

The deficiencies of the computer search lie not only in failure to retrieve references, but also in the retrieval of references which satisfy the search profile but by no means satisfy the question stated, or the enquirer's need. Precision calculations for the computer show an average value between 40 and 60% depending on the method of calculation.

The reasons for the retrieval of inaccurate items fall broadly into the two categories (i) indexing failures and (ii) profile failures. Those which occurred in this study are summarised in Appendix 7 and some are discussed in the following paragraphs.

3.2.1 Term Semantics and Inter-term Relationships

The problems associated with the meaning of terms and the relationship between them are complex and the distinction between the two is not well-defined. The former are particularly prominent in a system in which the indexing language is minimally controlled. Firstly, the meaning of a term can be highly dependent on the context of the document in which it is used. For example, the term'line' has a different meaning in the context of 'power lines' and'assembly lines'. In many cases the discrepancy is overcome by searching in only a limited part of the data base, but this is not a complete solution. In the same context authors or indexers will use one term in different ways, e.g. in a paper about train control the term 'remote' implies centralisation to some authors, but not to all.

Secondly, the consequence of searching the free-indexing fields as a string of unrelated terms instead of a series of distinct phrases is marked. The relationship between the phrases is not expressed and they frequently refer to completely different parts of the document described. Two terms each retrieved from different unrelated phrases lead to inaccuracies. A query concerning 'current' probes in the measurement of current, with the profile:

Field 1	Field 2
probe	current

retrieved a document decribed by 'Current controlled sources; Power supply circuit; Range switching; Temperature probes; Bridge circuits.'

In more extreme cases one indexing phrase may describe a very minor part of the document. Terms such as 'control' and 'circuit' are particularly subject to inaccuracies of this type. This was a common occurance, producing some noise in 23 of the searches made. It may sometimes be prevented by the co-ordination of search terms in a single profile field. For instance, in the example given above 'current proble' would constitute a single field and must, therefore, retrieve only those records in which the two words occur consecutively. In practice the output of the profiles in which this device was used was very small.

The preceding failures concerned the absence of an expressed relationship between terms, but the situation also arises in which the relationship between terms in a single phrase is expressed in the document indexing: usually by a preposition such as 'in or 'on'. The Boolean operators cannot distinguish this difference. Hence, in a search for papers about 'patient care after shock' a number of papers were retrieved which discussed 'shock occuring during patient care'. These inaccuracies occured less often in the investigation, but were noted on six searches. It is, however, sometimes difficult to distinguish between two types of failure.

3.2.2. Exhaustivity of Search Fields

The retrieval of items which do not accurately match the query also occurs when concepts of the query are omitted from the profile. This happens if a concept is difficult to express, for example, 'applications' 'developments' and 'high frequency'; all concepts which have already been discussed.

More commonly, any omission resulted from misunderstanding between the profile compiler and the information officer. This was a consequence of the artifical test situation in which in view of the large number of questions involved, it was necessary for the information officer to delegate the compilation of the profiles. This would not be the case if a continuing service was offered, but it serves to illustrate how easily an important detail of a query may be lost at the expense of the search results. In Appendix 10 therefore, a check list of necessary and useful enquiry information has been compiled.

4. SELECTION OF SUITABLE ITEMS FROM THE COMPUTER OUTPUT

The number of references retrieved in a computer search can vary from a single item to some hundreds depending on the breadth of the subject matter. In the latter case the search might be thought to be 'too successful' and must be screened to yield a practical number of references. The lists of references sent to the enquirer as a result of the manual searches indicate that 6 to 12 is, with a few exceptions, the number generally accepted as practical. The rigour with which the screening is carried out will depend on the size of the output, and where the output is very large only a small portion would be reviewed. In the following paragraphs, however, the basic criteria by which this was achieved for both the manual and computer searches are detailed.

4.1 Selection Criteria

The language of the paper is the most common criteria used for screening, and where a number of references in English are available they are often selected in preference to foreign language ones.

The papers selected may represent a range of the more general and specific aspects of the subject matter and a review paper may be included together with a variety of applications, techniques or devices. This many-sided approach is necessary if, as frequently happens, neither the enquirer's purpose or level of familiarity with the subject matter is known. If the enquirer's familiarity with the subject is known, then specialised papers, perhaps or papers with a particular slant will be rejected. Other criteria used slightly less frequently were the age of records, the country or organisation of origin, the standing of journals in which they are published or the type of publication, e.g. report, thesis, etc. All these affect how readily the full text may be obtained, although this is not the only reason for their use.

