A New Comparison Between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART)

G. Salton*

TR 71-115

Abstract

A new testing process is described designed to compare conventional retrieval (MEDLARS) and automatic text analysis methods (SMART). The results obtained with a collection of documents chosen independently of either SMART or MEDLARS indicate that a simple automatic extraction of keywords from document abstracts produces a 30 to 40 percent loss compared with MEDLARS indexing. A replacement of the unranked Boolean searches used in MEDLARS by the standard ranked output normally provided by SMART reduces the loss to between 15 and 20 percent. When an automatically generated word control list or a thesaurus is used as part of the SMART analysis, the results are comparable in effectiveness to those obtained by the intellectual MEDLARS indexing. Finally, the incorporation of user feedback procedures into SMART furnishes an improvement over the normal MEDLARS output of 15 to 30 percent.

One concludes again that no technical justification exists for maintaining controlled, manual indexing in operational retrieval environments.

---

*Department of Computer Science, Cornell University, Ithaca, New York 14850

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

1. Introduction

Over the last few years, a great deal of time and effort were devoted to experimentation dealing with the indexing and analysis of information content, and a number of studies were undertaken of methods for performing the analysis procedures automatically. As a result, considerable controversy has been created concerning the relative merits of controlled versus free language indexing, and manual versus automatic analysis methodology, respectively.

These questions, in fact, involve considerations other than the normal technical ones, in part because of the large financial stakes — every library and information center is vitally concerned with indexing and cataloging costs — and in part because of an emotional factor which emerges whenever any computerized procedure is suggested as a replacement for intellectual work performed by human beings. It is then not surprising that the evidence which has accumulated concerning the feasibility of mechanized indexing procedures has so far produced relatively few practical consequences.*

In fact, many observers feel that regardless of any particular test or evaluation output which may exist, an automatic analysis of document content appears improbable and unappealing:

> "... as a solution of the indexing problem for the whole range of the bibliographic record, the computer technique remains crude, makeshift and inadequate in my view" [2];

---

*For a survey of work in automatic text analysis, see for example reference [1].

> "this question of whether we shall in the future be using controlled or free languages in retrieval systems is, it seems, one of the fundamental issues; we seem at present to have little evidence of the superiority of one method over the other" [3].

The problem is complicated also by the existence of certain conventional special-purpose classification systems which are said to furnish nearly perfect results in their environments, and for which a substantial superiority over some automatic text processing methods is claimed. [4,5]

Since it is obviously not the case, that any automatic text manipulation method — no matter how ineptly used — will automatically produce gains over the conventional, intellectual indexing methodology, it is worth making an attempt to isolate those features in automatic text processing which appear to be of most value in indexing and retrieval Accordingly, the remainder of this study is concerned with a comparison between the MEDLARS retrieval system, based on the controlled indexing vocabulary in use at the National Library of Medicine [6], and the automatic SMART document retrieval system. [7,8] The design of the test procedure is covered in detail, and the several language processing features incorporated into the SMART system are individually evaluated.

2. The SMART - MEDLARS Test Design

A) General Considerations

When attempting a comparison between automatic text analysis and manual indexing procedures, the SMART and MEDLARS systems appear to be particularly appropriate. The MEDLARS system has been operating for many years on a large data base of several hundred thousand document citations in the medical area, using a controlled indexing vocabulary applied by trained indexers and searchers. The SMART system, on the other hand, includes sophisticated language analysis methods, and iterative, user-controlled search strategies, as opposed to the simple word matching used in many other automatic systems. Furthermore, both systems have previously undergone extensive tests, using the well-known <u>recall</u> and <u>precision</u> parameters for evaluation purposes.* [6,7,8]

While SMART and MEDLARS are then representative of the presently achievable state-of-the-art in automatic and conventional document analysis, an attempt to compare the effectiveness of the respective retrieval capabilities raises complicated questions regarding test design and evaluation criteria. In particular, it is necessary to choose a test collection of reduced size together with a representative set of user queries in such a way that no bias toward either system is introduced, while making it possible nevertheless to obtain reliable recall and precision measurements.

