A Comparison Between Manual and Automatic Indexing Methods

G. Salton

Technical Report
No. 68-11
March 1968

Department of Computer Science
Cornell University, Ithaca, N. Y.

A Comparison Between Manual and Automatic Indexing Methods

G. Salton*

Abstract

The effectiveness of conventional document indexing is compared with that achievable by fully-automatic text processing methods. Evaluation results are given for a comparison between the MEDLARS search system used at the National Library of Medicine, and the experimental SMART system, and conclusions are reached concerning the design of future automatic information systems.

1. Introduction

The design and operations of large-scale information systems has become of concern to an ever-increasing segment of the scientific and professional world. Furthermore, as the amount and complexity of the available information has continued to grow, the use of mechanized or partly mechanized procedures for various information storage and retrieval tasks has also become more widespread. As a result, a number of large information systems are now in operation in which at least the search operations — that is, the comparison of incoming search requests with stored information — is carried out automatically. Typical examples in the United States are the NASA Scientific and Technical Information Facility, and the MEDLARS system at the National Library of Medicine.

*Department of Computer Science, Cornell University, Ithaca, N. Y.

This study was supported in part by the National Science Foundation under grant GN-495.

While these operational information systems are thus able rapidly to search vast storage files, often containing many hundreds of thousands of items, most of the operations other than the search itself are performed manually with the help of human experts. In particular, all the content analysis and indexing operations, leading to the assignment of suitably chosen combinations of index terms to the stored documents and to incoming search requests are normally performed by specialists who know the given subject area, as well as the performance characteristics of the retrieval environment within which they operate.

Many of the information systems which base their operations on manual indexing but largely automatic search methods are quite successful in isolating, from the large mass of largely irrelevant stored material, many of the items which prove pertinent to the users' information needs. Nevertheless, the feeling that manual systems and procedures should be replaced by suitably chosen automatic methods has continued to grow, and a number of fully-automatic information storage and retrieval systems have been designed and put into operation, at least on an experimental basis. The SMART system represents one such effort to replace the intellectual indexing by sophisticated automatic text analysis procedures, and thereby to produce a retrieval environment in which all document and query handling procedures are performed automatically [1,2,3].

In the next few paragraphs, some of the evaluation measures that have been widely used to determine the effectiveness of information systems are introduced, and typical evaluation results obtained with the SMART system are given. Thereafter, the design of the SMART-MEDLARS test is examined, and evaluation results are given for the comparison between

SMART and MEDLARS searches, using a variety of different analysis and search methods. Suggestions are made for improving the performance of presently operating information systems, and for the design of future automatic retrieval services.

2. The Evaluation of Information Systems

Many different criteria may suggest themselves for measuring the performance of an information system. In the evaluation work carried out with the SMART system, the effectiveness of an information system is assumed to depend on its ability to satisfy the users' information needs by retrieving wanted material, while rejecting unwanted items. Two measures have been widely used for this purpose, known as <u>recall</u> and <u>precision</u>, and representing respectively the proportion of relevant material actually retrieved, and the proportion of retrieved material actually relevant [4,5]. (Ideally, all relevant items should be retrieved, while at the same time, all nonrelevant items should be rejected, as reflected by perfect recall and precision values equal to 1).

It should be noted that both the recall and precision figures achievable by a given system are adjustable, in the sense that a relaxation of the search conditions often leads to high recall, while a tightening of the search criteria leads to high precision. Unhappily, experience has shown that <u>on the average</u> recall and precision tend to vary inversely since the retrieval of more relevant items normally also leads to the retrieval of more irrelevant ones. In practice, a compromise is usually made, and a performance level is chosen such that much of the relevant material is retrieved, while the number of nonrelevant items which are also retrieved is kept within tolerable limits.

In the SMART evaluation system, these various possible operating ranges are taken into account by computing for each search request, and each processing method a variety of different statistics related to recall and precision. Specifically, four global statistics are generated, known as rank recall, log precision, normalized recall, and normalized precision respectively, as well as ten local statistics, consisting of the standard precision at ten different recall levels. The global statistics are used to represent the overall performance of a given search, whereas the local statistics furnish individual recall-precision pairs for specific operating ranges of the system. Paired comparisons are normally presented, consisting of the average performance over many search requests of two given search and retrieval systems [5].

One of the document collections used for evaluation purposes with the SMART system over the last few years is the set of 200 documents and 42 search requests in the filed of aerodynamics used earlier as part of the well-known Aslib-Cranfield experiments [4]. This collection is attractive for test purposes since a number of actual user queries were available, as well as sets of relevance judgments obtained from the scientists constituting the user population. Furthermore, English abstracts were furnished with each document, and it thus became possible to compare the effectiveness of the conventional retrieval operations based on a matching of the index term sets — manually assigned by trained indexers at Cranfield — with the performance of the fully-automatic language processing devices based on the manipulation of document abstracts, used by the SMART programs. Such a comparison could then produce evidence to indicate whether document identifiers automatically generated by language analysis methods, such as suffix cut-off procedures, thesaurus look-up, phrase generation methods, statistical term associations, syntactic analysis, and others, would perform equally as well as manually assigned index terms.

A typical comparison between the Cranfield indexing, and an automatic word stem matching process based on a matching of weighted word stems extracted from document abstracts and search requests, respectively, is shown in Fig. 1, averaged over the 42 Cranfield queries. The recall-precision graph of Fig. 1(a), and the corresponding tables of Fig. 1(b), indicate that the manual indexing is slightly superior to the simple automatic word-stem process. However, the statistical significance computations, included in Fig. 1(c), show that the differences in performance between the two systems are not significant. Specifically, each of the values shown in Fig. 1(c) represents the probability — computed by using either a standard t-test, or a sign test — that if the performance of the two systems (manual indexing and automatic word-stem match) were, in fact, equally high, then a test value as large as the one actually observed would occur in practice [5]. A probability of 0.05 is usually taken as an upper bound in judging whether a deviation in test values is significant or not. The probability values included in Fig. 1(c) are seen to be much higher than 0.05, and the assumption that the two systems are approximately comparable in effectiveness cannot safely be rejected.

The results of the SMART-Cranfield comparisons seem to indicate that even simple automatic text analysis procedures do not necessarily produce retrieval results which are much inferior to those obtained in a system based on manual indexing. In fact, the benefits of the index language control supplied in the conventional retrieval situation by the human indexers appears to be balanced by a deeper and more complex type of analysis available in the automatic environment, including selective term weighting and the use of relatively large sections of text to insure a high degree of indexing exhaustivity.

While the results of the SMART-Cranfield test are in line with many other evaluation figures obtained by SMART with different document collections in other subject fields [5], the test has nevertheless been criticized by some writers. In particular, it has been claimed that [6]:

a) the use of the standard recall-precision measures is questionable, since other possible criteria (cost, waiting time, etc.) are disregarded;

b) the relevance of a document with respect to a search query is not a stable criterion but varies with the user population, thus presumably producing different evaluation results for different sets of users;

c) the experimental controls used to identify the Cranfield user population, the query set, and the sets of relevance judgments may have been deficient;

d) the sources of variation affecting systems performance are not pinpointed, and no indication is given to permit a generalization of the test results to large, operational situations.

Certain recent experiments appear to indicate that some of these objections may be groundless — for example, different user populations seem to agree on the relative <u>ordering</u> of a set of documents in decreasing order of relevance with respect to a search request, thereby producing constant recall and precision values [7]. However, further comparisons between manual and automatic indexing systems are certainly of interest. The experiments carried out with a small subset of the MEDLARS collection were undertaken in an attempt to obtain further evidence in the ongoing comparison of conventional and automatic information systems.

3. The Test Design

A) The MEDLARS Evaluation Study

The SMART-MEDLARS experiments to be described are based on a small portion of a much larger systems evaluation study undertaken over the last few years within the National Library of Medicine [8,9]. In this larger study, 302 search requests actually processed by MEDLARS were carefully chosen to reflect both a stratified sample of the MEDLARS user population, and a representative proportion of the subject fields covered by MEDLARS. For these 302 searches, the help of the users was enlisted in order to obtain careful value judgments, made on a sample of the search output for each query. Specifically, a <u>precision base</u> (PB) was constructed by judging for relevance a sample of the documents retrieved by MEDLARS in response to each query; similarly, a <u>recall base</u> (RB) was obtained by taking documents identified in advance as being relevant to the query, plus documents retrieved and judged for relevance during the test. The recall and precision base documents were then used to compute for each search the recall and precision actually achieved during the search of the MEDLARS collection.