It is, incidentally, these criteria of selection which represent another distinction between retrospective search and the current-awareness service. Although they are applied to some SDI profiles at the user's request, the normal service provides a less restricted output.

These criteria are listed in Appendix 8 and also constitute much of Appendix 10 which acts as an aide memoire for ensuring that the profile compiler has the maximum possible description of the enquirer's requirements. In Section 6, which suggests some means of enhancing the computer search, the application of this information is discussed.

5. ADVANTAGES OF THE INTERACTION OF MANUAL SEARCHES

The previous section discussed the action taken with large scale computer output. In this section the situation at the other extreme is considered. The output might be very small, even non-existent, and cannot be improved except by rewriting the profile and repeating the search. In a manual search the adjustment of profile or requirements is carried out continuously throughout the search in the light of results. It is instructive therefore to study the results of this interaction in anticipation that some elements of the interaction may be simulated in an automatic search.

The selection of items in a manual search is not restricted to the entries which contain all the significant words of the query. The user will select entries in which a relevant paper is expressed irrespective of the terms used.

It has already been noted that the free-index term assignment is insufficiently exhaustive to satisfy the profile with many concepts. The headings and modifier lines in the printed subject indexes are less exhaustive and so have no greater probability of satisfying these queries. In order to avoid a non-productive search the user relaxes the conditions and this is commonly achieved by selecting entries which match only some of the concepts. This approach was adopted in eight of the final 22 searches with little resultant noise, since a precision of 62% was recorded.

An alternative approach adopted in four searches was the selection of index entries describing associated techniques or subjects, or documents indexed by attributes which differ from those of the query. They are selected in anticipation that the paper will prove to be relevant when more detail is available.

It might have been expected that the user would select index entries which include terms more specific than those of the query, but this was an uncommon occurance, which occured only three times in the whole investigation.

It can be concluded that the variation in the manually selected entries from the query statement generally lies in the reduction of the number of concepts sought, and we can consider the application of this result to the improvement of the computer searches.

6. IMPROVING THE PERFORMANCE OF THE COMPUTER SEARCH

The presentation of results and the discussion of the failures of the computer search have highlighted a number of aspects of the performance which might be improved. For example, by performing some of the operations in the output review procedure or in accommodating the query with many concepts. In the following paragraphs some suggestions are made for the implementation of such improvements.

6.1 Variation in profile field exhaustivity

Although the 3i's search system includes a facility for the weighting of search terms, this was not utilised in this investigation. Each profiler compiled only four profiles each month and since the majority of them had no previous experience, it was not reasonable to expect that they would master two alternative search techniques. This facility does, however, permit a more suitable search strategy for those queries which comprised a large number of concepts. Appropriate weights are assigned to each term or field and the document output is ranked by the sum of the weights of the matching terms. The items at the top of the list then satisfy the query completely and those lower down satisfy it only in part. If then no items satisfy the query, some less exact ones will still be retrieved. This result can be achieved in its simplest form by assigning each field a weight of 1.

The more complex problems encountered, of the overlap of concepts between fields, can also be overcome by the use of weighting in preference to the Boolean formulation. Although it should be pointed out that the problems were due to the rather restricted Boolean formulation that was allowed and would not have arisen had a more flexible version been available.

Throughout this study although the manual searches were conducted on the basis that the enquirer was given only as many references as he can reasonably cope with, the profiles were not compiled with this in mind. The users' requirements were not always known to the profiler and although the experienced profiler can, with a knowledge of the data base, vary the profile terms and exhaustivity to produce a desired output size, their success in this is severely limited by the 'all or nothing' requirement of the Boolean-based formulation permitted. Again, in this search system the weighting of search terms is one answer to this problem and the output size may be limited either to a specified number of references or by a predetermined threshold of matching weights.

6.2 Language of reference

For the majority of queries, it is essential that the papers retrieved should be written in a language which the enquirer can read or for which he can quickly obtain a translation. This fact has already been recognised by INSPEC and language information has been included in the records on the data base. When an attempt was made to use this information in the investigation, a particular difficulty was exposed. Only foreign languages are entered in the record. If the paper is in English the record field denoting language is generally left blank. Should the enquirer request only English Language papers, as is often the case, it becomes necessary to search for an empty or non-existent field. The search system used did not permit this. The data base processors were obviously not aware of the problem, which would be simply solved by inserting 'English' in the blank language fields before the data is processed.

6.3 Treatment Codes

INSPEC had also anticipated that it would be helpful to be able to retrieve only those papers which treat the subject matter in a particular way, for example a review paper, theoretical treatment or developments - a requirement which has already been discussed in some detail.

