Recent Studies in Automatic Text Analysis and Document Retrieval

G. SALTON

Cornell University, Ithaca, New York

ABSTRACT. Many experts in mechanized text processing now agree that useful automatic language analysis procedures are largely unavailable and that the existing linguistic methodologies generally produce disappointing results. An attempt is made in the present study to identify those automatic procedures which appear most effective as a replacement for the missing language analysis.

A series of computer experiments is described, designed to simulate a conventional document retrieval environment. It is found that a simple duplication, by automatic means, of the standard, manual document indexing and retrieval operations will not produce acceptable output results. New mechanized approaches to document handling are proposed, including document ranking methods, automatic dictionary and word list generation, and user feedback searches. It is shown that the fully automatic methodology is superior in effectiveness to the conventional procedures in normal use.

KEY WORDS AND PHRASES: automatic text processing, automatic indexing, information retrieval evaluation, computational linguistics

CR CATEGORIES: 3.42, 3.71, 3.72, 3.75

1. *Linguistics and Information Processing*

Experts in computational linguistics and automatic text handling have been convinced for many years that the outstanding problems in automatic language processing are not likely to be solved in the absence of powerful linguistic analysis techniques. The feeling is thus widespread that the first order of business must be the study of the structural and semantic properties of natural languages in the hope of gaining a sufficient understanding to lead to a solution of the open text processing problems [1, 2]. Unhappily, as Pacak and Pratt point out, the chances of gaining the necessary command of natural language structures in the foreseeable future are exceedingly dim because [3]:

"very significant weaknesses are still apparent [in automatic linguistic analysis]; all existing systems and/or models are small and core-bound . . . ; no system is able to deal with more than a small subset of English . . .; very few systems have tried to go beyond sentence boundaries in their analysis . . .; and all the mentioned theoretical proposals are highly tentative."

This study was supported in part by the National Science Foundation under grant GJ-314, and in part by the National Library of Medicine under grant LM-00704.

Author's address: Cornell University, Department of Computer Science, Ithaca, NY 14850.

The apparent situation is then a bleak one: on the one hand, a better understanding of syntax and semantics is said to be necessary to solve the open problems in natural language processing; on the other hand, the likelihood of making substantial and timely progress in computational linguistics is remote.

Fortunately, when the various text processing problems are actually considered in detail, it becomes clear that predictions of an impending dead end in language processing are premature. First of all, the text processing applications differ in scope and complexity. Second, it has not yet been shown that no replacement is possible for the missing linguistic analysis techniques.

An attempt may then be worthwhile to approach some of the text processing problems in the information retrieval area using either extra linguistic procedures that are not primarily based on sentence structure and semantics, or else using simple language analysis methods from which the more troublesome complexities are eliminated. These possibilities are investigated in the remainder of this study.

2. *Simplified Language Analysis*

The first approach to the language analysis problem consists of setting aside many of the difficulties belonging to the general area of "semantics"—for example, the disambiguation of polysemantic words; the recognition of synonyms; the treatment of punctuation marks; the interpretation of anaphoric, elliptical, and idiomatic constructions; the problem of indirect references; the discourse analysis of paragraphs and larger units of text; and so on—and concentrating instead on the structural properties of the language.

Unfortunately, the existing programs for syntactic text analysis are not easy to use because a unique correct analysis is not generally obtainable for each input sentence. However, the existing "phrase structure" analyzers can normally be relied upon to identify most noun phrases, prepositional phrases, and subject-verb-object structures with fairly good accuracy. For this reason, attempts have been made to incorporate simple phrase identification routines into a number of programs for automatic indexing, or abstracting of texts [4–7]. Many variations are possible; the following steps are often included [4]:

(1) Individual text words are first isolated.

(2) Syntactic units such as quantifiers, prepositions, clause introductions, are identified.

(3) Phrases are formed consisting of word strings beginning with one of the specified syntactic units—for example, a preposition or a relative pronoun—and ending before the beginning of the next specified syntactic unit.

(4) Certain backward connections between elements of different phrases are recognized consisting generally of the antecedents of certain pronouns and the governors of dependent prepositional phrases.

(5) Phrases which exhibit special structural properties, including mostly nouns and adjectives plus certain prepositions, are then extracted to serve as content identifiers.

Several evaluation studies of automatic indexing systems have been performed which include procedures similar to the ones outlined above. In general, one finds that a large proportion of the expressions assigned automatically would also have been chosen manually by subject experts charged with the indexing task [8]. On the other hand, when the phrase generation procedures using simplified syntax are

compared with other, simpler, content analysis methods which include no structural or semantic components, the surprising conclusion is that on the average better results are obtainable without the syntactic components than with them. For example, Cleverdon finds that [9]:

"the [Aslib-Cranfield] evaluation project indicates that the single term index languages are superior to any other type . . .; the phrase (concept) index languages were over-specific when used in natural language."

Several possible explanations suggest themselves to account for the failure of the simple syntactic devices in information retrieval environments. Most obvious may be the inadequacy of the syntactic analyzers used to generate the phrase identifiers. Many correct phrases may be expected to remain unidentified, while some phrases actually assigned may not be as helpful for content analysis as expected. Second, in a system designed to serve a heterogeneous user population, it may be unwise to overspecify the document content; in particular, since the interests of the user population are not always easily assessed in advance, a less specific content description in terms of individual concepts may better serve the average user than a narrow analysis in terms of phrases. Finally, most retrieval evaluation systems use performance criteria averaged over many user queries. In such a situation, a better than average performance for most queries may be preferable to a situation where some users score very highly while others do quite poorly.

Be that as it may, the conclusion which must apparently be drawn from the available evidence is that the language analysis procedures which offer themselves for use in automatic text processing systems do not produce results commensurate with the effort needed to implement them. A complete language analysis is not possible because the necessary linguistic know-how is not at hand, while the simple syntactic methodologies do not operate sufficiently effectively.

It remains to determine whether *nonlinguistic* computer techniques might be useful in replacing both the lack of linguistic know-how as well as the intellectual input provided by human beings. This question is investigated in the next section of this study.

3. *Manual Versus Automatic Indexing*

Over the years, a great deal of evidence has accumulated concerning the relative merits of manual and automatic indexing techniques [8–18]. When automatically generated keywords are compared with terms manually assigned by subject experts, one normally finds agreement for 60 to 80 percent of the assigned terms. Furthermore, the retrieval results obtainable with automatic indexing techniques are not, on the average, substantially different from those produced by conventional manual methods. As Swanson says [17]:

". . . even though machines may never enjoy more than a partial success in library indexing, . . . people are even less promising."

To determine whether fully automatic text processing methods could compete in effectiveness with conventional manual retrieval operations and to identify those automatic techniques most helpful in this respect, a number of experiments have been performed over the last few years relating the Medlars retrieval system operating at the National Library of Medicine in Washington with the experimental SMART system.

Medlars operates in the conventional manner by having trained subject experts assign key words, or index terms, to all incoming documents and search requests [19]. The vocabulary is controlled by a printed thesaurus known as Mesh (Medical Subject Headings), and retrieval is effected by comparing the list of key words attached to the documents with a Boolean formulation of the query terms. Specifically, all documents which exhibit the appropriate combination of key words are retrieved in response to the corresponding query, while documents which do not exhibit this combination remain in the file. As is true of all key word retrieval systems, the retrieval process merely separates the stored collection into two parts in that it distinguishes the retrieved items from those not retrieved. No ranking is obtained for either the retrieved documents or the nonretrieved, and the user has no convenient way for identifying among a potentially large set of retrieved items those which are likely to be most helpful.