While the design of the complete in-house MEDLARS test cannot be covered here, it is of interest to examine briefly some of the principal results: The overall average recall figure for MEDLARS was found to be approximately 0.58, while the overall precision was 0.50, thus indicating that, on the average, a typical search would retrieve almost sixty percent of the relevant material included in the collection, while only about half of the documents handed to the user in response to a search request would be nonrelevant. If one considers that the MEDLARS collection consists of a half-million documents, and that only a few hundred will, on the average, prove relevant to a given query, it is seen that the search system consistently and properly rejects many hundreds of thousands of nonrelevant items which the user obviously does not care to see, while retrieving, at the same time, a large proportion of the useful items.

The operating ranges for the present MEDLARS system are shown in the recall-precision graph of Fig. 2. In practice, the system operates in the center of the curve, since that is the area where neither recall nor precision are unreasonably low. Other operating areas are, of course, possible by sliding up and down the curve of Fig. 2. However, most users are not likely to prefer either the high precision — low recall or the low precision — high recall ends, particularly since the all points within easy reach of the presently implemented system are quite far away from the ideal operating range in the upper right-hand corner of the curve.

An indication of the recall and precision failures identified during the complete MEDLARS test is given in Table 1. It is seen that over thirty percent of both the recall and the precision failures are due to the fact that the manual query formulation does not adequately reflect the real user need. In addition, the indexing language in use produces many precision failures, and the document indexing is responsible for many recall failures. Finally, the lack of communication between user and system personnel during the search also causes a large number of errors.

A comparison of these test results with those applicable to the SMART runs is made following the exposition of the test design actually used.

B) Design of the SMART Test

For a variety of operational reasons, it was necessary to restrict the SMART tests to a small subset of the total MEDLARS test environment (300 queries, several thousand recall and precision base items, and over half-a-million documents to be searched). Specifically, eighteen queries were obtained from the National Library of Medicine, together with 273 of their associated RB and PB documents. The 273 documents actually used were chosen as follows:

Total documents evaluated by MEDLARS for the 18 queries in SMART subcollection (including 149 RB and 369 PB items)	518
Number unusable for SMART experiment because abstract or summary was not easily available	245
Total documents used in SMART subcollection	273

For the remaining 18 queries and 273 documents, the English abstracts were keypunched and the SMART runs were carried out in accordance with the standard SMART methods [1,2,3,5].

In order to make a comparison with the MEDLARS system possible, it was necessary, in addition, to choose a cutoff in the number of retrieved documents equivalent to that which MEDLARS would have obtained, had the SMART subcollection been used during the MEDLARS search. A typical cutoff computation is shown as an example in Table 2. Consider a typical request for which MEDLARS would have retrieved a total of 213 documents, including three RB items out of a total of six, thirty PB items, and 180 items retrieved but unassessed for relevance with respect to the given query. If it is assumed that the SMART subset contains all of the RB items, and twenty out of thirty PB items, then the retrieval cutoff is set at 23 documents for the recall calculations. (Since the SMART system ranks documents in decreasing correlation order with the search request, it is always possible to retrieve exactly the n highest ranking items). For the precision calculations, an additional adjustment is necessary since recall-base documents which are retrieved by MEDLARS are normally excluded from the MEDLARS precision calculations. For the example of Table 2, three such RB items had to be removed, the final cutoff being then 20 for precision purposes.

This procedure for determining the number of documents to be retrieved by SMART permits a direct comparison with the MEDLARS searches for the eighteen queries being processed. The following differences in the test environments must, however, be noted:

a) The original MEDLARS searches were conducted using the complete MEDLARS document collection, whereas the SMART searches were made with the subset for which keypunched abstracts were available; the possible effect of this reduction in collection size is discussed in the concluding section of this study;

b) The recall and precision bases used for the eighteen queries were larger for MEDLARS (518 items) than for SMART (273 items); the average MEDLARS recall-precision results of Table 4(c) show, however, that the MEDLARS performance for the two document subsets is comparable — indeed, MEDLARS obtains somewhat better results with the smaller set of 273 items — so that no further bias is introduced by the reduction in the size of recall and precision bases;

c) Since the MEDLARS <u>precision</u> calculations are based on the exact set of documents retrieved by MEDLARS in response to each search request, a comparison with the precision obtained by SMART can be made directly only if SMART retrieves exactly the same items as MEDLARS; in that case, the precision values are the same for the two systems; under all other circumstances, SMART retrieves items not also retrieved by MEDLARS, in which case, the corresponding documents are not normally assessed for relevance by MEDLARS, and a direct precision comparison becomes impossible; a precision adjustment must then be made before the respective values are comparable.

The precision adjustment actually made is based on the following argument: The apparent precision obtained by SMART takes into account only those documents which are retrieved by MEDLARS. But, MEDLARS does not retrieve <u>all</u> relevant items in its searches — in fact, the MEDLARS recall for the 18 test queries is only 0.64. Thus, the apparent SMART precision is based on the availability of only sixty-four percent of the relevant documents. The assumption is then made that the SMART precision would remain the same, were the full set of relevant documents to enter into the computation, instead of the sixty-four percent actually used; that is, it is conjectured that the proportion of relevant items retrieved would be the same for the unavailable relevant items (those not in the MEDLARS precision base) as for the available ones. Since

$$\frac{\text{Adjusted P}}{100\%} = \frac{\text{Apparent P}}{64\%} ,$$

the adjusted precision is obtained by multiplying the apparent precision by the factor 1.56. The complete argument is summarized in Table 3.

To summarize, the search results obtained by MEDLARS and SMART for the 18 queries are compared in the following manner: The cutoff value used by SMART to distinguish retrieved from nonretrieved items is exactly the one used in the corresponding MEDLARS search for the subset of 273 items; the recall calculations are based on the retrieval of the complete set of known relevant items, and the output values which result are directly comparable; the apparent precision calculations are based on an average MEDLARS recall of only sixty-four percent, and a suitable adjustment is made to account for the lack of relevance assessments in that part of the collection which is not retrieved by MEDLARS.

4. SMART-MEDLARS Comparison

The average recall and precision values obtained for the SMART and MEDLARS systems are shown in Table 4. The corresponding statistical significance calculations are given in Table 5. Tables 4(a) and 4(b) include, respectively, average recall and average precision values for the SMART runs, for each of three different language analysis systems: The word form dictionary, which makes it possible to match text words differing only in a final 's' (that is, singular and plural forms of the same word); the word stem dictionary, which includes a single entry for all words exhibiting the same word stem; and the thesaurus, which is used to recognize also words included within a single thesaurus class (such as synonyms and other related items).

For each dictionary system, three different SMART query sets are used for experimental purposes, termed "standard run", "negative delete", and "upweight", respectively. The results for the last two runs are examined in the next section. In each case, micro-averages are given as well as macro-averages. The former represent the averages obtained by comparing the total number of relevant retrieved over all 18 queries, to the total relevant or the total retrieved for all the queries; the latter are the actual per query averages and are normally more representative of the performance experienced by the average user [10].

A comparison of the average recall values for the 18 queries (Tables 4(a) and 4(c)) indicates that the micro-averages slightly favor MEDLARS, whereas the macro-averages slightly favor SMART. That is, MEDLARS is able to retrieve slightly more relevant documents overall, but SMART exhibits the better average recall per query. An examination of the recall values obtained for the individual queries listed in Table 6, reveals that MEDLARS may do very well, or very badly, in retrieving relevant documents, whereas SMART is more consistent in obtaining a performance which is generally neither perfect, nor very poor. Thus, perfect recall is obtained ten times by MEDLARS, but only seven times by SMART. In exchange, MEDLARS retrieves not a single relevant document in four instances, where this never happens for the SMART searches.