In an earlier test of SMART and MEDLARS, based on the performance of 18 user queries together with 273 documents, not all of the design problems could be successfully overcome. [9] On that occasion, average

---

*Recall is the proportion of relevant material actually retrieved, while precision is the proportion of retrieved material actually relevant. A perfect system which retrieves everything wanted, while simultaneously rejecting everything extraneous would be characterized by recall and precision values equal to 1.

recall figures of 0.64 and 0.72 were found for MEDLARS and SMART, respectively. Unhappily, the precision values for SMART could not be directly calculated, because relevance assessments were available only for those documents actually retrieved by MEDLARS in response to each query, but not for all documents retrieved by SMART. An assumption therefore had to be made concerning the relevance, or nonrelevance, of some of the items retrieved by SMART, leading to an average precision of 0.63 for MEDLARS, and an "adjusted" average precision of 0.57 for SMART.* This precision adjustment has been the subject of some criticism, and the resulting test results, indicating a small recall superiority for SMART and a small precision advantage for MEDLARS were never felt to be completely believable. [10]

To determine whether the automatic procedures incorporated into the SMART system were competitive with the more conventional MEDLARS indexing and search procedures, it has therefore become necessary to undertake a new SMART-MEDLARS comparison under more favorable test conditions. The following principal test criteria appeared to be of most importance:

a) the queries to be used for test purposes must be user search requests actually submitted to and processed by the MEDLARS system;

*The assumption was that the proportion of relevant retrieved by SMART among the "unknown" documents (those without relevance assessments) would be the same as that retrieved among the known documents that had previously been judged for relevance.

b) the test collection must consist of documents originally included in the MEDLARS data bank chosen in such a way that each document is germane to at least one of the query topics; in these circumstances each document will therefore possess a substantial a priori likelihood of being retrieved in response to one or more of the test queries;

c) these potentially retrievable documents must be chosen independently of either SMART or MEDLARS, and without prior knowledge of the retrieval results obtainable by either system with these documents;

d) full relevance assessments must be available for each document with respect to each query, to permit a direct computation of recall and precision figures.

A description is given in the next few paragraphs of the manner in which the test collection and the relevance assessments were obtained.

B) Collection Generation and Relevance Assessments

The main requirement was to obtain a set of MEDLARS queries together with a MEDLARS subset of potentially relevant documents chosen without bias toward either retrieval system. The following procedure was actually used:

a) a set of 30 search requests previously used for the in-house evaluation of MEDLARS [6] was chosen, and the MEDLARS search results consisting of lists of documents retrieved by MEDLARS were obtained;

b) for each of the 30 queries, one or more of the documents previously identified as relevant by the respective MEDLARS users during the original MEDLARS test were chosen as starting items;

c) these relevant documents were then used as entry points to the 1964, 1965 and 1966 issues of the Science Citation Index (SCI), and new documents citing the original relevant ones were obtained from the Citation Index;

d) for each query, a total of fifteen SCI citations were identified, the assumption being that these documents must be potentially relevant to the query, since each of them cites a document previously known to be relevant;

e) a check was made to insure that each SCI document constituting the new extended MEDLARS collection was in fact included in the full MEDLARS collection at the time the original MEDLARS searches were performed; documents not so included were replaced by new ones from the Science Citation Index;

f) the abstracts pertaining to the resulting 450 documents were keypunched, and the usual SMART text processing methods were utilized to analyze query and document texts, and to perform the search and retrieval operations.

In order to obtain the normal performance indications in terms of recall and precision, it became necessary to generate full relevance judgments of each of the 450 documents with respect to the 30 test queries. Unfortunately, the original MEDLARS requestors could not be relied upon to perform additional work in connection with queries that had been processed several years earlier. For this reason, a medical school student was hired in order to serve as an assessor of relevance for the extended MEDLARS collection. Before actually starting with the judging process, the new judge was given for assessment 6 of the original MEDLARS queries — different from the 30 queries used in the present test — together with a collection of MEDLARS documents for which the relevance characteristics had been determined by the respective MEDLARS users during the earlier MEDLARS test.