For this purpose 'treatment codes' are included in the records where appropriate. This policy had not been adopted when the early records of the archive were indexed. Thus, if a treatment code is necessary to satisfy the logic of a profile, none of the early records will be retrieved, even though they may treat the subject in the required manner. This problem is overcome only if a match on the treatment code is not essential to the retrieval but adds to the final output rank of the documents which include it.

The 'treatment' factor also has a more complicated aspect. The variety of papers which the manual research provides, including general and specialised papers and a selection of applications or techniques, is difficult to simulate. Although a clear definition from the enquirer can greatly improve matters, it does not provide the complete solution.

6.4 Books

The book as a source of general articles was frequently used in the manual search. They are quickly selected from the library shelves and are more likely to cover the well-established subject matter than are journal articles. They also often provide bibliographies. The computer search

does not easily manipulate the book entries. The indexing is very general and is rarely specific even to the extent of chapter headings. A section or chapter in a book, although relevant cannot be retrieved by a profile designed for the retrieval of journal papers and reports. It is questionable whether it is of value to include books in the searchable data base when a different search strategy is required.

6.5 Other search fields

The information officers' assessments stressed a number of other details which are sometimes of interest in a retrospective search. Amongst these are included the date of publications, the country or organisation of origin and the source journal or type of literature. All this information is available on the INSPEC tapes and could with some attention to suitable search strategies, be useful search fields.

6.6 Summary

It can be seen that the majority of features to which the manual searces owe their flexibility and selectivity can be simulated in a computer search by use of more of the indexing fields available on the INSPEC files and by use of more flexible search formulations. A more flexible Boolean logic or weighted search system would solve many of the problems presented by profile compilation in this investigation, but some of the errors will also be overcome when the INSPEC thesaurus, displaying the relationship between the terms is available to the user.

7. COSTS

The average computer processing cost of each automatic search was £23. A further £2 might be added in respect of the profile compiler's effort but this is only a small proportion of the cost. The cost of the computer search is of an order of magnitude higher than the manual search.

The manual searches in the printed index took an average 45 minutes and would therefore be unlikely to represent a cost of more than £2. (The cost of assessing abstracts is ignored for the comparison since it is common to both methods of search).

The effectiveness of the computer search must be high to justify the cost, but in this experiment the complete failure rate by this method was eight times that of the manual search. Although means may be suggested for reducing this failure rate, the noise would be also increased and additional manual selection effort would be required.

8. CONCLUSIONS

The overall performance of the computer search was not good when compared with the search carried out manually. There was a high incidence of completely unsuccessful searches and a number which produced very large outputs, of the order of two or three hundred references.

The cost of the computer search was of an order of magnitude greater than the manual search in this instance and it is known that the price charged was significantly lower than would have been charged for an operational service.

Two or three weeks elapsed between the dispatch of the queries and receipt of the output and although this was due to the distance involved it is unlikely that even an inhouse operation would achieve a turn round on batch processed searches of less than a day. Some difficulty was experienced in the later stages of the investigation in collecting sufficient suitable queries and it could not be expected that enough queries would be recorded in a single day or possibly even a week for batch processing to be economic.

The investigation showed that there is more to a retrospective query than the subject matter. The enquirer does not normally state the use to which he intends to put the information, the languages he is able to read, or the type of literature to which he has ready access, and in written communications he only infrequently indicates the level of treatment he requires in the documents retrieved or any particular bias. Information of this type is extremely important to the user of the printed index or profile compiler, but is often not consciously determined by the information officer taking the query. The presence of two intermediaries between the enquirer and his profile therefore, led to the loss of some of this information, but in an operational service this situation would not arise.

The searches most successfully completed by the computer search are those whose subject matter comprises a single concept and not those combining a series of concepts as might have been anticipated. The indexing in the free-indexing field of the data base was not sufficiently exhaustive to satisfy the most exhaustive profiles. In order to increase the retrieval it is necessary to relax the conditions and select items indexed by only some of the profile terms. This is an approach which is most commonly followed in a manual search to increase the retrieval where printed index entries are insufficiently exhaustive. Unfortunately, an increase in the number of items retrieved almost inevitably means an increase in noise, and greater manual selection effort is required to screen the output to provide the user with the most useful list of references.

Ultimately, a reduction of profile length produces a list of references for screening which is equal in length to the entries under a printed (subject) index heading. The unsatisfactory performance of the long term profiles served to emphasise that care must be taken when a profile is written, that query concepts are not repeated in more than one field.