These figures point to a fundamental difference between manual indexing systems, and the automatic text processing schemes used in SMART: Often the human intermediary charged with the formulation of the search statement in the manual system is exceptionally clever in determining the user's information needs; at other times, however, these needs are misunderstood, thus accounting for the searches with zero recall. In addition, the manual indexing system is, of course, highly dependent on the richness and completeness of the indexing language, and on the thoroughness and accuracy with which the document indexing is performed.

In the automatic text analysis, on the other hand, the complete text of a document abstract is normally used for analysis purposes, and it is very rare indeed that the resulting content identifiers do not reflect the actual document content at least to some extent. In addition, the automatic environment makes it possible to use complex weighting and matching procedures designed to increase the effect of certain important content identifiers at the expense of others that are less crucial. At the same time, the basic dependence on the initial vocabulary is also responsible for the fact that some relevant items are difficult to retrieve, thus accounting for the less than perfect performance of the SMART searches.

The statistical significance output of Table 5 shows clearly that the recall differences between SMART and MEDLARS are not statistically significant; indeed, the sign test probabilities are equal to 1 for each dictionary. Thus, the average recall performance is just about identical for the two systems.

The precision figures for the standard SMART runs and the MEDLARS searches are contained in Tables 4(b) and 4(c) respectively. As expected, the apparent SMART precision is much smaller than the corresponding MEDLARS precision. However, when the adjustment factor is included, it is seen that the adjusted precision is only slightly lower for SMART than for MEDLARS, the differences in performance being again not statistically significant.

Overall, the average performance data of Table 4 lead to the conclusion that the MEDLARS and SMART performance is comparable for the 18 queries, with SMART showing a slightly better recall while MEDLARS exhibits a somewhat higher precision.

One factor, not taken into account in the average performance figures of Table 4, is the ability of the SMART system to rank the documents in decreasing correlation order with the search requests. Thus, in the comparison with MEDLARS, no distinction is made between different rankings of relevant documents that are retrieved. Such a ranking is, however, important to a user interested in retrieving the relevant items ahead of the nonrelevant ones.

To test the ranking ability of the automatic SMART process, a separate test was therefore made by comparing the "rank recall" measure [10] computed from the SMART ranks, with the rank recall obtained from a hand-ranked output list produced manually within the National Library of Medicine for test purposes. The hand-ranked output lists were available for fourteen of the eighteen queries, shown in the evaluation output of Table 7. It is seen here again that the performance of the two ranking systems — manual ranking with MEDLARS and automatic SMART ranking — is about equally effective, the average rank recall being slightly better for SMART than for MEDLARS.

A number of conclusions to be drawn from the foregoing test results are examined following the comparison of the various SMART runs.

5. Comparison of SMART Analysis Methods

Several different language analysis procedures were used for the SMART runs conducted with the MEDLARS subcollection. Specifically, runs were made using document titles only or full document abstracts, and three different dictionaries — known respectively as the word form, word stem, and thesaurus dictionaries — were used for language normalization. The first two dictionaries were generated by machine using for this purpose the standard SMART procedures; the thesaurus was generated by hand [11].

The differences in the dictionary makeup account, in general, for the differences in performance observed in the recall and precision measures of Table 4. Both the recall and precision values are nearly the same for word form and word stem dictionaries in the standard SMART runs. The thesaurus dictionary, on the other hand, which would normally be expected to produce better results than either of the suffix dictionaries produces only slightly better precision but slightly worse recall. The thesaurus groupings were actually constructed by a staff member without special knowledge of the medical terminology, and the corresponding performance is thus not typical of the thesaurus results obtained by SMART with document collections in different subject fields [5].

The results obtained by using document titles instead of full abstracts for analysis purposes are, however, fully in accord with comparable data previously obtained for different subject areas. The graph of Fig. 3(a) shows, in particular, that titles are much less effective than abstracts, particularly at the high recall end of the curve. Furthermore, the significance output of Fig. 3(c) indicates that the performance differences are, in fact, statistically significant, at least for the global measures.

Two additional minor modifications were made in the query set for test purposes. The first consisted in removing from the query statements all negative phrases included to denote what the user did _not_ wish to retrieve. Such negative phrases are not presently recognized by the standard SMART analysis methods, with the result that negative statements are actually interpreted as positive subject descriptions. The second query modification consisted in repeating in the query statements certain technical words occurring in the collection with low frequency. This modification produces an increased weight for the corresponding document identifiers, and may thereby generate a query statement which matches the user's interests more closely than the original. The several types of query formulations resulting from the modification procedure are shown for three queries in Table 8.

Both of the query alterations were performed manually, although machine programs might have been written to accomplish the same tasks. The new queries represent formulations that could realistically occur in an environment of informed users who would be instructed not to use negative subject descriptions, and who would emphasize the important technical terms in their query formulation.

The evaluation results for "negative deletion" and "upweighting" are included in Table 4 for the recall and precision averages corresponding to the MEDLARS searches, and in Fig. 4 in the form of recall-precision graphs. It is seen that the upweighting process improves both recall and precision by five to ten percent over the complete range of the recall-precision curve. The average recall and average adjusted precision values for the upweighted SMART queries are also superior to the comparable measures for the corresponding MEDLARS searches. The negative phrase deletion does not, however, exhibit the same uniformly beneficial effects, although some improvement in precision is noticeable at the low recall end of the curve. The significance data of Fig. 4(c) show that the changes in search effectiveness between original and altered queries are not sufficiently pronounced to be statistically significant.

An examination of the search results for the individual queries shows that the negative phrase deletion does not perform equally well for all queries. In particular, the procedure fails to improve the retrieval if the deletion process reduces the query to only a very short statement, no longer representative of user needs. It may also fail in cases where a given thesaurus grouping includes a variety of different concepts, where some of these concepts occur in a negative phrase, while others occur in a positive sense within the same query. In that case, the deletion of the negative phrases produces a decrease in the weight of important terms, which may consequently reduce the search effectiveness.

The upweighting process for important technical terms generally produces an improvement in search effectiveness. However, the improvement may be less uniform than expected. For some queries, it is easy to pick appropriate terms whose weight should be increased; for example, in query 01, listed in Table 8, the term "lens" may be expected to be much more essential for the subject description than, for example, the term "vertebrate". Other query statements may, however, occur for which the important terms are much more difficult to locate; in such cases, the search improvements due to upweighting may remain small, or may be nonexistent.

The two query modification procedures incorporated into the SMART system are only two possible methods which may improve the result of the automatic searches. Similar methods can, of course, also be used for the semi-manual MEDLARS searches. The prospects for such potential improvements in retrieval effectiveness are taken up in the concluding section.

6. Conclusions

The MEDLARS test comparisons which are described in this study lead to the same conclusions previously reached in other test environments with the SMART evaluation system [5]: Fully-automatic text analysis and search systems do not appear to produce a retrieval performance which is inferior to that obtained by conventional systems using manual document indexing and manual search formulations. While the manual indexing and search formulations can lead to exceptionally fine results when the indexer and/or the searcher are completely aware of the relationships between the stored collection and the user needs, the search results may also be very poor when these conditions are not met. The automatic process, on the other hand, with its exhaustive input data and complex analysis methods performs rarely very poorly, and may often produce completely satisfactory retrieval action.

Two important questions may be asked concerning the practical implications of the foregoing test results: First, is it reasonable to expect that identical results would hold if the automatic text processing methods were applied to the operational MEDLARS environment comprising half-a-million or more documents; and, second, can anything be done to improve the search effectiveness of presently existing automatic and manual information systems beyond those reflected in the recall-precision graphs of Figs. 1 to 4.