It is seen from Table 1 that the average amount of overlap in the relevant document sets obtained for the six sample queries (labelled A through F in Table 1) between the original MEDLARS users and the new judge was almost 69 percent. This is a much higher agreement than can be expected for random sets of user populations, and indicates that the new assessor can safely serve as a substitute during the judging process.*

The relevance characteristics obtained for the thirty test queries by the new judge are detailed in Table 2. It may be seen that for one query (query 8), no relevant documents were identified by the assessor, and no documents were retrieved by MEDLARS in response to this query. Since recall and precision values can thus not be computed, query 8 was dropped from the evaluation process, and recall and precision results exhibited in this report are computed for 450 documents averaged over 29 queries.

---

*An earlier complete evaluation of the relevance judging process indicates that for an average relevance agreement of only 31 percent between original requestors and subject experts, obtained for 48 queries and 1200 documents, the differences in average recall and precision values is almost everywhere less than one percent, with a maximum difference of less than five percent. [11]. The earlier study also gives reasons why averaged recall-precision evaluation results remain generally invariant for normal differences in the document relevance characteristics.

Query Number	Number of Relevant MEDLARS User	Number of Relevant New Judge	Number Relevant in Common	Percent Overlap
A	15	12	12	80%
B	12	13	9	56%
C	22	15	15	68%
D	11	12	10	77%
E	9	7	6	60%
F	6	6	5	71%
			Average Overlap	68.67%

Verification of Relevance Assessments

Table 1

Table 2 also indicates that whereas a total of 127 documents included in the test collection were retrieved by the MEDLARS system in response to the 29 queries, the relevance judge identified a total of 284 items as relevant. In order to obtain a proper performance comparison, the SMART system searches must then produce a total of 127 documents over the 29 test queries. For either system, a recall ceiling of 127/284 thus exists, and the maximum possible recall (number of relevant retrieved divided by total relevant in collection) will therefore equal 0.4471. The maximum precision attainable (number of relevant retrieved divided by total retrieved) will, however, remain equal to 1.0000 as usual.

3. Basic Word Extraction versus Controlled Indexing

The SMART retrieval procedures differ from the normally implemented methodologies in several important respects. First, the documents and search requests are processed by the SMART system without any prior manual analysis, using one of several different programmed automatic content analysis methods. Second, an index of similarity, or correlation coefficient, is computed between each document and the corresponding query, as a function of the amount of overlap between analyzed documents and queries; this makes it possible to order the output document citations in decreasing correlation order so as to permit the user to look at the most promising items first. Third, feedback operations are implemented, designed to construct better queries, more likely to represent the user's need, based on information

Query Number	Documents Retrieved by MEDLARS	Total Number of Relevant Documents
1	13	12
2	1	9
3	7	6
4	0	8
5	2	12
6	1	4
7	1	10
8*	0	0
9	12	6
10	6	13
11	9	9
12	0	4
13	9	9
14	6	10
15	0	13
16	0	7
17	12	10
18	8	14
19	3	12
20	5	13
21	6	12
22	0	5
23	6	12
24	3	10
25	0	12
26	9	14
27	0	4
28	4	14
29	1	11
30	3	9
30	127	284

(Query 8 was dropped because of no relevant items in collection.)

Query Characteristics

Table 2

supplied by the user to the system following a previous search operation.

Consider first the simple word extraction procedures. The following principal steps are involved:

a) the individual words included in document abstracts and in query texts are isolated;

b) certain words, such as function words listed in a negative dictionary are removed from the document and query word lists;

c) one of two types of suffix cut-off procedures is used to reduce the word lists to word stem lists:

   i) the suffix "s" method removes only final "s" endings;

   ii) the regular word stem process cuts off all normal suffixes to produce word stems;

d) weights are assigned to the word stems based on the frequency of occurrence of the stems in the document abstracts or query formulations;

e) the weighted word stem vectors representing documents and queries respectively are compared, and a correlation coefficient is computed for each query-document comparison reflecting the similarity between the corresponding vectors;

f) the document citations are presented to the user in decreasing order of the correlation coefficients.