The SMART system, on the other hand, operates without any manual content analysis [20, 21]. Document excerpts—normally abstracts—and query texts are introduced instead into a computer, and a variety of automatic text analysis procedures is used to produce for each item a "concept vector" consisting of weighted terms or concepts representative of the document content. Typically, about one hundred different concepts might be used to identify a given document. Following a comparison between document and query vectors, a similarity or correlation coefficient is computed for each query-document pair, and documents are then submitted to the user in decreasing order of the corresponding similarity coefficients. Thus the user may choose to look only at the top retrieved item—the one presumed to be most relevant—or the top five, or the top ten.

A comparison between SMART and Medlars may be particularly appropriate in judging the effectiveness of automatic text processing methods in a retrieval environment, because Medlars represents a well-known conventional system which has been operating for many years on a large data base of several hundred thousand documents, while SMART includes automatic language analysis techniques and iterative, user-controlled search strategies to replace the intellectual input provided by trained indexers and searchers in conventional environments. Thus the systems are representative of the sophistication presently achievable in operational and automatic document retrieval environments respectively.

The design of the SMART-Medlars tests is covered in the remaining sections of this study together with the principal retrieval results.

4. *The SMART-Medlars Tests*

4.1 Introduction. If the results of any test procedure are to be taken as representative, it becomes necessary to maintain identical retrieval environments for both SMART and Medlars and to transfer to the SMART system the operational characteristics pertaining to Medlars. The following principal conditions must then be fulfilled:

(1) The queries to be used for test purposes must be user search requests actually submitted to and processed by the Medlars system.

(2) The test collection must consist of documents originally included in the Medlars data bank, chosen in such a way that any advance knowledge concerning the retrievability of any given document by either system is effectively ignored.

(3) The number of documents considered to be retrieved by SMART in response

to a given query (the cutoff value) must correspond to the Medlars cutoff, so that the same number of documents is used for evaluation purposes by both systems.

The evaluation parameters used in this study to assess retrieval performance are the well-known *recall* and *precision* measures which have served in the past for a number of similar investigations. Recall is defined as the proportion of relevant material actually retrieved, whereas precision is the proportion of retrieved material actually relevant. More specifically,

$$\text{recall} = \frac{\text{number of items retrieved and relevant}}{\text{total relevant items in the collection}} \quad (1)$$

$$\text{precision} = \frac{\text{number of items retrieved and relevant}}{\text{total number of retrieved items}} \quad (2)$$

Ideally, one wants to retrieve all relevant items, while at the same time withdrawing no extraneous ones, thereby achieving perfect recall and precision values equal to one. In practice, less than perfect output is normally obtained and values much less than one are the rule. Furthermore, the performance measures are not normally displayed for each individual user query. Instead, average values are computed over many queries to reflect the performance level for a random query submitted to the system.

The computation of recall and precision requires a knowledge not only of what is retrieved but also of what is relevant, or germane, to a given query. For the recall computation, in particular, information must be available concerning the identity and number of relevant items in the collection. For small test collections of the type used in the present experiments, it is often possible to obtain exhaustive relevance assessments of each document with respect to each query, prepared either by the query authors or by subject experts familiar with the subject area in question.

A summary of the characteristics of the Medlars collections used in the present experiments is contained in Table I. The first three collections and query sets are distinct, accounting for a total of 76 different user queries and 1,575 documents. The

TABLE I. SMART-MEDLARS TEST COLLECTIONS

Collection type	*Number of*		*Document origin*	*Type of relevance judgments*
	Queries	*Documents*		
Original Medlars	18	273	Chosen from among items previously used in an in-house evaluation of Medlars	Selective judgments by query authors available only for some of the retrieved documents
Extended Medlars	29	450	Obtained independently of SMART or Medlars by reference to the *Science Citation Index*	Exhaustive judgments obtained from nonauthor subject experts
Ophthalmology I	29	852	Obtained independently of SMART or Medlars from nine journals in ophthalmology	Exhaustive judgments from nonauthor subject experts
Ophthalmology II	17	852	Same as Ophthalmology I	Selective judgments by query authors for 10 items per query (5 retrieved by Medlars and 5 by SMART)
	76	1,575		

last collection, labeled "Ophthalmology II," is identical with "Ophthalmology I" except for the differences in the relevance assessments and in the number of queries used in the tests.

A number of earlier SMART-Medlars tests are reviewed in the next few sections, and the design of an expanded experiment using documents in ophthalmology is covered in detail [22, 23].

4.2 Original Testing Process. A first comparison between SMART and Medlars, performed some years ago, was based on 18 Medlars queries together with a subset of 273 documents. Both the queries and the documents had been used earlier for an in-house evaluation of the Medlars search system performed at the National Library of Medicine [24]. Each document incorporated in the collection had been judged for relevance with respect to one of the search queries by the corresponding query author; however, full relevance judgments of all documents with respect to all queries were not available.

The SMART process used originally consisted in keypunching the abstracts of all documents, and in automatically generating three distinct concept vectors for each document and query to represent information content:

(a) a *word form*, or suffix "s," vector consisting of weighted words extracted from the document abstracts, with final "s" endings removed;

(b) a *word stem* vector consisting of selected word stems with weights, obtained from the texts of the original document abstracts by a suffixing process designed to remove all normal suffixes;

(c) a *thesaurus* vector obtained by using a manually constructed dictionary, or thesaurus, which classifies related words or word stems into a thesaurus class; the document vectors consist of thesaurus class identifiers obtained from the original words or word stems by a dictionary look-up process.

Following the automatic analysis, query and document vectors are matched in the SMART system and a similarity coefficient is obtained for each query-document pair. A cutoff is then chosen for each query, equivalent to the exact number of documents retrieved by the corresponding Medlars query, and recall and precision values are computed. The average recall values obtained for the 18 test queries by Medlars and SMART respectively are shown in column 2 of Table II. It is seen that SMART produces a recall improvement of 8 to 12 percent compared with Medlars.[1]

The precision computations were unfortunately difficult to perform in the earlier test, because relevance assessments were not available for all documents retrieved by SMART with respect to each query, but only for the documents retrieved by Medlars. Typically, for a given query, about 5 documents out of 10 would be re-

[1] Statistical significance data are included for all recall-precision comparisons. These reflect the probability that under a hypothesis of statistical equivalence between the respective sets of recall or precision values, differences between samples as large as the ones observed would occur by chance. The *Wilcoxon signed rank test* (WSR) postulates that only the ranking of the differences between sample values is important, not the actual values of these differences; WSR further assumes that the two sample sets belong to the same family of distributions. The *sign test* (SIGN) takes into account neither the values nor the ranks of the differences between sample values but only their sign, and no assumptions are made concerning the distribution of the samples.

In either case, a probability output smaller than 0.05 is generally assumed to indicate that the differences in sample values are *statistically significant*, and the hypothesis of equivalence must then be rejected. Tests for which statistically significant performance differences were obtained are circled in the recall-precision tables.