The failures of the computer search lie mainly in two areas. Firstly, the inability to retrieve references in response to long profiles and, secondly, in the retrieval of inaccurate items as a result of the lack of expressed relationships between terms.

The logical formulations available in the 3i's system displays not only those term-relationship problems typical of Boolean logic, but also problems resulting from the inflexible format. This can be overcome by sophisticated profile compilation or term weighting but neither procedure would be expected of the inexperienced profiler.

Analysis of the manual searches made showed that few intellectual decisions are applied in an effort to increase the number of references retrieved, but much effort and many decisions are made to select the most suitable references from the longer lists of items retrieved. This fact is encouraging for the improvement of the computer search as methods are available for simulating the relaxation of profile requirements and the screening of output.

These operated in batch-mode, could enhance the performance considerably but they are still restricted by their one-off nature. The advantages gained by user system interaction and its consequent profile modification which are so important in the manual search cannot be achieved in batch processed searches but are to be sought in on-line searching.

9. RECOMMENDATIONS FOR FURTHER STUDY

This investigation has shown that the main areas in which further studies are necessary are in methods of satisfying the exhaustive query and in increasing the selective powers of the computer search.

The former will be achieved by investigation of the ranking of output and search logics which do not have the 'all or nothing' limitation of the Boolean statement. The optimum use of the classification codes would also be closely linked with these studies. The classification like the free-indexing is unique to the INSPEC data base and so although some studies have been carried out elsewhere concerning the best approaches to retrospective searching, the results obtained are not necessarily also applicable to the particular form of indexing used by INSPEC. For example, the optimum use of the discrete indexing phrase is not clearly understood.

The study of the optimum use of the treatment and language codes on the INSPEC data base are not subjects for extensive projects but deserve some attention to detail. Satisfying a query for a range of treatments is undoubtedly a problem. Searching numerical ranges and comparative terms again need some investigation.

The usefulness of searching for books and papers by the same profile was brought into doubt by this investigation and would bear further consideration.

USER: _____ DATE 23-6-72 PAGE: 1 OF: 1

PROFILE DESCRIPTION, COMMENTS. RETROSPECTIVE SEARCH EXPERIMENT

LANDING LIGHTS FOR AIRFIELDS

ATTACH CODING FORMS FOR ALL CORRECTIONS – INDICATE DATES OF CORRECTIONS

		CO. NO.	PROF. NO.	DATA BASE (REPORT)	DATA BASE (ACCTG)	T S R	PROF. UNITS	THRESH.	COMMENTS
O	I	x 999	x x x	x IEE	x	x	x	x	
F	I	T, A,	F, S,	C, H, I, R, B					

		PARAM	LOGIC	SIGN	WGT	MODE	TYPE	TERM VALUE (TO CONTINUE ON NEXT CARD REPLACE S/ BY -/)
1	S/	1	C				S	SALSG
2	S/						S	SBL4G
3	S/	2	C			T*R	T	AIRFIELD
4	S/					T*H	T	AIRFIELD
5	S/					T*R	T	AIRPORT
6	S/					T*H	T	AIRPORT
7	S/					T*R	T	AERODROME
8	S/					T*H	T	AERODROME
9	S/					T*R	T	AIR-PORT
10	S/					T*H	T	AIR-PORT
11	S/					T*R	T	AIR-FIELD
12	S/					T*H	T	AIR-FIELD
13	S/	3	C			T*R	T	LIGHT
14	S/					T*H	T	LIGHT
15	S/					T*R	T	LAMP
16	S/					T*H	T	LAMP
17	S/							
18	S/							
19	S/							
20	S/							

PROFILE CODING SHEET

APPENDIX 1

COMPUTER SEARCH OUTPUT

DOCUMENT NO. PROFILE THR ACCUM WGT DATA BASE DATE CARD 1
200838 999/076 0001 00003 IEE 13JUL72 OF 2

TITLE: The development of an airfield taxiway light

AUTHOR: Richards, M.J.,

AFFILIATION: GEC, Wembley, England,

SOURCE: (J) Lighting Res. Technol. (GB) VOL.2, NO.3 (156-60) 1970

CLASSIFICATION: R4630

SUBJECT HEADINGS: Lighting ||of|| airfield taxiways, developments and specification aspects) /Lamps (for airfield taxiway lighting, specification requirements) /Air traffic control (lighting requirements, runway and taxiway illumination levels)/