In order to compare the SMART retrieval procedures with the MEDLARS searches, it is necessary to keep the number of retrieved documents constant for both systems. This can be achieved in several different ways.

The most obvious possibility consists in picking for each SMART search a cut-off identical to the number of documents retrieved by MEDLARS for the same query. This procedure optimizes the results for MEDLARS, since the Boolean searches used at the National Library of Medicine do not present a specification of the number of items desired as output. At the same time, the SMART system is forced to operate with a MEDLARS cut-off for which no particular justification exists, even though the SMART ranking feature could normally provide a more rational output.

The recall-precision results averaged over the 29 queries are shown in Table 3. It may be seen from the Table that the average recall is about forty percent lower for SMART than for MEDLARS, whereas the precision loss is between thirty and forty percent. Obviously, the simple word, or word stem extraction used with a cut-off equivalent to that obtained in a Boolean search is not as effective as the MEDLARS indexing.

When the ranked output provided by SMART is used, the situation improves drastically. Specifically, instead of retrieving for each query the exact number of documents obtained by the Boolean search output used with MEDLARS, it is possible to retrieve all those items whose correlation with the queries exceeds some predetermined value. The results obtained with this procedure while still maintaining the simple word extraction method, are shown in Table 4. The correlation cut-off of 0.2403 for suffix "s", and 0.2501 for word stem is chosen in such a way that over the 29 queries the total number of documents retrieved by SMART is still equal to the 127 items retrieved overall by MEDLARS. However, the number of items

Analysis Method	Cutoff Determining Number Retrieved	Recall	Percent Difference from MEDLARS	Precision	Percent Difference from MEDLARS
MEDLARS (controlled terms)	Boolean search	.3117		.6110	
SMART word form (suffix "s")	same as MEDLARS	.1814	-42%	.3867	-37%
SMART word stem	same as MEDLARS	.1814	-42%	.4141	-32%
Maximum possible values		(.4471)		(1.0000)	

Natural Language Word Extraction Versus MEDLARS
Controlled Terms

(Averages for 450 documents, 29 queries)

Table 3

retrieved for a particular query is no longer the same for SMART and MEDLARS; rather, SMART now retrieves all those documents which exhibit a high query-document similarity.

The SMART output of Table 4 shows a deficiency of only 16 percent in recall and 19 percent in precision when the word stem extraction with the ranked feature is compared with the MEDLARS indexing. The ranked output has therefore reduced the deficiency by half, thus indicating that the utilization of the Boolean search technique is not optimal in normal document retrieval systems. Instead, a vector matching process, providing a numeric similarity coefficient between queries and documents can be utilized to obtain a more effective output product.

Since the SMART output is still below the MEDLARS standard, it becomes necessary to introduce additional techniques to improve the effectiveness of the automatic procedures. Two main possibilities suggest themselves: the first consists in introducing some form of dictionary — automatically, or manually derived — for word control purposes; the other utilizes the feedback search techniques provided by the SMART system. These methods are further described in the next two sections.

4. Use of Stored Dictionaries Versus Manual Indexing

The SMART output of Tables 3 and 4 was obtained by using word stems extracted from document abstracts and query formulations without any word control beyond the exclusion of obvious function words. Previous evaluations of the SMART procedures have shown that some word control is useful in improving the effectiveness of the automatic text analysis

Analysis Method	Cutoff Determining Number Retrieved	Recall	Percent Difference from MEDLARS	Precision	Percent Difference from MEDLARS
MEDLARS (controlled terms)	Boolean search	.3117		.6110	
SMART word form (suffix "s")	Correlation .2403	.2613	-16%	.4960	-19%
SMART word stem	Correlation .2501	.2622	-16%	.4901	-19%
Maximum values		(.4471)		(1.0000)	

Natural Language Word Extraction with SMART
Ranked Output Versus MEDLARS

(SMART correlation set to retrieve same
total number of documents (127) as MEDLARS)

Table 4

procedures. [12] Accordingly, various types of dictionaries, and word control lists are normally used with the SMART system.