TABLE II SMART-MEDLARS COMPARISON: ORIGINAL TEST
(Averages for 273 documents, 18 queries, SMART cutoff same as Medlars for each query)

Analysis methods	*Recall* % *Difference*	(*WSR*) (*SIGN*)	*Apparent precision* % *Difference*	(*WSR*) (*SIGN*)	*Adjusted precision* % *Difference*	(*WSR*) (*SIGN*)
Medlars	0.643		0.625		0.625	
SMART						
Word form (suffix "s")	0.704 +9%	(.3124) (.5000)	0.368 −41%	.0007 .0001	0 571 −9%	(.2174) (.3145)
Word stem	0.718 +12%	(.2385) (.5000)	0.367 −41%	0007 .0001	0.570 −9%	(1906) (3145)
Thesaurus	0.695 +8%	(3611) (.5000)	0 393 −37%	.0011 .0002	0 611 −2%	(.3973) (.5000)

trieved in common by SMART and Medlars. For each of the common 5, the relevance characteristics would be available; however, nothing would be known about the other 5 SMART items which were not also retrieved by Medlars. The "apparent precision" values shown in column 3 of Table II were obtained by assuming that all unknown items retrieved by SMART with respect to each query would be called nonrelevant, whereas the "adjusted precision" values of column 4 result from an assumption that the proportion of relevant items retrieved by SMART among the unknown documents would be the same as that among the known documents which had actually been judged for relevance.

It is seen from Table II that a small excess in SMART recall is compensated by a small deficiency in the adjusted precision, thus indicating that the order of magnitude of the performance parameters was approximately the same for the two systems. This is confirmed by the fact that the performance differences are in each case statistically nonsignificant. Unhappily, the precision adjustment used to obtain the final SMART precision values had to be based on an assumption whose validity could never be proved beyond doubt. For this reason, a series of new tests was undertaken in an effort to obtain a more detailed picture of the effectiveness of the various SMART text processing capabilities.

4.3 RANKING AND FEEDBACK SEARCHES. As a replacement for the linguistic analysis which appears too complex for a mechanized approach, the following three automatic techniques appear to be of most importance:

(1) The ranking feature incorporated into SMART makes it possible for the user to obtain access first to what is presumed to be most relevant; this suggests that the number of documents to be looked at—that is, the number retrieved—should be related to the magnitude of the corresponding query-document correlation.

(2) The ability to store a complete document collection and compute interdocument similarities for pairs of document vectors provides new ways of generating dictionaries and word lists to be utilized in analyzing document content.

(3) An iterative search strategy where information about items retrieved in an

earlier search serves to generate improved query formulations for use during subsequent search operations appears to provide great advantages in retrieval effectiveness.

Consider first the *ranking question.* The Boolean search formulations normally used in operational retrieval situations are not likely to produce optimal retrieval results because the number of retrieved documents does not normally depend on any existing or implied user requirement but rather on a query formulation over which the user may have little control. Thus for some queries, a very large number of documents may typically be retrieved, whereas for others nothing or very little is obtained. By forcing the SMART output into a Medlars framework, as is the case when each SMART search is limited to retrieving the same number of items as the corresponding Medlars query, an unreasonably restrictive search policy is used which may produce less effective retrieval results than originally expected.

A more reasonable process for the SMART system would consist in using a *correlation cutoff* by retrieving all those documents whose correlation coefficient with the query exceeds a given value. In order to make a comparison of results possible between Medlars and SMART, it is then necessary to choose the SMART correlation coefficient in such a way that the total number of documents retrieved over all the queries is identical with the total number retrieved by Medlars. However, for individual queries the number retrieved may be permitted to vary in accordance with the value of the query-document similarity.

The document ranking process is also essential for the implementation of the *relevance feedback* process which is used with the SMART system for the automatic construction of improved query formulations [25, 26]. An initial search is performed first using the query originally submitted. The user then submits relevance judgments for some of the documents retrieved early in the search, and a new query formulation is automatically constructed which will be more similar to the items identified as relevant and less similar to the nonrelevant items than the original query. Specifically, the feedback query is generated from the original query by addition or weight increases of terms from the documents termed relevant, and by a corresponding deletion or weight decreases of terms from the nonrelevant items.

The query updating formula actually used in the present experiments may be expressed in the following terms:

$$q_{i+1} = q_i + \sum_{1}^{10} \mathbf{r}_i - \sum_{1}^{2} \mathbf{s}_i , \tag{3}$$

where q_i is the query on the ith iteration, each $\mathbf{r}_i$ is a relevant document retrieved in the top 10 above cutoff, and each $\mathbf{s}_i$ refers to one of two nonrelevant retrieved above cutoff. That is, up to 10 relevant and up to two nonrelevant items participate in the feedback process, assuming that all are retrieved above cutoff. If the cutoff used should provide fewer than 12 retrieved documents, a correspondingly smaller number of items are used for feedback purposes.

It should be obvious that, in principle, as many feedback iterations can be carried out as desired; in practice, the return in terms of improved retrieval output will diminish as the number of feedback searches increases, so that two to three iterations will normally suffice.

The feedback operations incorporated into SMART are comparable to a language analysis operation in the sense that the generation of a feedback query amounts to reindexing or reformulating the query content. Furthermore, while it is possible in

theory to use feedback techniques in conventional systems—for example, by asking the requestor to rephrase a query using information contained in previously retrieved documents—it is not easy to come up with practical ways for implementing the idea. One immediate problem is the near-universal utilization of inverted file organizations in which the content description of a given document vector is scattered in many different parts of the file. In these circumstances, multiple file accesses are needed to retrieve complete document vectors, thus preventing relevance feedback-type operations for practical purposes.

4.4 Automatic Dictionary Construction. The construction of a thesaurus useful for document language normalization may be broken down into two parts: first, a choice must be made concerning the terms actually used for content description; and second, the chosen content words must be grouped into equivalence classes in such a way that all terms included in a given thesaurus class are represented in the document vectors by a common identifier. The first operation is used to eliminate from the content descriptions all terms which do not, in fact, represent document content, or which cannot serve to discriminate among the documents of a collection—typically, function words such as articles, prepositions, conjunctions, and so on, and other high-frequency words in the subject area under consideration. The subsequent grouping operation assembles into a common class synonymous or otherwise closely related terms, whose equivalence can thus be recognized during the text analysis operations.

A large number of procedures are available for the automatic construction of term classifications [13, 14, 27–31], and some of the proposed methodologies produce thesaurus classes which are not inferior in a retrieval environment to manually generated classifications. The main problem with the automatic classification techniques is the comparatively large cost in time and effort which must be expended to produce the automatic thesaurus. In many cases, the number of terms to be classified is of the order of some tens of thousands and the number of vector comparisons for n terms is of order n^2, or at any rate of order kn—too large for a practical utilization. Additional work is then needed with cheap, automatic classification techniques before these methods can become attractive in practice [32].

While the term grouping problem remains to be worked on, useful dictionaries can be constructed without any term grouping at all, by merely entering into the dictionary all those terms which are thought to be effective for purposes of content identification, or alternatively, by removing terms of low utility, sometimes called "common words." Classically, the identification of common words is an intellectual operation which involves a consideration of the semantic qualities of the individual words, as well as of the usage patterns of the words in the documents of a given collection. More often than not, a decision to use a given term for content identification, or alternatively to reject it, must in these circumstances be based on hunches or at best on informed guesses.

A number of experiments have been performed recently with a view to replacing the intellectual decision process normally required for the identification of common words by an automatic, controllable process [33, 34]. Specifically, the method is based on the notion that when a common word is assigned as a document identifier, the corresponding term will occur in many of the documents, thus rendering the documents more similar to each other; contrariwise, when a content word is assigned, it will serve to discriminate among the documents, thus producing individual document vectors which are less similar to each other.