SCIENCE ABSTRACTS REF. NO./TREATMENT: B7039836

ABSTRACT: Developments in airfield operational procedures demand,

DOCUMENT NO. PROFILE THR ACCUM WGT DATA BASE DATE CARD 2
200838 999/076 0001 00003 IEE 13JUL72 OF 2

among other things, an ability to taxi even the largest aircraft safely from the runway to the unloading area under the worst conditions of visibility. This has produced a demand for high-intensity taxiway lights to meet a stringent Ministry of Technology specification. The paper describes the development of such a fitting based on the use of tungsten halogen lamps and incorporating a new, very compact and efficient optic. Prototype fittings, now undergoing airfield tests, have given, to date, a very satisfactory performance

PROJECT AB RETRIEVAL CENTRE

MANUAL SEARCH RECORD. QUESTION NO. 83

Initial Statement of Enquiry

1. Original letter etc.

   By phone

2. Statement of query

   Papers on loss angle measurement - testing of insulation etc of motors, machines and transformers - extrapolation of results to predict breakdown

3. Enquirer

4. Enquirer's background

5. Restrictions eg language, size of output, treatment, date, bibliographic form. (These restrictions may be applied by the information officer in his experience on behalf of the enquirer). These are not restrictions or relaxations imposed by the computer during the search. see II.

6. Reason for selection of query for test

   Very suitable

7. Query received by KB search conducted by KB
   Date. 7/7/72

Question No.
Search Procedure

3. Volumes of Science Abstracts searches

Electrical and Electronics Abstracts 1969-1971

9. Index headings used.

loss angle measurement

10. List of abstract numbers selected from modifier lines. Items are assigned to one of the following categories. (A = accurate S = Suitable i.e. sendable N = Not).

i)	S (and NA) = S	Accept	All items in categories
ii)	A and S	Accept	(i) - (iii) are logged
iii)	A (and NS) = A	Reject	in column I (A,S,NS)
iv)	NA	Reject	

Heading				Heading			
loss angle measurement							
Abs No.	I	II	Reason III	Abs No	I	II	Reason III
00092	A+S	NA					
4534	S	A+S	✓				
5625	S	NS	In German				
1205	S	A+S	✓				
15731	S	NS	In French				
19645	S	NA	Abstract only				
39013	A+S	A+S	✓				
37389	A+S	A+S	✓				
22236	S	NS	Polish				
37365	S	A+S	✓				
3813	S	NA					
16973	S	NA	Striplines				
808	S	NS	Rumanian				
12387	S	A+S	✓				
5526	S	NS	Italian				

11. Restrictions or relaxations imposed by the searcher.

Preferably in English, Not patents

12. Abstracts of accepted items from 10 are examined and each item is assigned to one of the 4 categories on the basis of the information in the abstract in col II. Give reasons (III)

13. Items found in Science Abstracts in any other way than directly through the index and how.

14. Items from other services.

15. Details sent to enquirer - copy attached

16. Time for looking at modifies lines 30 mins

QUESTIONS	Computer Search			Manual Search		
	No. of references retrieved (a)	No.selected by information officer (b)	Precision (a)/(b)%	No. of references retrieved (a)	No.selected by information officer (b)	Precision (a)/(b) %
90. Use of computers for the calculation of ground potential distribution in high and medium voltage substations with a ground grid and rods.	5	0	0	8	1	13
91. Capacitance braking for 3-phase a.c. motors up to about 10 HP	0	0	0	9	6	67
92. Installation of standby generator sets and connection to feeders and isolation.	0	0	o/o	14	11	79
93. Impedance of coils.	1	0	0	9	9	100
94. Partial discharges in dielectrics			100 ++	15	14	93
95. Fault diagnosis in combinational networks	0	0	0	10	10	100
96. Divided-winding-rotor generation.	6	6	100	4	4	100
97. Design of coils windings etc. for transformers and motors.	56	13	23	14	8	57
98. Precautions taken in laboratory wiring.	0	0	o/o	10	2	20
100. Future uses of electrical heating.	184	22	12	9	8	89