The first question cannot be answered with full certainity, since it is obviously not likely that keypunched abstracts should ever become available for the full MEDLARS collection. To what extent the present results can safely be extrapolated to searches performed with the full MEDLARS collection depends to a large extent on whether the set of properly rejected nonrelevant documents included in the MEDLARS collection falls into subject categories which are clearly far away from the query subjects. Obviously, if the nonrelevant documents not included in the SMART subset but included in the full collection could be assumed to be easier to reject than the nonrelevant actually included in the subset, then the SMART results for the full collection should be the same as those obtained for the subset alone. If, on the other hand, there are many more hard-to-reject nonrelevant items in the full collection, than in the subset, the results obtained by SMART on the subcollection may not be directly transferable to those obtainable on the full collection. An estimate for the amount of degradation to be expected in such a case may be obtained by adding to the SMART subset new documents which are nearly — but not quite — relevant to the search requests, and repeating the searches with the augmented collection. Based on the previous test results obtained with the SMART system in other subject areas, it is this writer's guess that the degradation, if any, will be small. This assertion remains, however, to be tested.

The problem relating to the fundamental improvements of both the SMART and MEDLARS searches is easier to treat. The originators of the internal MEDLARS test have, in fact, some pertinent suggestions to make concerning possible changes to be implemented in the search formulations, indexing language, and user-system interaction:

a) Concerning an appropriate query formulation "...the prime requirement is a complete statement of what the requester is looking for in the requester's own natural language, narrative form; [the query formulation must not] be deliberately phrased.... in a form that the requester believes will approximate a MEDLARS search strategy" [9, p. 117];

b) Concerning the indexing language to be used "we recommend a shift in emphasis away from the external advisory committee on terminology and towards the continued analysis of the terminological requirements of MEDLARS users as reflected in the demands placed upon the system [9, p. 193];

c) Concerning user-system interaction during the search "the greatest potential for improvement in MEDLARS exists at the interface between user and system; a significant improvement in the statement of requests can raise both the recall and the precision...." [9, p. 193].

That these suggestions are all well taken has been shown by the retrieval comparisons previously made with the SMART system [5]. Indeed, the search formulations suggested as ideal for MEDLARS are exactly the ones already used for all SMART searches. Furthermore, the dictionary construction principles derived for the SMART system also point in the direction of greater responsiveness to collection makeup and user needs, and away from committee control [12]. Finally, user-controlled iterative searches have been implemented successfully with the SMART system for several years [13,14,15].

It is difficult to predict exactly how much improvement in search effectiveness may result from the introduction of these various search and retrieval aids. The test results obtained under experimental conditions with the SMART system appear to indicate that the potential improvement will not exceed ten to fifteen percent, leading to a recall and precision performance of 0.70 or 0.75, instead of the present 0.50 to 0.60.

Such a performance would still be far short of what is desirable. However, it is encouraging to note that the present situations are well enough understood to make it reasonable to suggest avenues for the design of future improved systems, including viable automatic search and analysis procedures in place of some of the uncertain manual ones now in use.

References

[1] G. Salton and M. E. Lesk, The SMART Automatic Document Retrieval System — An Illustration, Communications of the ACM, Vol. 8, No. 6, June 1965.

[2] M. E. Lesk, Operating Instructions for the SMART Text Processing and Document Retrieval System, Scientific Report No. ISR-11 to the National Science Foundation, Section II, Department of Computer Science, Cornell University, June 1966.

[3] G. Salton, et al., Information Storage and Retrieval, Scientific Reports to the National Science Foundation, Nos. ISR-11, ISR-12, ISR-13, Department of Computer Science, Cornell University, June 1966, June 1967, and January, 1968.

[4] C. W. Cleverdon and E. M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 2 — Test Results, Aslib-Cranfield Research Project, Cranfield, 1966.

[5] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1968.

[6] A. M. Rees, Evaluation of Information Systems and Services, in Annual Review of Information Science and Technology, Vol. 2, C. Cuadra, editor, Interscience Publishers, New York, 1967.

[7] A. M. Rees and D. G. Schultz, A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching, Final Report to the National Science Foundation, Center for Documentation and Communication Research, Case Western Reserve University, October 1967.

[8] F. W. Lancaster, Evaluating the Performance of a Large Operating Information Retrieval System, Proceedings of the Second Electronic Information Handling Conference, Thompson Book Company, Washington, 1967.

[9] F. W. Lancaster, Evaluation of the Operating Efficience of MEDLARS, Final Report, National Library of Medicine, January, 1968.

References
(contd)

[10] G. Salton, The Evaluation of Computer-based Information Retrieval Systems, Proceedings of the FID 1965 Congress, Spartan Books, Washington 1966.

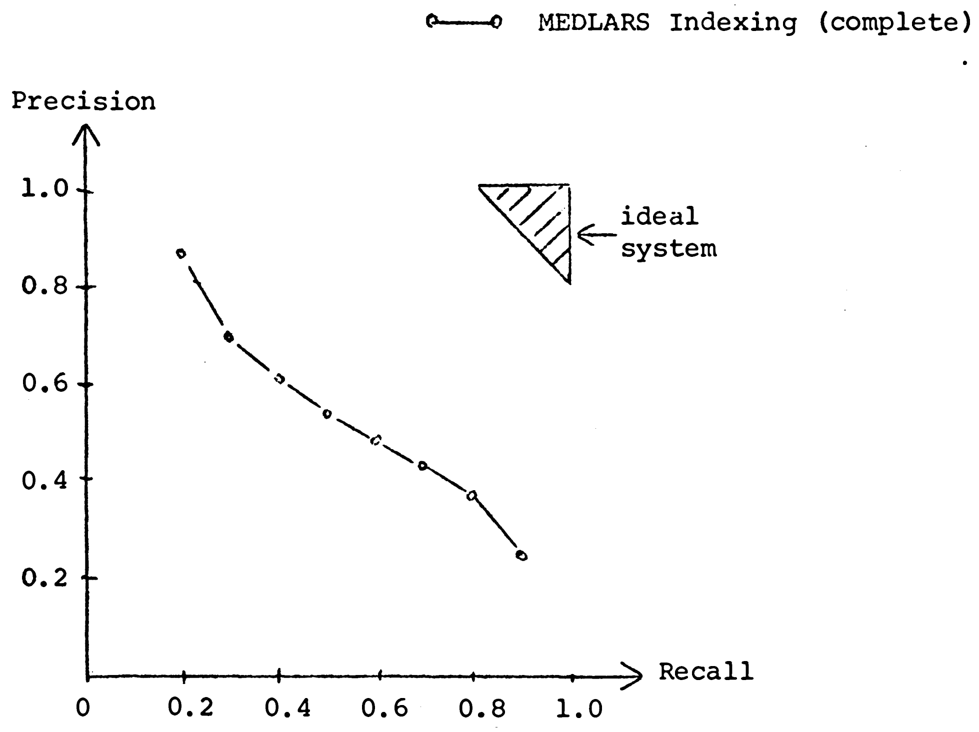
[11] E. M. Keen, Suffix Dictionaries and Thesaurus, Phrase, and Hierarchy Dictionaries, Information Storage and Retrieval, Report No. ISR-13 to the National Science Foundation, Sections VI and VII, Department of Computer Science, Cornell University, January 1968.

[12] G. Salton, Information Dissemination and Automatic Information Systems, Proceedings of the IEEE, Vol. 54, No. 12, December 1966.

[13] J. J. Rocchio, Jr., Document Retrieval Systems — Optimization and Evaluation, Harvard University Doctoral Thesis, Scientific Report No. ISR-10 to the National Science Foundation, Harvard Computation Laboratory, March 1966.

[14] J. J. Rocchio and G. Salton, Information Search Optimization and Iterative Retrieval Techniques, Proceedings of the AFIPS Fall Joint Computer Conference, Vol. 27, Spartan Books, November 1965.

[15] G. Salton, Search and Retrieval Experiments in Real-Time Information Retrieval, Proceedings of the IFIP Congress '68, Edinburgh, August 1968, (also Technical Report 68-8, Department of Computer Science, Cornell University, Ithaca, N. Y., February 1968.)