Two different dictionaries providing language control are tested for the present SMART - MEDLARS evaluation:

a) the first one is an automatically constructed <u>word discriminator list</u> which excludes any term determined to be a common term, or a nondiscriminator;

b) the second is a regular <u>thesaurus</u>, generated in part by manual methods, which groups the thesaurus entries into affinity groups, to provide synonym recognition.

Consider first the construction of the automatic discriminator dictionary. A blueprint of the generation process is outlined in Table 5. A word list is first constructed of all the words included in a sample document (or abstract) collection in the topic area under discussion (medicine). Following removal of final "s" endings, this list included 13,471 entries in the present instance. Terms of frequency 1 are deleted next, as well as high-frequency terms occurring in at least 25 percent of the document abstracts. This leaves about 6,200 dictionary entries. Finally an automatic procedure is used to recognize additional nondiscriminators — that is, words which are not useful in providing discrimination among the documents of the collection. [13,14]

A <u>nondiscriminator</u> is defined as a term which increases the inter-document similarity — that is, the average similarity coefficient between the documents — when it is incorporated with the normal document vectors. Contrariwise, a discriminator is one which reduces the inter-document similarity by rendering the documents less similar to each other.

Procedure	Number of Entries Removed	Number of Entries Remaining
Formation of word list from texts of document abstracts following removal of "s" endings		13,471
Deletion of words of frequency one	7,245	6,226
Deletion of terms occurring in 25 percent or more of the abstracts	30	6,196
Deletion of terms automatically determined to be nondiscriminators	255	5,941

Automatic Discriminator Dictionary Characteristics

Table 5

The automatic recognition procedure identifies 255 additional terms as nondiscriminators, leaving a total of 5,941 entries in the dictionary used for the current experiments.

Some of the terms, automatically identified as nondiscriminators are listed in Table 6, together with certain terms found to be discriminators. The terms listed in Table 6 are arranged in decreasing "nondiscrimination order"; that is, those terms which cause the greatest increase in inter-document similarity, and which therefore provide the least amount of discrimination between the documents are shown at the top of the list.

The effectiveness of the automatic word discrimination dictionary can be assessed by considering the recall and precision output of Table 7. A query-document correlation of 0.2109 is used for the SMART runs, set to retrieve exactly 127 documents over all 29 queries. It may be seen that the automatic dictionary provides a ten percent improvement in recall and a twenty percent precision advantage over the standard word stem extraction method. Furthermore, the order of magnitude of the SMART output is now approximately the same as that obtained with the controlled MEDLARS indexing (minus 8 percent in recall, and minus 4 percent in precision).

The results obtained with the automatic discriminator dictionary thus confirm the previous evaluation output, that fully automatic text processing methods can be used to obtain retrieval output of an effectiveness substantially equivalent to that provided by conventional, manual indexing.

	Dictionary Entry		Number of Document Occurrences	Total Number of Occurrences	Average Occurrence per Abstract
1.	cell		208	785	3.77
2.	case		253	521	2.06
3.	have		256	332	1.30
4.	effect		207	346	1.67
5.	may		222	327	1.47
6.	normal		203	369	1.82
7.	treatment	Non	181	300	1.66
8.	has	Discriminators	225	305	1.36
9.	result		227	285	1.26
10.	other		234	304	1.30
	⋮		⋮	⋮	⋮
251.	pattern		52	72	1.38
252.	order		34	36	1.06
253.	do		37	43	1.16
254.	involved		39	41	1.05
255.	among		42	47	1.12
1.	DNA		43	201	4.67
2.	antigen		34	128	3.76
3.	nickel	Discrimin-	21	78	3.71
4.	HGH	ators	11	46	4.18
5.	amyloidosi(s)		18	73	4.06
6.	tumor		24	66	2.75
7.	hepatiti(s)		13	49	3.77
8.	oxygen		31	104	3.35

Typical Discriminating and Nondiscriminating

Dictionary Entries

Table 6

The discriminator dictionary process outlined in Tables 5 and 6 does not, however, include any term grouping method, and does not provide therefore any help in recognizing synonyms and other closely related terms.