APPENDIX 4

QUESTIONS	Computer Search			Manual Search		
	No.of references retrieved (a)	No.selected by information officer (b)	Precision (a)/(b)%	No.of references retrieved (a)	No.selected by information officer (b)	Precision (a)/(b)%
64. Use of thyristors in speed control of motors in the Apollo Space Programme.	0	0	o/o	1	0	0
67. Prosthetics for knee and hip amputations i.e hinge-joints for ex-articulation.	10	8	80	7	4	56
76. Landing lights for airfields - design of lamps, installation and cabling	7	7	100	11	8	73
78. Continental practice for electricity meters - designs and use of polyphase meters.	2	0	0	9	1	11
81. Bibliography on synchronous induction motors, rotor excited from static sources	0	0	o/o	21	5	48
82. I-diagrams - the measurement of noise and interference on digital waveforms by oscilloscope methods.	35	0	0	3	0	0
83. Loss angle measurment, testing of insulation etc. of motors/machines/ transformers and extrapolation of results to predict breakdown	4	0	0	15	6	40
84. Technological trends in electronics e.g. LSI, IC's, computers, communications, data transmission facsimile and components.	83	67	37	6	3	50
85. Jointing of undersea cables under high pressure	0	0	o/o	7	4	56
86. Speed control of fan motors (split phase capacitor start) using thyristors.	9	5	56	18	8	44
88. Manufacture of very small high speed (over 10,000 rpm.) d.c. motors running on 4-6 volts using a wet cell charagable battery.	0	0	o/o	14	6	43
89. Automatic synchronisation of generators.	3	1	33	0	0	o/o

Retrieval and precision figures

Questions		Computer search retrieval		
		No. of references retrieved (a)	No. selected by information officer (b)	Precision a/b %
1.	Medical thermography	38	30	79
2.	Small-scale applications for linear motors	0	0	o/o
3.	Electronic speed control of gas and steam turbines	1	1	100
4.	Centralised train control	38	20	53
5.	Current probes used in measurement of current	73	60	82
6.	V.L.F. undersea propagation	-	-	- +
7.	Measurement of iron losses using wattmeter and 'ghost coil'	2	1	50
8.	F.E.T. oscillator circuits	19	12	63
9.	Cable balancing using a compute	24	0	0
10.	Antilock brakes	20	18	90
11.	Dopplereffect - acoustical, optical and radar - especially re-unscrambling the return signal from various types of interference.	-	-	- +
12.	Insulators for overhead lines and transformers	3	0	0
13.	Quadrature amplitude modulation	-	-	- +
14.	Vehicle guidance by signals received from underground cables	0	0	o/o
15.	Demagnetizing devices			20 ++
16.	Medical ultrasonics: particularly narrow beamwidths transducers and techniques for measurement of polar diagrams	-	-	- +
17.	Design of 3-phase to single-phase static balances, especially applicable to single-phase furnaces	-	-	- +
18.	Speed-control of polyphase induction motors using triacs	0	0	0
19.	Protective multiple earthing	10	10	100
20.	Autoclosing circuit-breakers	-	-	- +

	Computer search retrieval		
	No. of references retrieved (a)	No. selected by information officer (b)	Precision a/b %
21. Floating point hardware	30	24	80
22. Modular numbers	3	2	67
23. Live-line washing of insulators	1	1	100
24. Computer-controlled traffic lights	8	7	88
25. Engineering technician education	5	5	100
26. Electrical installation in nuclear power station buildings	1	0	0
27. Lightning protection of buildings using the reinforcement in reinforced concrete	0	0	o/o
28. Eddy-current couplings	-	-	-
29. Defibrillation after electric shock	1	0	0
30. Boxcar integrators	5	4	80
31. Dimming of fluorescent lamps	1	0	0
32. Suppression of interference from portable electric tools	0	0	o/o
33. Solid state control of Ward Leonard machines	2	2	100
34. Sodium-filled cables	9	7	78
35. Gasturbine alternators	14	13	93
36. Peak load lopping using diesel stand-by generators	11	7	64
37. Metal foils for electrolytic capacitors	0	0	%
38. Electric actuators			60 ++
39. 'Lincompex' radio-telephony system	4	4	100
40. Telegraph message switching with electronic stores	1	1	100
41. Power transformer audio noise reduction using antiphase noise techniques	0	0	o/o
42. Speed control of induction motors using thyristors	22	21	96
43. Synthetic testing of circuit breakers	28	27	96
44. Description of the Decca 'Navigator' System of radio-navigation	2	2	100
45. Lead/acid battery maintenance during periods of non-use	2	1	50
46. Power system interconnection			5 ++
47. Plasma arc cutting of steels	5	4	80