MEDLARS Recall-Precision Range for Complete Test
(302 queries, 500,000 documents)

Fig. 2

Source of Failure	Percentage of 797 Recall Failures	Percentage of 3038 Precision Failures
Index Language (lack of specific term or false coordination of terms)	10%	36%
Searching (search formulation too exhaustive or too specific)	35%	32%
Indexing (document indexing insufficiently exhaustive, or too exhaustive, or omission of important term)	37%	13%
Lack of User-System Interaction	25%	17%
Miscellaneous	1%	2%

Typical Recall and Precision Failures for Complete MEDLARS Test
(302 queries, 500,000 documents)

Table 1

A) <u>Recall Computation</u>	
Retrieved by MEDLARS	3 RB (out of 6) 30 PB 180 unassessed
SMART Subset	all RB 20 PB
SMART Cutoff	20 PB + 3 RB = 23
Assuming 4 RB retrieved (out of 23 items)	Recall 4/6 = 0.66
B) <u>Precision Computation</u>	
Initial cutoff	23
Intermediate cutoff (after removal of RB)	23 - 3 = 20
Number retrieved by SMART in top 20	4 RB, 9 PB relevant
Final cutoff	20 (remove 4 RB and replace by 4 new items)
SMART retrieval 9 relevant out of 16 1 relevant out of 4 new items	Precision $\frac{9+1}{20} = 0.50$

Sample SMART-MEDLARS Cutoff Computation

Table 2

Precision Adjustment

1. SMART can reach MEDLARS precision value only if it retrieves exactly the same items as MEDLARS (since calculations are based on PB)

2. SMART PB base consists only of items retrieved by MEDLARS, and MEDLARS does not retrieve all relevant (MEDLARS recall on SMART subset = 0.64)

3. SMART and MEDLARS are independent systems, and assuming all relevant were available in SMART collection, some of them would be retrieved by SMART

4. Assuming that the percentage of relevant retrieved were to remain the same if all relevant were available to SMART

$$\frac{\text{Apparent Precision}}{\text{Percent relevant in SMART collection (64\%)}} = \frac{\text{Adjusted Precision}}{\text{all relevant (100\%)}}$$

5. Adjustment

$$\text{Adjusted P} = \text{Apparent P} \cdot \frac{100}{64}$$

Explanation for Precision Adjustment

Table 3

	SMART (standard)			SMART (negative delete)			SMART (upweight)		
	Word Form	Thes.	Word Stem	Word Form	Thes.	Word Stem	Word Form	Thes.	Word Stem
"Micro" Average	0.644	0.632	0.655	0.667	0.644	0.667	0.770	0.690	0.770
"Macro" Average	0.704	0.695	0.718	0.665	0.700	0.718	0.802	0.692	0.799

a) SMART Recall Averages
(18 queries, 273 documents)

	SMART (standard)			SMART (negative delete)			SMART (upweight)		
	Word Form	Thes.	Word Stem	Word Form	Thes.	Word Stem	Word Form	Thes.	Word Stem
Apparent "Micro"	0.395	0.410	0.389	0.355	0.395	0.385	0.445	0.440	0.430
Apparent "Macro"	0.368	0.393	0.367	0.353	0.394	0.342	0.431	0.430	0.421
Adjusted "Micro"	0.583	0.605	0.574	0.524	0.583	0.568	0.656	0.649	0.634
Adjusted "Macro"	0.571	0.611	0.570	0.549	0.613	0.531	0.670	0.669	0.655

b) SMART Precision Averages
(18 queries, 273 documents)

	Recall	Precision
Micro - average Macro - average (273 documents)	0.678 0.643	0.640 0.625
Micro - average Macro - average (518 documents)	0.671 0.558	0.568 0.573

c) MEDLARS Recall - Precision Values
(18 queries)

Recall-Precision Tables

Table 4

Analysis Methods being Compared	Probabilities (A over B) t-Test	Sign Test	Comparisons favoring A, B, even
A. MEDLARS Search B. SMART Word Stem			
Recall	0.5676	1.0000	A: 11
Apparent Precision	0.0001	0.0002	B: 13
Adjusted Precision	0.2706	0.6291	even: 12
A. MEDLARS Search B. SMART Thesaurus			
Recall	0.6883	1.0000	A: 9
Apparent Precision	0.0004	0.0005	B: 15
Adjusted Precision	0.8139	1.0000	even: 12
A. MEDLARS Search B. SMART Word Form			
Recall	0.5675	1.0000	A: 12
Apparent Precision	0.0001	0.0002	B: 13
Adjusted Precision	0.2911	0.6291	even: 11

Significance Computations for SMART-MEDLARS Comparisons
(18 queries, 273 documents, standard runs)

Table 5

Recall Range	Number of Queries in Range MEDLARS	SMART (stem)
1.0	10	7
0.99-0.50	1	7
0.49-0.01	3	4
0.0	4	0

Query Distribution in Various Recall Ranges

Table 6

Query Number	No. of PB Relevant	Rank Recall MEDLARS	SMART	No. of queries favoring MEDLARS	SMART	Neither
02	5	0.88	0.94		✓	
04	13	0.91	0.83	✓		
05	10	0.92	0.98		✓	
06	8	0.40	0.76		✓	
07	3	0.43	1.0		✓	
09	11	0.97	0.93	✓		
10	11	0.90	0.83	✓		
13	10	0.97	0.63	✓		
14	2	1.0	1.0			✓
16	4	0.71	0.71			✓
18	2	0.60	1.0		✓	
32	15	0.75	0.75			✓
40	3	1.0	1.0			✓
187	9	0.70	0.69	✓		
Average Rank Recall		0.80	0.86	5	5	4

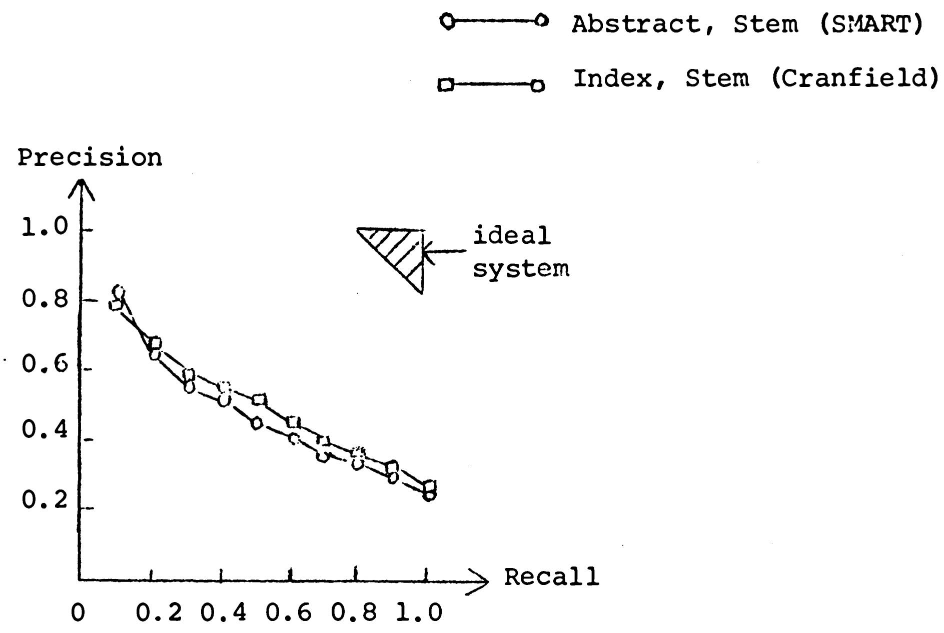
Rank Recall Comparison for "Handranked" MEDLARS
with SMART Word Stem Process
(14 queries using PB documents only)

Table 7

Query Number	Original Query	Negative Phrase Delete	With Term Upweighting
01	The crystalline lens in vertebrates, including humans, but not drug therapy or surgery.	The crystalline lens in vertebrates, including humans.	[Original Query], crystalline lens, crystalline lens.
03	Electron microscopy of lung or bronchi. Pleura or pleural diseases may be excluded.	Electron microscopy of lung or bronchi.	[Original Query], electron, electron, lung, bronchi.
13	Blood or urinary steroids in human breast or prostatic neoplasms. Drug therapy, toxicology, etc., to be excluded.	Blood or urinary steroids in human breast or prostatic neoplasm.	[Original Query], steroids, steroids, breast, prostate.