A thesaurus may, however, be used as part of the analysis methodology which operates in such a way that individual terms ocurring in document abstracts and search requests are replaced by the corresponding thesaurus class numbers. Thus, a document containing the term "production" can be matched with a query containing "manufacture", assuming that both terms are entered in the same thesaurus class.

Methods exist for automatically grouping the terms into thesaurus classes [15,16]. For the present test, the thesaurus used with the extended MEDLARS collection was, however, constructed partly by hand by a trained expert.*

The following steps were used to construct the extended MEDLARS thesaurus:

a) a list of all words contained in the MEDLARS document abstracts was automatically produced, together with a concordance and word frequency list;

---

*Thesauruses have been constructed for use with the SMART system in many subject areas, and for several different natural languages. The manual process used for this purpose has been criticized because the thesaurus construction "is largely nonprocedural and nonreproducible" [17]. In fact, the same "thesaurus construction principles" are used for all manual SMART thesauruses, and the procedure is largely reproducible [7, p.28-29].

Analysis Method	Cutoff	Recall	Percent Difference from MEDLARS	Percent Difference from SMART Stem	Precision	Percent Difference from MEDLARS	Percent Difference from SMART Stem
MEDLARS (Controlled terms)	Boolean Search	.3117			.6110		
SMART	Correlation						
word stem	.2501	.2622	-16%		.4901	-19%	
automatic discriminator list	.2109	.2872	- 8%	+10%	.5879	- 4%	+20%
thesaurus	.3720	.3232	+ 4%	+23%	.6106	0%	+25%

SMART Automatic Discriminator Dictionary and Thesaurus
Versus MEDLARS Controlled Terms
(SMART correlation set to retrieve a total of 127 documents)

Table 7

(b) the following terms were then eliminated:

i) 220 high frequency, common function words;

ii) all words of frequency one;

iii) all functors (prepositions, pronouns, auxiliary verbs, conjunctions, articles, and so on);

iv) words of general meaning (such as, for example, "appear", "associated", "comparable", "necessary", "take", and so on);

v) single letters;

vi) names of persons, unless also the name of a disease;

Stedman's Medical Dictionary and Webster's Seventh New Collegiate Dictionary were consulted in the process;

(c) words with a unique meaning in the corpus were assigned concept numbers (that is, thesaurus class numbers) in decreasing frequency order; the alphabetical word list was consulted in the process to detect alternate spellings (for example, fetus, foetus), as well as different word forms of the same stem; such terms were assigned the same concept numbers, as were synonyms of already classified terms of comparable frequency of occurrence; using the punched card form of the thesaurus, new words freshly classified were interfiled, and an up-to date listing of the existing thesaurus could be made at any time;

(d) for high-frequency terms, the names of body parts, their diseases, and the corresponding operations were given separate thesaurus class numbers (for example, kidney, nephritis, nephrectomy); for low frequency terms, these various concepts were grouped into the same class, using the previously mentioned thesaurus construction principles (for example, leucocyte, leucocytosis; word groups (phrases) which co-occurred exclusively were given the same thesaurus class number (e.g. sella turcica).

Following the completion of the first pass in the thesaurus construction, each thesaurus category was reviewed, and some classes were broken up into two or more classes if the combined frequency of the included words was too high; contrariwise, certain low frequency classes were combined into a single class. The final thesaurus includes 3766 entries broken down into 1737 concept classes. Some typical classes are reproduced in Table 8.