This situation is reflected in the example of Table III. In the top part of Table III, each document in the space is represented by an X, and the distance between two X's is inversely proportional to the similarity of the corresponding document vectors. The left side of Table III represents the original situation where, for the sake of the example, three terms are assumed to be common to documents D_i and D_j. The correlation coefficient (measuring the number of common terms divided by the number of distinct terms) is then equal to 0.1764. In the center of the table, two common words are added to the document vectors, thus producing for D_i and D_j, 5 terms in common out of 12, or a correlation of 0.2631. Obviously, the document space is now more "bunched up." On the right side of Table III, two content words are added which are not likely to be the same for D_i and D_j. The overlap for the two documents is now reduced to 3 terms out of 12, producing a correlation of only 0.1428 and a space much more spread out than the original.

The strategy needed to distinguish common terms from content terms is now clear: a function is computed representing the density Q of the document space (Q might, for example, be measured as the sum of the correlations between each document and the center of the space). A new term i is now added to the document space; if the new density function Q_i is greater than Q, then term i is a common word, or a nondiscriminator, and the difference $Q_i - Q$ measures the strength of the term as a nondiscriminator. Contrariwise, if the space spreads out when the new term is added and Q_i is smaller than Q, then term i is a content word and $Q - Q_i$ measures its strength as a discriminator.

By ranking the discriminators in decreasing order of $Q - Q_i$, a list is obtained exhibiting the best discriminators first; similarly, an ordering in decreasing $Q_i - Q$ order for the nondiscriminators moves the "best" nondiscriminators on top. Such a list including the 12 highest discriminators as well as the 12 best nondiscriminators is shown in Table IV for a collection of 852 documents in the area of ophthalmology.

The automatic construction of a discriminator dictionary then requires the computation of a space density function for each term in the collection. Terms responsible for an increase in space density are identified as "common words" and are removed from the dictionary used to produce the document vectors. A typical dictionary construction process is illustrated in Table V for the previously mentioned ophthalmology collection.

Initially, all distinct words contained in the 852 abstracts of the ophthalmology collection are listed. When final "s" endings are deleted, the so-called "word form" dictionary is obtained. Terms of frequency one in the collection are deleted next, as are terms occurring in 25 percent or more of the documents, the notion being that terms of frequency one will not contribute to many query-document matches, whereas the high-frequency terms cannot discriminate among the documents. The discriminator detection algorithm is used next for the 5100 remaining terms, and about 200 additional terms are removed as nondiscriminators. Additional terms can be removed as shown in the last line of Table V, the deletion occurring in increasing discriminator order—that is, terms having the lowest value as discriminators are removed first.

In the next section, retrieval results are exhibited comparing the various automatic SMART procedures—correlation cutoff, automatic discriminator dictionaries, and feedback searches—with the conventional Medlars output.

4.5 General Retrieval Results. Two document collections in the area of medicine, chosen independently of either the SMART or the Medlars systems, are

TABLE III. DOCUMENT SPACE TRANSFORMATIONS

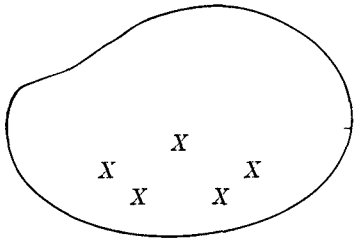
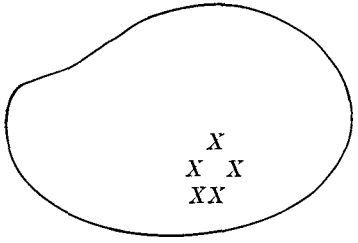
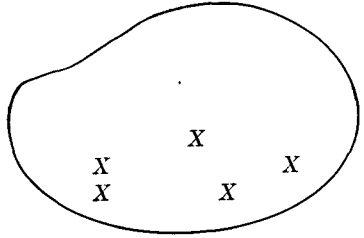
Original documents	*Documents after addition of nondiscriminators (common terms)*	*Documents after addition of discriminators (content words)*
1. Document space		
 original space	 more dense	 less dense
2. Document vectors D_i, D_j		
$(x_i^1, x_i^2, \cdots, x_i^k, \cdots, x_i^n, \cdots)$ $(x_j^1, x_j^2, \cdots, x_j^k, \cdots, x_j^n, \cdots)$	$(x_i^1, x_i^2, \cdots, x_i^k, \cdots, x_i^n, x_i^{n+1}, x_i^{n+2}, \cdots)$ $(x_j^1, x_j^2, \cdots, x_j^k, \cdots, x_j^n, x_j^{n+1}, x_j^{n+2}, \cdots)$	$(x_i^1, x_i^2, \cdots, x_i^k, \cdots, x_i^n, x_i^{n+1}, x_i^{n+2}, \cdots)$ $(x_j^1, x_j^2, \cdots, x_j^k, \cdots, x_j^n, x_j^{n+1}, x_j^{n+2}, \cdots)$
3. Similarity Computation		
(3 common terms out of 10) $S(D_i, D_j) = \frac{3}{10 + 10 - 3}$ $= .1764$	(5 common terms out of 12) $S(D_i, D_j) = \frac{5}{12 + 12 - 5}$ $= .2631$	(3 common terms out of 12) $S(D_i, D_j) = \frac{3}{12 + 12 - 3}$ $= .1428$

TABLE IV. HIGHEST RANKING DISCRIMINATOR AND NONDISCRIMINATOR ENTRIES (852 abstracts in ophthalmology)

Nondiscriminators				*Discriminators*			
Term	*Document frequency*	*Total frequency*	*Average frequency*	*Term*	*Document frequency*	*Total frequency*	*Average frequency*
1. Patient	201	408	2.03	1. Rubella	10	47	4.70
2. At	194	292	1.51	2. Capillary	19	54	2.84
3. Use	179	247	1.38	3. Laser	11	32	2.91
4. Have	194	257	1.32	4. Collagen	12	40	3.33
5. Retinal	134	275	2.05	5. Cyst	17	42	2.47
6. Present	184	219	1.19	6. Cholinesterase	6	26	4.33
7. Has	171	231	1.35	7. Fiber	16	50	3.13
8. Effect	150	259	1.73	8. Cyclodialysis	4	12	3.00
9. Result	179	234	1.31	9. Implant	18	36	2.00
10. Found	174	228	1.31	10. Uveitis	21	45	2.14
11. Report	141	172	1.22	11. Vessel	36	82	2.28
12. Occular	125	194	1.55	12. Spray	2	25	12.50

TABLE V. TYPICAL AUTOMATIC WORD LIST GENERATION (OPHTHALMOLOGY COLLECTION)

Dictionary generating process	*Automatic dictionary*	*Number of entries*	
		Removed	*Remaining*
Formation of word list from document abstracts following removal of final "s" endings	Word form (suffix "s")	—	8,672
Deletion of terms of frequency one	—	3,535	5,137
Deletion of terms occurring in 25% or more of the documents	Automatic word list	29	5,108
Deletion of terms automatically identified as nondiscriminators (minimum space density)	Automatic discriminator dictionary	180	4,928
Deletion of additional terms in decreasing nondiscriminator order	Reduced discriminator dictionary	3,928	1,000

used experimentally. The "extended Medlars" collection of 450 documents was obtained by picking entries from the *Science Citation Index* corresponding to documents previously identified as relevant by number of Medlars user queries [23]. The ophthalmology collection, on the other hand, includes articles published in 1967 in nine major journals in ophthalmology.[2] In each case, a verification procedure was used to ascertain that each document of the sample collections was in fact also included in the operational Medlars collection at the National Library of Medicine. Documents not so included must be removed before the beginning of the actual testing process.