	Computer search retrieval		
	No. of references retrieved (a)	No. selected by information officer (b)	Precision a/b %
48. Long term reliability of discrete-component electronic semiconductor equip. (cf 'bath-tub' characteristics of vacuum tubes)	5	4	80
49. Load-frequency control in interconnected power systems	6	6	100
50. Developments in oscilloscopes	5	1	20
51. London's power supply system	3	2	67
52. Anglo-French hvdc link	0	0	0
53. Electrical power development in Spain	0	0	o/o
54. Push-button telephones - their design and uses especially in data transmission	9	8	89
55. Lifts in high-rise buildings	6	5	83
56. Ultrasonic atomisers for the production of aerosols	3	0	0
57. Surge protection (diverters) for electrical systems	2	2	100
58. Computer typesetting	20	20	100
59. Conduction in graphite-impregnated rubbers and plastics	43	9	21
60. Computer programs or software for retrieval from documentation data bases	18	9	50
61. Invalid carriages for disabled persons	2	0	0
62. Electrical parameters of earth soil etc. upper layers of interest re radiowave propogation, especially maps of G.B. and Germany			0 ++
63. Ripple control of power systems and domestic equipment, lighting etc.	0	0	o/o
65. Use of electronics in the fishing industry e.g. data processing and low light T.V.	0	0	o/o
66. Electrical services in hospitals	18	10	56
68. Substation earthing	11	6	55
69. Standardisation in video recording	11	5	46
70. General papers on stepper motors	0	0	o/o

	Computer search retrieval		
	No. of references retrieved (a)	No. selected by information officer (b)	Precision a/b %
71. Constant voltage transformers	5	5	100
72. Manufacture of incandescent and fluorescent lamps	8	6	75
73. Geothermal power supplies	9	8	89
74. Use of wood poles for carrying ehu transmission lines	0	0	o/o
75. Mass soldering methods used in the production of printed circuit boards especially dip and flow methods	0	0	o/o
77. On-line computers in traffic control (systems in Glasgow, London and Toyko)			
79. Autopilots for ships	13	9	69
80. Automatic control of vulcanisers	0	0	o/o
87. Material properties of GaAs	-	-	-
99. Thyristor controlled d.c. machines	42	19	45

\+ The output of this search was seconded for a seperate project or invalidated by a procedural error.

++ Where the output was of the order of 200 or 300 references the precision was estimated on the basis of a random selection from the output.

APPENDIX 5

Summary of reasons for retrieval failures in the computer search

1. Profile contributed failures

   (i) Too many profile fields

   (ii) Repetition of concepts in different fields

   (iii) Use of 'vague terms' e.g. technological trends

   (iv) Use of comparative terms and value ranges e.g. small scale.

   (v) Insufficient synonyms

   (vi) Misuse of truncation and exact matching

   (vii) Spelling errors

2. Document indexing

   (i) Superficial indexing of books and reviews

   (ii) Indexing by different attributes

Detailed discussion of the exhaustivity of profile fields

Exhaustive profiles as a cause of profile failure.

A common factor in many of the search profiles which failed to retrieve was a larger number of profile fields i.e. four or more fields which must be satisfied. As the number of profile fields increases the probability of effecting a match decreases substantially and if any references are to be retrieved the number of fields must be kept to a minimum.

Minimising the number of profile fields.

A more detailed consideration of the longer profiles revealed that many were unnecessarily long. The occured most commonly where two fields contained synonymous or hierarchically related terms. For example, a query concerning the Lincompex radio-navigation system was expressed by a profile which included the two distinct fields 'Lincompex' and 'radio-navigation'. Since Lincompex is itself a radio-navigation system it is unnecessary to include the latter profile field and the shorter profile would not lack precision. Duplication of this type is more easily avoided when a controlled language is used mainly because the profile compiler has to hand a structured list of terms. A similar structured list would be most useful for reducing these errors. Where a free-language base is being searched.

A more complex variation on this problem arises when a single profile field conveys a concept which is already expressed in the combination of two or more of the remaining fields. Again this is an occurance particular to the free language vocabulary which could be reduced if the compiler has access to a structured display. In this system however, the compiler, although aware of the problem, may be unable to correct it because of the restricted nature of the Boolean logic and very complex profiles or a more flexible logic formulation are the only solutions.

The INSPEC database provides the user with a choice of search fields whose structure differs but whose subject information content overlaps, e.g. classification codes and free-indexing terms. Experience has shown that precision in retrieval is achieved by utilising the calssification codes to select a subset of the data base in which the free indexed profile terms are sought. There are however, occasions when the classification codes can play a more positive role in the profile compilation. In a number of queries some or all of the concepts are precisely expressed in a classification code and further definition by profile terms is unnecessary. For example, the concept of 'speed control' is defined by the code 73.22 (Control of variables/ speed) and the two additional profile fields 'speed' and 'control'

lengthen the profile to no purpose. Where two overlapping search fields are available it is important that their use is complementary.

One further contributory factor in the longer profiles was the use of profile terms which are implied by the nature of the data base searched. For example, it might reasonably be expected that devices described in records retrieved from Section B of the INSPEC data base would be either electrical or electronic and not mechanical. To specify these terms is therefore unnecessary. When a single data base is used this point is trival but as services are extended to cover a wider variety of the data bases available it will become more important.