Samples of Query Modification by Negative
Phrase Deletion and Upweighting

Table 8



a) Recall-Precision Graph

SMART Word Stem		Cranfield Indexing	
R	P	R	P
0.1	0.8239	0.1	0.8045
0.2	0.6518	0.2	0.6581
0.3	0.5578	0.3	0.5908
0.4	0.5093	0.4	0.5498
0.5	0.4522	0.5	0.5171
0.6	0.4143	0.6	0.4506
0.7	0.3800	0.7	0.4035
0.8	0.3431	0.8	0.3649
0.9	0.3005	0.9	0.3233
1.0	0.2551	1.0	0.2799
RNK REC= 0.2998		RNK REC= 0.3122	
LOG PRE= 0.4655		LOG PRE= 0.4674	
NOR REC= 0.8644		NOR REC= 0.8897	
NOR PRE= 0.6704		NOR PRE= 0.6831	

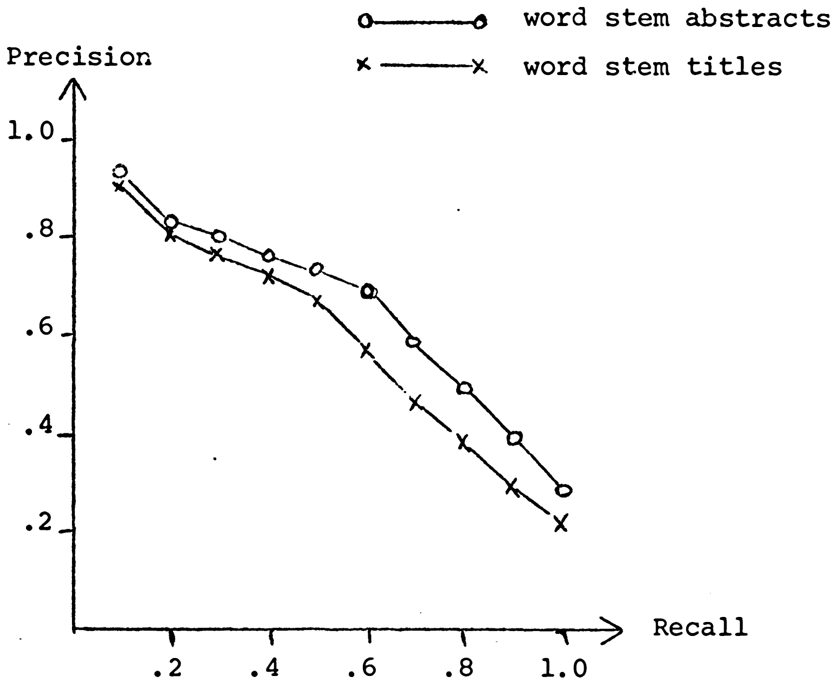
b) Recall-Precision Tables

Evaluation Measures	Probabilities t-Test (Indexing	Sign Test over Abstract)
Precision at		
R = 0.1	0.5151	0.7011
R = 0.3	0.4358	0.7283
R = 0.5	0.1163	1.0000
R = 0.7	0.3682	0.8679
R = 0.9	0.4044	0.2559
Rank Recall	0.6622	0.0470
Log Precision	0.9341	1.0000
Normed Recall	0.1491	0.1081
Normed Precision	0.7268	1.0000
Combined	0.0415	0.0465

c) Significance Output

Recall-Precision Comparisons for Cranfield Indexing and SMART Word Stem Process
(averages 42 queries, 200 document abstracts, cosine numeric)

Fig. 1



a) Recall-Precision Graph

Word Stem Abstracts R	Word Stem Abstracts P	Word Stem Titles R	Word Stem Titles P
0.1	0.9167	0.1	0.8973
0.2	0.8132	0.2	0.8154
0.3	0.8008	0.3	0.7610
0.4	0.7646	0.4	0.7178
0.5	0.7311	0.5	0.6718
0.6	0.6947	0.6	0.5694
0.7	0.5893	0.7	0.4683
0.8	0.4904	0.8	0.3904
0.9	0.3962	0.9	0.2909
1.0	0.2814	1.0	0.2319
RNK REC= 0.4092		RNK REC= 0.3023	
LOG PRE= 0.6926		LOG PRE= 0.6446	
NOR REC= 0.9104		NOR REC= 0.7508	
NOR PRE= 0.8199		NOR PRE= 0.7169	

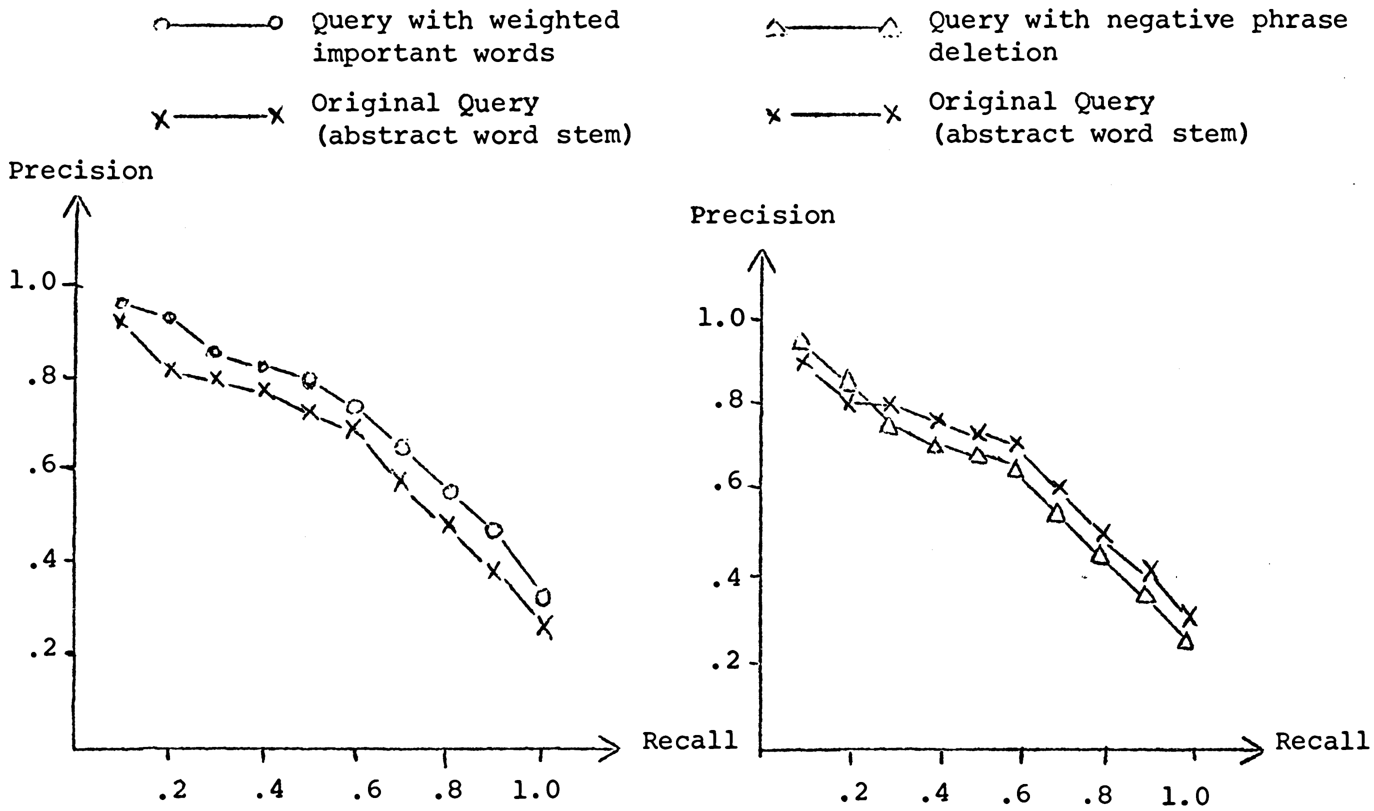
b) Recall-Precision Tables

Evaluation Measures	Probabilities (Abstracts over Titles) t-Test	Sign Test
Precision at		
R = 0.1	0.5066	1.0000
R = 0.3	0.3076	0.2266
R = 0.5	0.2311	1.0000
R = 0.7	0.0438	0.0213
R = 0.9	0.1305	0.1435
Rank Recall	0.1446	0.0042
Log Precision	0.4790	0.0127
Normed Recall	0.0005	0.0074
Normed Precision	0.0085	0.0127
Combined	0.0000	0.0000

c) Significance Output

Recall-Precision Data for Abstract-Title Comparisons
(MEDLARS collection, 18 queries, 273 documents)