A total of 58 queries originally submitted to the Medlars system by actual Medlars users was used for test purposes; of these, 29 related to ophthalmology and 29 to general biomedical topics. When research and professional people function as query authors, as is the case for most user queries submitted to the Medlars system, it is not practical to request full relevance assessments of each query with respect to each document from the query authors. For this reason, full relevance assessments had

[2] *Acta Ophthalmologica, American Journal of Ophthalmology, Archives of Ophthalmology, British Journal of Ophthalmology, Experimental Eye Research, Investigative Ophthalmology, Ophthalmologica, Transactions of the American Academy of Ophthalmology and Otolaryngology,* and *Vision Research.*

to be obtained from outside subject experts—a medical school student for the extended Medlars and a resident in ophthalmology for the ophthalmology collection.

The average retrieval results based on these nonauthor relevance judgments are summarized in Table VI for the extended Medlars collection, and in Table VII for ophthalmology. The following principal results are immediately apparent:

(1) When the number of retrieved items per query is exactly the same for SMART as for Medlars, the automatic analysis procedures included in SMART produce a deficiency in recall and precision ranging from 20 percent to almost 50 percent compared with the conventional Medlars indexing; the differences in the values are statistically significant in this case, so that the Medlars indexing is clearly superior.

(2) When a correlation cutoff is used by SMART, set in such a way as to retrieve the same total number of documents for the 29 queries as Medlars—127 documents for extended Medlars and 602 for ophthalmology—the deficiency of the SMART runs decreases to an average of about 10 percent in recall and precision, and the differences in performance are no longer statistically significant; this implies that the SMART and Medlars performance results are not sufficiently distinct to support a claim for the unequivocal superiority of either system.

(3) The SMART feedback searches produce improvements over the normal Medlars output ranging from a few percentage points for a single feedback search to 30 percent for two feedback iterations; in most cases, the performance differences are not statistically significant.

(4) The basic SMART word stem procedure which assigns weighted word stems to documents and search requests is already competitive with the controlled Medlars indexing when feedback searches are used, thus indicating that a word stem extrac-

TABLE VI. SMART-MEDLARS COMPARISON—EXTENDED MEDLARS COLLECTION
(450 documents, 29 queries; SMART correlation cutoff set to retrieve a total of 127 documents)

Analysis methods	*Medlars cutoff*				*Correlation cutoff*				*Two feedback searches*			
	Recall	*(WSR)*	*Precision*	*(WSR)*	*Recall*	*(WSR)*	*Precision*	*(WSR)*	*Recall*	*(WSR)*	*Precision*	*(WSR)*
	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*
Medlars (controlled terms)	.3117		.6110									
SMART												
Word form (suffix "s")	1814	.0021	3867	.0007	2613	(.2216)	.4960	(1302)	3525	(3383)	6740	(.2665)
	−42%	.0005	−37%	.0021	−16%	(3450)	−19%	(.1050)	+13%	(.5000)	+13%	(.5000)
Word stem	1814	.0013	.4141	.0009	.2622	(2420)	4901	.0494	3433	(3939)	.6892	(.2021)
	−42%	.0005	−32%	.0005	−16%	(4194)	−19%	.0392	+10%	(.5000)	+13%	(.3318)
Automatic discriminator dictionary A5	.2462	.0051	4518	.0032	2872	(3102)	.5879	(4056)	.3801	(.1870)	7230	(.1216)
	−21%	.0154	−26%	.0154	−8%	(2781)	−4%	(2706)	+22%	(.4223)	+18%	(.3238)
Thesaurus	.2181	.0060	.4512	.0018	.3232	(4946)	.6106	(3102)	4029	(.1578)	.7438	(0914)
	−30%	.0032	−26%	.0032	+4%	(3450)	0%	(.4223)	+29%	(.0849)	+22%	(.3450)

TABLE VII. SMART-MEDLARS COMPARISON—OPHTHALMOLOGY COLLECTION
(852 documents, 29 queries; nonauthor relevance assessments)

Analysis methods	*Medlars cutoff*				*One feeedback search*[1]				*Two feedback searches*[1]			
	Recall	*(WSR)*	*Precision*	*(WSR)*	*Recall*	*(WSR)*	*Precision*	*(WSR)*	*Recall*	*(WSR)*	*Precision*	*(WSR)*
	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*	% *Difference*	*(SIGN)*
Medlars (controlled terms)	.4272		.4454									
SMART												
Word form (suffix "s")	.2240	0007	.2728	0023	.4025	(4861)	.4144	(1239)	.4181	(3975)	.4839	(.3658)
	−48%	.0036	−39%	.0036	−6%	(.5000)	−7%	(0155)	−2%	(.4159)	+9%	(.2706)
Word stem	.2802	0056	.2932	.0089	.4125	(4741)	.4367	(2505)	.4318	(3365)	.4402	(.1379)
	−34%	.0207	−34%	.0207	−3%	(4159)	−2%	(.0388)	+1%	(.4159)	−1%	(.0539)
Automatic word list A4	.2843	.0134	3156	.0109	4251	(4310)	.5249	(.2821)	.4440	(3195)	6055	(.0892)
	−33%	0318	−29%	.0318	0%	(.5000)	+18%	(5000)	+4%	(5000)	+36%	(.5000)
Automatic discriminator dictionary A5	.3159	.0125	3039	0072	4624	(.2374)	.4041	(.1480)	4794	(.1650)	4456	(4144)
	−26%	0207	−32%	0207	+8%	(.1050)	−9%	(.5000)	+12%	(.1050)	0%	(.2207)
Thesaurus	3355	0149	3262	0182	4230	(4851)	4403	(.3102)	.4475	(3074)	.4727	(.4445)
	−21%	0318	−26%	0318	−1%	(.2517)	−1%	(4223)	+5%	(.0669)	+6%	(.4223)

[1] SMART feedback searches use correlation cutoff set to retrieve 602 items.

tion method can be used advantageously when supplemented by suitable search refinements.

(5) The SMART results obtained with the automatic discriminator dictionary and with a manually constructed thesaurus are competitive with the Medlars controlled indexing without any use of feedback techniques—assuming only that the SMART ranking is used to best advantage to control the output size; when the feedback techniques are added to the language normalization provided by the SMART dictionaries, improvements up to 30 percent are possible over the conventional Medlars performance.

The main conclusion derivable from the data of Tables VI and VII may be summarized as follows.

Simple word extraction from document abstracts or texts followed by Boolean searches of the type now implemented in many conventional retrieval environments are not likely to produce retrieval results equivalent in effectiveness to standard manual indexing techniques; however, a variety of different, generally nonlinguistic methods are easily implemented on a computer—including document ranking procedures, text normalization with stored dictionaries and thesauruses, and interactive feedback searches—which will produce retrieval results whose effectiveness exceeds that of the conventional manually controlled methodologies.