APPENDIX 7

Summary of reasons for retrieval of non-relevant items from computer search.

1. In the statement of the query

   (i) Inadequately expressed

2. In the profile

   (i) A term has a different meaning in the context of the document retrieved

   (ii) A term has a different meaning or function in the context of the indexing phrase from which it is retrieved

   (iii) The Boolean operators do not adequately express the relationships between terms

   (iv) Not all the concepts of the query are expressed in the profile

   (v) A spelling error

   (vi) An error in a classification code

   (vii) Truncated terms retrieve unexpected words

3. In the document indexing

   (i) Spelling errors

   (ii) Exhaustive indexing

   (iii) Misrepresentation of the document

APPENDIX 8

A Summary of the Criteria Used by the Information Officer in the Selection of Items Suitable for the Enquirer

1. In screening the papers which accurately satisfy the subject matter of the enquiry.
(These criteria are drawn from the analysis of the information officer's assessment of the output of both the computer and manual searches)

   (i) Language
   (ii) Level of treatment, e.g. studenttext or very advanced
   (iii) Bias, e.g. applications, equipment or theory
   (iv) Age of paper
   (v) Country of origin
   (vi) Organisation in which the work was carried out
   (vii) Bibliographic form, e.g. journal article, report or book
   (viii) The availability of the full text
   (ix) The total number of items retrieved
   (x) Standing of the journal cited

2. In selecting papers which are suitable for the enquirer but do not accurately satisfy the subject matter of the query.
(These criteria are drawn from the analysis of the manual searches)

   (i) Reduction in the exhaustivity of the query
   (ii) Associated subjects or techniques
   (iii) Indexing breakdown by different attributes
   (iv) Synonymous terms
   (v) Reduced or increased specificity of terms

Suggested criteria for queries best suited to a batch-processed computer search or a manual search

1. For a batch-processed computer search

   (i) If an exhaustive search for the compilation of a bibliography is required
   (ii) If the limited time span of the data base is sufficient
   (iii) If a single concrete term is sought which possibly cuts across the classification and is not a subject-index heading.

2. For a manual search

   (i) If the query can be expressed by a subject index heading and a single qualifier.
   (ii) A query which is subjective in nature e.g.'recent advances'
   (iii) A query which implies a large number of alternatives which cannot easily be listed e.g. 'Electric hand tools'.
   (iv) A query for basic information which are best answered by books.
   (v) Where range data is required e.g. over 10,000 rpm, small scale.

Useful query information for successful profile compilation

(i) Subject matter - clearly defined

(ii) The use which is to be made of the references, e.g. to write a review for a popular magazine or to compile a complete bibliography.

(iii) An approximation of the size of output which would best please the enquirer.

(iv) The level of treatment which the enquirer requires or can understand, e.g. student text or highly advanced.

(v) Any special applications or bias of treatment, e.g. theory or equipment.

(vi) Languages which the enquirer can read or readily have translated.

(vii) Details of the approximate period when the work in the subject field was carried out.

(viii) Any restrictions on the dates of publications retrieved.

(ix) Preferences for country or organisation of origin of items.

(x) If time is an important factor, the journals and type of literature from the full-text is available are important.

Guidelines for tape use and profile compilation

A. Notes on the INSPEC data base

(i) English language items are not usually explicitly stated as such. If searches are required on this data a search for an empty field is necessary. Alternatively the field may be filled automatically before processing.

(ii) Treatment codes were not included on the records on the early tapes. Treatment codes must not be 'necessary' to the retrieval if documents from the older tapes are desired.

(iii) The indexing of the data base does not support exhaustive profiles with many 'necessary' fields. A match on four or more fields is unusual.

(iv) The indexing of book material is not always sufficiently detailed for books and journal papers to be sought using the same profile.

(v) The use of multiword profile terms does not generally produce high recall.

B. Notes on profile construction

(i) It is essential that the profile compiler has such information about the enquiry as is necessary (Appendix 9)

(ii) The data base must have a suitable subject coverage and cover a suitable time period for the query e.g. the majority of work in a particular field may have to be carried out before the commencement of the data base.

(iii) Care should be taken that a query which represents two approaches to a problem is expressed as two profiles or separate parts of a profile.

(iv) Subjective profile terms such as 'developments' and 'applications' are rarely successful.

(v) The use of comparative terms e.g. vhf and numerical ranges in a profile is not usually successful.

(vi) Care should be taken that a profile does not require a match on a single concept in more than one field.