Fig. 3



a) Recall-Precision Graphs for Comparison of Original Queries with Altered Queries

Measures	Original	Upweight	Negative
Precision at			
R = 0.1	0.9167	0.9627	0.9288
R = 0.3	0.8008	0.8500	0.7939
R = 0.5	0.7311	0.8004	0.7246
R = 0.7	0.5893	0.6441	0.5794
R = 0.9	0.3962	0.4776	0.3898
RNK REC	0.4092	0.4546	0.3973
LOG PRE	0.6926	0.7260	0.6874
NOR REC	0.9104	0.9200	0.8294
NOR PRE	0.8199	0.8476	0.7674

b) Recall-Precision Tables

Evaluation Measures	Upweight over Original (Sign Test)	Original over Negative (Sign Test)
Precision at		
R = 0.1	0.6875	1.0000
R = 0.3	1.0000	1.0000
R = 0.5	0.7539	1.0000
R = 0.7	1.0000	0.7266
R = 0.9	0.0574	0.7539
RNK REC	0.4545	0.7539
LOG PRE	0.1796	0.5078
NOR REC	0.4240	0.7539
NOR PRE	0.3018	0.7539
Combined	0.0011	0.1215

c) Significance Output

Recall-Precision Comparisons for Original Queries and Altered Queries by Negative Phrase Deletion and Upweighting
(MEDLARS collection, 18 queries, 273 documents, abstract stem process)

Fig. 4

Appendix A

Recall-Precision Comparisons for Individual Queries

Recall for MEDLARS

Query	Cutoff	Total Relevant	MEDLARS Retrieved	MEDLARS Recall	SMART Retrieved			SMART Recall		
					Word Form	Thes	Word Stem	Word Form	Thes	Word Stem
1	27	11	11	1.0	7	6	7	0.64	0.55	0.64
2	8	3	3	1.0	3	3	3	1.0	1.0	1.0
3	18	9	9	1.0	4	5	4	0.44	0.56	0.44
4	17	3	1	0.33	1	1	1	0.33	0.33	0.33
5	11	4	0	0	1	2	2	0.25	0.5	0.5
6	21	5	5	1.0	5	5	5	1.0	1.0	1.0
7	12	2	2	1.0	2	2	2	1.0	1.0	1.0
8	6	11	0	0	6	6	6	0.55	0.55	0.55
9	23	11	11	1.0	5	4	5	0.46	0.36	0.46
10	14	1	1	1.0	1	1	1	1.0	1.0	1.0
13	13	3	3	1.0	2	1	2	0.67	0.33	0.67
14	4	3	1	0.33	1	1	1	0.33	0.33	0.33
16	14	7	7	1.0	7	7	7	1.0	1.0	1.0
18	3	4	0	0	3	3	3	0.75	0.75	0.75
32	21	3	2	0.66	3	3	3	1.0	1.0	1.0
40	8	4	1	0.25	3	3	3	0.75	0.75	0.75
187	14	2	2	1.0	1	1	1	0.5	0.5	0.5
303	10	1	0	0	1	1	1	1.0	1.0	1.0

	MEDLARS	SMART		
		Word Form	Thes	Word Stem
"Micro" average	0.678	0.644	0.632	0.655
"Macro" average	0.643	0.704	0.695	0.718

Basic SMART-MEDLARS Recall Comparisons

Table A-1

Query	Total Retrieved	MEDLARS Relevant	MEDLARS Precision	SMART Relevant Word Form	SMART Relevant Thes	SMART Relevant Word Stem	SMART Precision Word Form	SMART Precision Thes	SMART Precision Word Stem
1	17	15	0.882	11	10	10	0.647	0.588	0.588
2	6	5	0.883	2	3	2	0.333	0.500	0.333
3	9	7	0.778	3	5	4	0.333	0.556	0.444
4	17	13	0.765	5	7	5	0.294	0.412	0.294
5	11	10	0.909	8	7	8	0.727	0.636	0.727
6	20	8	0.400	8	8	8	0.400	0.400	0.400
7	11	3	0.273	3	3	3	0.273	0.273	0.273
8	6	1	0.167	0	0	0	0	0	0
9	12	11	0.917	3	1	3	0.250	0.083	0.250
10	14	11	0.786	6	7	5	0.429	0.500	0.357
13	12	8	0.667	3	1	3	0.250	0.083	0.250
14	3	2	0.667	2	2	2	0.667	0.667	0.667
16	7	5	0.714	4	3	4	0.571	0.429	0.571
18	3	2	0.667	1	2	1	0.333	0.667	0.333
32	21	15	0.714	9	11	9	0.429	0.524	0.429
40	8	3	0.375	3	3	3	0.375	0.375	0.375
187	13	9	0.692	4	5	4	0.308	0.385	0.308
303	10	0	0	0	0	0	0	0.	0

	MEDLARS	SMART Apparent Word Form	SMART Apparent Thes	SMART Apparent Word Stem	SMART Adjusted Word Form	SMART Adjusted Thes	SMART Adjusted Word Stem
"Micro" average	0.640	0.395	0.410	0.389	0.583	0.605	0.574
"Macro" average	0.625	0.368	0.393	0.367	0.571	0.611	0.570

Basic SMART-MEDLARS Precision Comparison

Table A-2

Query	Cutoff	Total Relevant	MEDLARS Retrieved	MEDLARS Recall	SMART Retrieved Word Form	SMART Retrieved Thes	SMART Retrieved Word Stem	SMART Recall Word Form	SMART Recall Thes	SMART Recall Word Stem
1	27	11	11	1.0	9	7	8	0.82	0.64	0.73
2	8	3	3	1.0	3	3	3	1.0	1.0	1.0
3	18	9	9	1.0	5	5	5	0.56	0.56	0.56
4	17	3	1	0.33	1	1	1	0.33	0.33	0.33
5	11	4	0	0	1	2	2	0.25	0.50	0.50
6	21	5	5	1.0	5	5	5	1.0	1.0	1.0
7	12	2	2	1.0	2	2	2	1.0	1.0	1.0
8	6	11	0	0	6	6	6	0.55	0.55	0.55
9	23	11	11	1.0	5	4	5	0.46	0.36	0.46
10	14	1	1	1.0	0	1	1	0	1.0	1.0
13	13	3	3	1.0	2	1	2	0.67	0.33	0.67
14	4	3	1	0.33	1	1	1	0.33	0.33	0.33
16	14	7	7	1.0	7	7	7	1.0	1.0	1.0
18	3	4	0	0	3	3	3	0.75	0.75	0.75
32	21	3	2	0.67	3	3	3	1.0	1.0	1.0
40	8	4	1	0.25	3	3	3	0.75	0.75	0.75
187	14	2	2	1.0	1	1	1	0.5	0.5	0.5
303	10	1	0	0	1	1	1	1.0	1.0	1.0

	MEDLARS	SMART Word Form	SMART Thes	SMART Word Stem
"Micro" average	0.678	0.667	0.644	0.678
"Macro" average	0.643	0.665	0.700	0.729

Recall with Negative Phrase Deletion

Table A-3

Query	Total Retrieved	MEDLARS Relevant	MEDLARS Precision	SMART Relevant			SMART Precision		
				Word Form	Thes	Word Stem	Word Form	Thes	Word Stem
1	17	15	0.882	12	12	12	0.706	0.706	0.706
2	6	5	0.883	2	3	2	0.333	0.5	0.333
3	9	7	0.778	4	4	5	0.444	0.444	0.556
4	17	13	0.765	5	7	5	0.294	0.412	0.294
5	11	10	0.909	8	7	8	0.727	0.636	0.727
6	20	8	0.400	8	8	8	0.4	0.4	0.4
7	11	3	0.273	3	3	3	0.273	0.273	0.273
8	6	1	0.167	0	0	0	0	0	0
9	12	11	0.917	3	1	3	0.25	0.083	0.25
10	14	11	0.786	0	6	4	0	0.429	0.286
13	12	8	0.667	3	2	5	0.25	0.167	0.417
14	3	2	0.667	2	2	2	0.667	0.667	0.667
16	7	5	0.714	4	3	4	0.571	0.429	0.571
18	3	2	0.667	1	2	1	0.333	0.667	0.333
32	21	15	0.714	9	11	9	0.429	0.524	0.429
40	8	3	0.375	3	3	3	0.375	0.375	0.375
187	13	9	0.692	4	5	4	0.308	0.385	0.308
303	10	0	0	0	0	0	0	0	0