It should be noted that the absolute magnitude of the figures included in Tables

VI and VII is not indicative of any particular performance level, although the differences in the values for two or more methods or systems do accurately reflect the *relative* performance levels. The reason for this complication is the existence of *recall and precision ceilings* due to the fact that the number of documents retrieved for a given query is not, of course, the same as the number of relevant documents specified for that query by the outside relevance judge. The data of Table VIII indicate that for the 852 documents of the ophthalmology collection, a total of 602 were judged to be relevant to the 29 queries. When the number of documents identified as relevant to a given query is smaller than the Medlars cutoff—as is the case for query 1 of Table VIII—then the precision must necessarily remain below one; contrariwise, when the number of relevant is larger than the number retrieved, as it is for queries 3, 4, and 5, then a ceiling is imposed on the recall. For the ophthalmology collection, the average recall and precision ceilings lie between 70 and 75 percent, indicating that the best automatic retrieval results of Table VII reach about 65 percent of the maximum attainable recall and 83 percent of the possible precision.

TABLE VIII. Recall and Precision Ceilings—Ophthalmology Collection
(Nonauthor judgments; 852 documents, 29 queries)

Query number	*Number retrieved by Medlars*	*Number judged relevant*	*Maxium relevant retrievable*	*Recall ceiling*	*Precision ceiling*
1	20	14	14	1.0000	.7000
2	11	1	1	1 0000	.0909
3	5	6	5	.8333	1.0000
4	36	92	36	.3913	1.0000
5	8	60	8	.1333	1.0000
6	16	10	10	1.0000	.6250
7	54	59	54	.9153	1.0000
8	6	2	2	1.0000	.3333
9	19	15	15	1.0000	.7895
10	8	8	8	1 0000	1.0000
11	37	5	5	1.0000	.1351
12	12	23	12	.5217	1.0000
14	5	1	1	1.0000	.2000
15	7	53	7	.1321	1.0000
16	28	24	24	1.0000	.8571
17	20	1	1	1.0000	.0500
19	17	14	14	1.0000	.8235
20	11	54	11	.2037	1.0000
21	12	7	7	1.0000	.5833
22	10	54	10	.1852	1 0000
23	6	65	6	.0923	1.0000
24	10	43	10	.2326	1.0000
25	10	33	10	.3030	1.0000
26	11	12	11	.9167	1.0000
27	5	1	1	1.0000	.2000
28	174	149	149	1.0000	.8563
29	8	1	1	1.0000	.1250
31	6	4	4	1.0000	.6667
34	30	44	30	.6818	1.0000
			Average ceiling	0.7428	0.7254

4.6 USE OF SELECTED AUTHOR RELEVANCE JUDGMENTS. The results of the previous section were obtained by using relevance assessments produced by non-author subject experts. When relevance assessments are obtained from query authors directly, the retrieval evaluation must be adjusted since complete relevance information is not then normally available. On the other hand, when only some items are actually judged for relevance, difficulties in interpreting the results of the kind previously mentioned for the output of Table II may arise.

An attempt was made to overcome the problem by submitting to each of the 29 ophthalmology query authors 10 of the documents retrieved previously in response to his query. Specifically, 5 documents were randomly chosen from among those retrieved by the Medlars search, and 5 other documents were randomly chosen from among those obtained by the standard SMART word stem run. Each author was then asked to assess the relevance of each of the 10 documents with respect to his query. Three different answers were possible: the item is of major value in relation to the author's information need, of minor value, or of no value. Results were tabulated for the case where both major and minor value items are considered to be relevant and for the additional case where only the major value items are assumed relevant. Usable answers were actually obtained for 17 out of 29 queries; the results in Tables IX, X, and to XI thus correspond to the performance of only 10 documents out of 852 for each of 17 user queries.

When only partial relevance assessments are available, the recall and precision measures cannot be computed in the normal manner using eqs. (1) and (2). Instead, a precision value must be calculated derived solely from the 10 available documents

TABLE IX. RELATIVE RECALL COMPARISON—OPHTHALMOLOGY COLLECTION
(Author relevance judgments—5 SMART and 5 Medlars documents averaged for 17 queries)

Analysis methods	*Word form*		*Word stem*		*Automatic word list*		*Automatic discriminator list*		*Thesaurus*	
	Recall Percent	*(WSR) (SIGN)*	*Recall Percent*	*(WSR) (SIGN)*	*Recall Percent*	*(WSR) (SIGN)*	*Recall Percent*	*(WSR) (SIGN)*	*Recall Percent*	*(WSR) (SIGN)*
(1) Medlars cutoff										
Medlars	2490	(3283)	2294	(4070)	.1931	(1870)	1274	(.2871)	.2568	(2108)
SMART	2774	(2744)	2509	(3872)	.2794	(2744)	.1696	(1445)	.3068	(5000)
	+11%		+9%		+45%		+33%		+19%	
(2) Correlation cutoff										
Medlars	1666	(0486)	2107	(2386)	1588	(1540)	.1372	(1013)	.1294	(0092)
SMART	2941	(0195)	2676	(2539)	2588	(1719)	.2735	(1719)	.3264	(0327)
	+76%		+27%		+61%		+100%		+151%	
(3) One feedback search										
Medlars	1666	(1432)	1519	(1871)	1588	(0333)	.1078	(0515)	.1294	(0047)
SMART	2372	(2539)	1960	(2539)	2754	(.0547)	.2696	(0547)	.3254	(.0107)
	+42%		+29%		+73%		+150%		+151%	
(4) Two feedback searches										
Medlars	1960	(.0712)	1519	(0618)	.1588	(0333)	1078	(.0515)	.1294	(.0047)
SMART	2754	(.2744)	1960	(2539)	.2754	(.0547)	2696	(.0547)	.3254	(.0107)
	+41%		+29%		+73%		+150%		+151%	

TABLE X. PRECISION COMPARISON—OPHTHALMOLOGY COLLECTION
(Author relevance judgments—10 documents per query averaged for 17 queries)

Analysis methods	*Word form*		*Word stem*		*Automatic word list*		*Automatic discriminator list*		*Thesaurus*	
	Recall *Percent*	*(WSR)* *(SIGN)*	*Recall* *Percent*	*(WSR)* *(SIGN)*	*Recall* *Percent*	*(WSR)* *(SIGN)*	*Recall* *Percent*	*(WSR)* *(SIGN)*	*Recall* *Percent*	*(WSR)* *(SIGN)*
(1) Medlars cutoff Medlars precision *0.4823*	.3523 −27%	.0067 .0037	.3705 −23%	.0240 .0287	3235 −33%	0087 .0017	2000 −59%	.0002 0000	.3117 −35%	0024 0021
(2) Correlation cutoff Medlars precision *0.4823*	.2764 −43%	.0034 0032	.2882 −40%	.0039 0112	.3058 −37%	0093 .0032	.2352 −51%	.0007 .0001	.2823 −41%	0032 0009
(3) One feedback Medlars precision *0.4823*	2705 −44%	.0014 .0065	.2705 −44%	.0009 .0037	.3294 −32%	.0109 .0287	.2588 −46%	.0009 0017	.2882 −40%	.0032 0009

for each query, and a new definition of the recall becomes necessary. The following precision computations were actually used:

$$\text{Medlars precision} = \frac{\text{total number of relevant retrieved by Medlars (of the 10 judged for relevance)}}{\text{total retrieved judged for relevance (equal to 10)}} \quad (4)$$

$$\text{SMART precision} = \frac{\text{total number of relevant retrieved by SMART (of the 10 judged for relevance)}}{\text{total retrieved judged for relevance (equal to 10)}} \quad (5)$$

While a fairly standard procedure could be used for the precision, an estimation of the recall is more complicated because a knowledge of the total number of documents that would be judged relevant by the query authors is lacking. A *relative recall* measure is therefore defined which relates the number of relevant retrieved by Medlars to the number of relevant retrieved by SMART, and vice versa.[3] Specifically, of the 5 documents retrieved by Medlars which had been judged for relevance, consider the M that are judged relevant by the query author. Let S_M of the M items be retrieved by SMART; then the SMART relative recall is defined as S_M/M. Similarly, consider the 5 items retrieved by SMART for which relevance assessments are on hand. Let S of these be called relevant and assume that M_S of these S items are also retrieved by Medlars; then the Medlars relative recall is defined as M_S/S.[4] In other words,

$$\text{SMART relative recall} = \text{Number of relevant retrieved by SMART out of the total relevant previously retrieved by Medlars} = S_M/M \quad (6)$$

[3] The use of the relative recall was suggested by M. E. Lesk of Bell Laboratories.

[4] Queries for which either M or S equals zero are not included in the set of 17 corresponding to the computations of Tables IX to XI.

TABLE XI. RELATIVE RECALL AND PRECISION—OPHTHALMOLOGY COLLECTION
(Author relevance judgments—*major* relevant only averaged for 17 queries)

Analysis methods		*Word form*		*Word stem*		*Automatic word list*		*Automatic discriminator list*		*Thesaurus*	
(1) Medlars cutoff											
Recall:	Medlars	.1245	.0697	1294	(1311)	.1470	(.1206)	.0784	(1473)	1666	(.0865)
	SMART	2784	0547	2098	(3769)	2705	(3770)	1607	(1094)	.3176	(.0898)
	Percent	+124%		+62%		+84%		+105%		+91%	
Precision:	Medlars	.2588	0442		0119		0011		.0011		.0060
	SMART	.1705	0176	1529	.0065	.1411	0002	.1000	.0005	.1588	.0032
	Percent	−34%		−41%		−45%		−61%		−39%	
(2) Correlation cutoff											
Recall:	Medlars	.0539	0086	0588	(0212)	.1127	(0963)	0882	(0253)	.1176	0332
	SMART	.3333	.0351	2843	(.1445)	2333	(2539)	.3186	(0898)	.3235	.0195
	Percent	+518%		+384%		+107%		+261%		+175%	
Precision:	Medlars	.2588	.0093		.0017		0040		.0011		.0054
	SMART	.1411	.0032	.1176	.0005	.1352	.0037	.1352	.0002	.1529	.0032
	Percent	−45%		−23%		−48%		−48%		−41%	
(3) One feedback search											
Recall:	Medlars	.0539	.0086	.0489	0086	.0930	.0076	.0588	.0122	.1176	.0332
	SMART	.3627	.0352	.3264	.0352	3509	0195	2794	.0195	.3431	0195
	Percent	+573%		+567%		+277%		+375%		+192%	
Precision:	Medlars	.2588	0115		.0054		.0192		.0011		.0072
	SMART	.1470	.0032	.1156	0017	.1705	0065	.1352	.0002	1588	.0059
	Percent	−43%		−55%		−34%		−48%		−39%	

$$\text{Medlars relative recall} = \frac{\text{Number of relevant retrieved by Medlars out of the total relevant previously retrieved by SMART}}{} = M_S/S \qquad (7)$$

Since the Medlars relative recall depends in each case on the number of relevant retrieved by SMART, a different Medlars recall value will correspond to each of the SMART retrieval runs.

A comparison of relative recall values, using both minor and major relevance, is shown in Table IX. It may be seen that when the normal Medlars cutoff is used, the SMART relative recall is about 25 percent better on the average than the Medlars average recall. The improvement reaches 80 to 90 percent on the average for the more sophisticated SMART methods, such as correlation cutoff and feedback searches, and the differences then are statistically significant. While the best SMART results are obtained once again for the dictionary and thesaurus runs, the feedback technique does not appear to provide much improvement in output in this case. The reason, of course, lies with the fact that the relevant documents previously retrieved which are used to construct the improved feedback queries must belong to the set of 5 actually judged for relevance. Unfortunately, these 5 were not chosen to be the top 5 retrieved by SMART, but rather a random set of 5. It is seen that the SMART feedback process which depends on the use of the document ranking feature does not operate advantageously when randomly ranked items are used for query modification.

Table X shows an average deficiency of about 30 to 40 percent for the SMART precision compared with the Medlars precision of 0.4823. The problem here again is that the SMART ranking feature is not used in the choice of the documents evaluated for relevance by the query authors. However, whereas the recall advantage of SMART grows from about 10 percent for the initial word stem run to over 150 percent for most of the thesaurus runs, the precision disadvantage remains fairly constant overall.

When major relevance only is taken into account, the same pattern is accentuated as shown by the output of Table XI. A constant precision deficiency of about 40 percent for SMART is compensated by a relative recall advantage increasing from an average of about 90 percent for the standard Medlars cutoff to about 290 percent on the average for the correlation cutoff and almost 400 percent for one iteration of feedback search. Most of the performance differences included in Table XI are statistically significant. Once again, the more sophisticated SMART procedures improve the recall output without simultaneously causing a greater loss in precision.

4.7 Conclusions. Average performance differences with Medlars are shown for a variety of SMART search methods in Table XII, and for the several SMART dictionaries in Table XIII. The information in Tables XII and XIII is derived from Tables VI, VII, IX, X, and XI, and is averaged for 5 SMART dictionaries in Table XII and for various search strategies in Table XIII. The following main conclusions appear warranted:

(1) The SMART ranking procedures as well as the feedback search methods produce considerable performance improvements over the Boolean search output used in conventional systems (see Table XII).

TABLE XII. SMART–Medlars Performance Differences
(Averages for several SMART dictionaries)

SMART Search Process	*Extended Medlars collection*		*Ophthalmology collection*					
			Nonauthor judgments		*Author judgments (all relevant)*		*Author judgments (major relevant only)*	
	R	*P*	*R*	*P*	*R*	*P*	*R*	*P*
Medlars cutoff	−33.75%	−30 25%	−32 4%	−32 0%	+23.4%	−35.4%	+93 2%	−44.0%
Correlation cutoff	−9.00%	−10 50%	−26 2%	−26 8%	+83.0%	−42 4%	+289 0%	−41.0%
One feedback search	+13.75%	+12 25%	−0.4%	−0 2%	+89 0%	−41 2%	+396.8%	−43.8%
Two feedback searches	+18 50%	+17 25%	+4 0%	+10.0%	+88 8%	−40.0%	—	—

TABLE XIII. SMART–Medlars Performance Differences
(Averages for several SMART search methods)

SMART dictionary	*Extended Medlars collection*		*Ophthalmology collection*					
			Nonauthor judgments		*Author judgments (all relevant)*		*Author judgments (major relevant only)*	
	R	*P*	*R*	*P*	*R*	*P*	*R*	*P*
Word form (suffix "s")	−15.00%	−14.33%	−18 66%	−12 33%	+42.50%	−38 00%	+405 00%	−40.66%
Word stem	−16.00%	−12 66%	−12 00%	−12.33%	+23 50%	−35 66%	+337.66%	−39.66%
Automatic word list	—	—	−9.66%	+8.33%	+63.00%	−34.00%	+156 00%	−42 33%
Automatic discriminator dictionary	−2.33%	−3.00%	−2 00%	−13 66%	+108.25%	−52.00%	+247.00%	−52.33%
Thesaurus	+1.00%	−1.33%	−5.66%	−7 00%	+118.00%	−38 66%	+152.66%	−39 66%

(2) The feedback searches implemented by SMART lead to considerable gains over the standard Medlars search output (see Table XII).