	MEDLARS	SMART Apparent			SMART Adjusted		
		Word Form	Thes	Word Stem	Word Form	Thes	Word Stem
"Micro" average	0.640	0.355	0.395	0.390	0.524	0.583	0.575
"Macro" average	0.625	0.353	0.394	0.385	0.549	0.613	0.599

Precision with Negative Phrase Deletion

Table A-4

Query	Cutoff	Total Relevant	MEDLARS Retrieved	MEDLARS Recall	SMART Retrieved Word Form	SMART Retrieved Thes	SMART Retrieved Word Stem	SMART Recall Word Form	SMART Recall Thes	SMART Recall Word Stem
1	27	11	11	1.0	11	7	11	1.0	0.64	1.0
2	8	3	3	1.0	3	3	3	1.0	1.0	1.0
3	18	9	9	1.0	5	4	5	0.56	0.44	0.56
4	17	3	1	0.33	1	1	1	0.33	0.33	0.33
5	11	4	0	0	1	2	2	0.25	0.50	0.50
6	21	5	5	1.0	5	5	4	1.0	1.0	0.80
7	12	2	2	1.0	2	2	2	1.0	1.0	1.0
8	6	11	0	0	6	6	6	0.55	0.55	0.55
9	23	11	11	1.0	10	9	9	0.91	0.82	0.82
10	14	1	1	1.0	1	1	1	1.0	1.0	1.0
13	13	3	3	1.0	3	2	3	1.0	0.67	1.0
14	4	3	1	0.33	1	0	1	0.33	0	0.33
16	14	7	7	1.0	7	7	7	1.0	1.0	1.0
18	3	4	0	0	3	3	3	0.75	0.75	0.75
32	21	3	2	0.66	3	3	3	1.0	1.0	1.0
40	8	4	1	0.25	3	3	3	0.75	0.75	0.75
187	14	2	2	1.0	2	2	2	1.0	1.0	1.0
303	10	1	0	0	1	0	1	1.0	0	1.0

	MEDLARS	SMART Original Word Form	SMART Original Thes	SMART Original Word Stem	SMART Negative Word Form	SMART Negative Thes	SMART Negative Word Stem	SMART Upweight Word Form	SMART Upweight Thes	SMART Upweight Word Stem
"Micro"	0.678	0.644	0.621	0.655	0.667	0.644	0.667	0.770	0.690	0.770
"Macro"	0.643	0.704	0.690	0.718	0.665	0.700	0.718	0.802	0.692	0.799

Recall for Upweighted Queries

Table A-5

Query	Total Retrieved	MEDLARS Relevant	MEDLARS Precision	SMART Relevant Word Form	SMART Relevant Thes	SMART Relevant Word Stem	SMART Precision Word Form	SMART Precision Thes	SMART Precision Word Stem
1	17	15	0.882	15	12	15	0.882	0.706	0.882
2	6	5	0.883	2	3	2	0.333	0.500	0.333
3	9	7	0.778	4	5	5	0.444	0.556	0.556
4	17	13	0.765	7	5	6	0.412	0.294	0.353
5	11	10	0.909	8	8	8	0.800	0.800	0.800
6	20	8	0.400	8	8	7	0.400	0.400	0.350
7	11	3	0.273	3	2	3	0.273	0.182	0.273
8	6	1	0.167	0	0	0	0	0	0
9	12	11	0.917	8	6	7	0.667	0.500	0.583
10	14	11	0.786	6	7	5	0.429	0.500	0.357
13	12	8	0.667	4	3	4	0.333	0.25	0.333
14	3	2	0.667	2	1	2	0.667	0.33	0.667
16	7	8	0.714	4	4	4	0.571	0.511	0.571
18	3	2	0.667	1	2	1	0.333	0.667	0.333
32	21	15	0.714	8	12	9	0.381	0.571	0.429
40	8	3	0.375	3	3	3	0.375	0.375	0.375
187	13	9	0.692	6	7	5	0.462	0.538	0.385
303	10	0	0	0	0	0	0	0	0

	MEDLARS	SMART Original Word Form	SMART Original Thes	SMART Original Word Stem	SMART Negative Word Form	SMART Negative Thes	SMART Negative Word Stem	SMART Upweight Word Form	SMART Upweight Thes	SMART Upweight Word Stem
"Micro"	0.640	0.395	0.410	0.389	0.355	0.395	0.385	0.445	0.440	0.430
"Macro"	0.625	0.368	0.393	0.367	0.353	0.394	0.342	0.431	0.430	0.421

Precision for Upweighted Queries

Table A-6

Appendix B

Text of MEDLARS Queries*

01 1 The crystalline lens in vertebrates, including humans, but not drug therapy or surgery. Crystalline lens, crystalline lens.

02 2 The relationship of blood and cerebrospinal fluid oxygen concentrations or partial pressures. A method of interest is polarography. Blood oxygen, cerebrospinal oxygen.

03 3 Electron microscopy of lung or bronchi. Pleura or pleural diseases may be excluded. Electron, electron, lung, bronchi.

04 4 Tissue culture of lung or bronchial neoplasms. Lung, bronchial.

05 5 The crossing of fatty acids through the placental barrier. Normal fatty acid levels in placenta and fetus. Fatty acids, placenta.

06 6 Ventricular septal defect occurring in association with aortic regurgitation; aortic stenosis probably not involved. Septal defect, aortic regurgitation.

07 7 Radioisotopes in heart scanning. Mainly used in diagnosis of pericardial effusions. Also used to study tumors, heart enlargement, aneurysms and pericardial thickening. Technetium, RIHSA, radioactive hippurate, cholegraffin are used. Radio-isotopes, radioisotopes.

08 8 The effects of drugs on the bone marrow of man and animals, specifically the effect of pesticides. Also, the significance of bone marrow changes. Pesticides, bone marrow.

09 9 The use of induced hypothermia in heart surgery, neurosurgery, head injuries and infectious diseases. Hypothermia, hypothermia.

10 10 Neoplasm immunology, excluding plant tumors, granulomas, on-cogenic viruses. Neoplasm immunology.

Appendix B
(contd)

13	11	Blood or urinary steroids in human breast or prostatic neoplasms. Drug therapy, toxicology, etc. to be excluded. Steroids, steroids, breast, prostate.
14	12	Effect of azath ioprine on systemic lupus erythematosus, particularly in regard to renal lesions. Azath, ioprine, azath, ioprine.
16	13	Bacillus subtilis phages and genetics, with particular reference to transduction. Subtilis, subtilis.
18	14	Renal amyloidosis as a complication of tubercolosis and the effects of steroids on this condition. Only the terms kidney diseases and nephrotic syndrome were selected by the requester. Prednisone and prednisolone are the only steriods of interest. Prednisone, prednisolone, renal amyloidosis.
32	15	Homonymous menianopsia in visual aphasia, particularly measurement and assessment. Gerstmann's syndrome and agnosia are also of interest. Hemianopsia, hemianopsia.
40	16	Separation anxiety in infancy (i.e. up to two years of age) and in preschool children, particularly separation of a child from its mother. Separation anxiety, separation anxiety.
187	17	Nickel in nutrition: requirements for methods for analysis; relation with enzyme systems; toxicity of, in humans and laboratory animals; deficiency signs and symptoms; level in various foodstuffs; level in blood and tissues. Nickel, nickel, nickel.
303	18	The toxicity of organic selenium compounds. Selenium, selenium, toxicity.

---

*Underlined words were added to upweight important concepts and used in upweighting experiment.