(3) The automatic SMART discriminator dictionary and the thesaurus produce better performance than the word or word stem extraction process alone (see Table XIII).

(4) The SMART language normalization methods which are used to produce dictionaries and thesauruses lead to retrieval results at least equivalent in average effectiveness to the conventional manual indexing (see Table XIII).

(5) Future retrieval systems might use vector matching techniques leading to ranked output, as well as interactive search techniques for the formulation of more effective query statements.

(6) The standard syntactic and semantic language analysis techniques and the intellectual input conventionally provided by expert indexers might be replaced by automatically constructed analysis tools derived from existing document collections.

ACKNOWLEDGMENT. The writer gratefully acknowledges the assistance of Dr. Joseph Leiter and Mr. Constantine J. Gillespie of the National Library of Medicine, Professor F. W. Lancaster of the University of Illinois, and Dr. M. E. Lesk of Bell Laboratories, all of whom were helpful in generating the SMART-Medlars test design and the subsequent evaluation process.

REFERENCES

1. GARVIN, P. L., et al. Some opinions concerning linguistics and information processing. Rep. PB 190 639, Center for Applied Linguistics, May 1969. Available from National Technical Information Service, Washington, D.C.
2. EDMUNDSON, H. P. New methods in automatic extracting. *J. ACM 16*, 2 (Apr. 1969), 264–285.
3. PACAK, M., AND PRATT, A. W. The function of semantics in automated language processing. Symp on Information Storage and Retrieval, U. of Maryland, Apr. 1971.
4. BAXENDALE, P. An empirical model for machine indexing. Third Institute on Information Storage and Retrieval, American U , Washington, D.C., Feb. 1961, pp. 207–218.
5. CLARKE, D. C , AND WALL, R E. An economical program for the limited parsing of English, Proc AFIPS 1965 FJCC, Vol. 27, Pt. 1, Spartan Books, New York, pp. 307–319.
6. DAMERAU, F J. Automatic parsing for content analysis. *Comm. ACM 13*, 6 (June 1970), 356–360.
7. RUSH, J. E., SALVADOR, R., AND ZAMORA, A. Automatic abstracting and indexing: Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *J. ASIS 22*, 4 (July–Aug. 1971), 260–274.
8. SALTON, G. Automatic text analysis. *Science 168*, 3929 (17 Apr. 1970), 335–343.
9. CLEVERDON, C. W., AND KEEN, E. M. Factors determining the performance of indexing systems; Vol. 2—test results. Aslib Cranfield Res. Proj., Cranfield, England, 1966.
10. SALTON, G., AND LESK, M. E. Computer evaluation of indexing and text processing. *J. ACM 15*, 1 (Jan. 1968), 8–36.
11. SALTON, G. Automatic processing of foreign language documents. *J. ASIS 21*, 3 (May–June 1970), 187–194.
12. DENNIS, S. F. The design and testing of a fully automatic indexing-searching system for documents consisting of expository text. In *Information Retrieval—A Critical View*, G. Schecter, Ed., Thompson Book Co., Washington, D.C., 1967.
13. GIULIANO, V. E , AND JONES, P E. Study and test of a methodology for laboratory evaluation of message retrieval systems. Rep. ESD-TR-66-405, Arthur D. Little, Cambridge, Mass., 1966.
14. SPARCK JONES, K. *Automatic Keyword Classification for Information Retrieval.* Butterworth and Co., London, 1971.

15. STEVENS, M. E. Automatic indexing: A state of the art report. NBS Monograph 91, U.S. Bureau of Standards, Washington, D.C., March 1965.
16. STEVENS, M E., GIULIANO, V. E., AND HEILPRIN, L. B. Statistical association methods for mechanized documentation. NBS Misc. Pub. 269, U. S. Bureau of Standards, Washington, D.C , Dec. 1965
17. SWANSON, D. R. Searching natural language text by computer. *Science 132*, 3434 (Oct. 21, 1960), 1099–1104.
18. SWANSON, D. R. Interrogating a computer in natural language. Proc. IFIP Cong. 1962, North-Holland Publishing Co., Amsterdam, p. 288–393.
19. The Principles of Medlars. National Library of Medicine, Bethesda, Md., 1970. Available from Superintendent of Documents, Washington, D.C.
20. SALTON, G. *Automatic Information Organization and Retrieval.* McGraw-Hill, New York, 1968.
21. SALTON, G. *The SMART Retrieval System—Experiments in Automatic Document Processing* Prentice-Hall, Englewood Cliffs, N.J , 1971.
22 SALTON, G. A comparison between manual and automatic indexing methods. *American Documentation 20*, 1 (Jan 1969), 61–71
23. SALTON, G. A new comparison between conventional indexing (Medlars) and automatic text processing (SMART). *J.ASIS 23*, 2 (March–April 1972), 75–84.
24. LANCASTER, F. W. Evaluation of the Medlars demand search service. National Library of Medicine, Bethesda, Md., Jan. 1968.
25 SALTON, G. Search and retrieval experiments in real-time information retrieval. In *Information Processing 68* (Proc IFIP Cong.), North-Holland Publishing Company, Amsterdam, 1969, pp. 1082–1093.
26. SALTON, G The performance of interactive information retrieval. *Information Processing Letters 1*, 2 (July 1971), 35–41.
27. BORKO, H. The construction of an empirically based mathematically derived classification system. Rep. SP-588, System Development Corp., Santa Monica, Calif., Oct. 1961.
28. AUGUSTSON, J G., AND MINKER, J An analysis of some graph theoretical cluster techniques. *J ACM 17*, 4 (Oct 1970), 571–588.
29. DOYLE, L. B. Breaking the cost barrier in automatic classification, Rep. SP-2516, System Develpment Corp., Santa Monica, Calif , July 1966.
30. GOTLIEB, C. C., AND KUMAR, S. Semantic clustering of index terms. *J. ACM 15*, 4 (Oct. 1968), 493–513.
31. DATTOLA, R. T. Experiments with a fast algorithm for automatic classification. In *The SMART Retrieval System—Experiments in Automatic Document Processing*, G. Salton, Ed., Prentice-Hall, Englewood Cliffs, N J., 1971
32. JOHNSON, D. B., AND LAFUENTE, J. M. A controlled single-pass classification algorithm with application to multi-level clustering. Sci. Rep. ISR-18, Sec. XII, Dept. of Computer Science, Cornell U., Ithaca, N.Y., Oct 1970.
33. BONWIT, K., AND ASTE TONSMAN, J. Negative dictionaries. Sci. Rep. ISR-18, Sec. VI, Dept. of Computer Science, Cornell University, Ithaca, N Y., Oct. 1970.
34. SALTON, G. Experiments in automatic thesaurus construction for information retrieval. Proc. IFIP Congress 71, Ljubljana, North-Holland Publishing Co., Amsterdam, 1972, pp. 115–123.

RECEIVED DECEMBER 1971; REVISED MAY 1972