The effectiveness of the word normalization procedures achievable by using the SMART thesaurus is illustrated on the last line of Table 7. It may be seen that the thesaurus offers an improvement in average recall and precision over the standard word stem process of about 25 percent. The retrieval effectiveness is again about the same as that achievable with the MEDLARS indexing, with a slight advantage for SMART.

The conclusion is that with a minimal amount of language normalization, provided, for example by a word stem discrimination list or a thesaurus, the automatic SMART language processing is fully equivalent to the conventional MEDLARS indexing administered by trained indexers.

5. Use of SMART Feedback Searches

It is well known that improvements in retrieval output are obtainable by generating better query formulations than those originally submitted by the system user. This can be achieved before any search operation is actually carried out by displaying for the user's attention excerpts of available dictionary, or word control list entries, thereby suggesting to the requestors additional possibilities for formulating the search queries.

Total number of thesaurus entries	3766
Total number of thesaurus classes	1737
Average number of entries per class	2.2

Class Number	Entries	Class Number	Entries
134	Fetal Fetus Fetuses Foetal Foetus	731	Plasma-Calcium CA15 CA Calcium
144	Fears Anxiety Anxieties Anxious	960	Iron FE Fe-Din Ferric
300	Pitressin Vasopressin	1430	Reimplantation Implantation Implanted Implants
415	Anaemia Anemia	1531	Dysarthria Stuttering

Thesaurus Characteristics and Sample Classes

Table 8

Alternatively, a search may be performed using the initial queries, and selected output data — for example, titles or abstracts of previously retrieved documents — may be used to reformulate the queries for use in subsequent search operations.

Feedback operations based on quality of output are not directly built into the MEDLARS procedures; the Boolean search output consisting of a batch of unranked documents does not, in any case, make it simple to furnish feedback data. However, three different search formulations are constructed by the MEDLARS searchers as a matter of course, and the "best" output — that is, the one supplying a manageable number of output documents — is then submitted to the user. This type of operation is then equivalent to feedback based on output quantity (rather than quality).

The SMART system, on the other hand, utilizes the ranked document output for its relevance feedback operations. [18,19] Specifically, two or three of the highest ranked documents retrieved in an earlier search operation are submitted to the user for a rough assessment of usefulness. The queries are then automatically adjusted by addition of terms from the documents termed relevant, and simultaneous deletion of terms from the nonrelevant items. In a previous evaluation of the relevance feedback procedure it was found that "one feedback stage improves the output by 10 to 22 percent, while two stages produce an advance of 13 to 36 percent in recall and precision." [20]

The effectiveness of the relevance feedback output is illustrated for the extended MEDLARS collection in Tables 9 and 10. Table 9 contains the data for the word forms and word stem extraction methods,

Analysis Method	Cutoff Determining Number Retrieved	Recall	Percent Difference from MEDLARS	Precision	Percent Difference from MEDLARS
MEDLARS (controlled terms)	Boolean Search	.3117		.6110	
SMART word form	Correlation				
0 - initial search	.2403	.2613	-16%	.4960	-19%
1 - iteration feedback	.3434	.3328	+ 7%	.6398	+ 5%
2 - iteration feedback	.3405	.3525	+13%	.6740	+13%
SMART word stem	Correlation				
0 - initial search	.2501	.2622	-16%	.4901	-19%
1 - iteration feedback	.3380	.3235	+ 4%	.6385	+ 5%
2 - iteration feedback	.3260	.3433	+10%	.6892	+13%

Natural Language Word Extraction With Ranked

Output and Feedback Searches Versus MEDLARS

(SMART correlation set to retrieve same total number of documents (127) as MEDLARS)

Table 9

while Table 10 illustrates the operations of the automatic discriminator dictionary and the thesaurus. In each case the initial run performance is given together with one and two stages of relevance feedback.

It may be seen from Table 9 that the 15 to 20 percent deficiency noted earlier for the SMART word stem extraction process is turned into an advantage over MEDLARS of 4 to 7 percent after one feedback operation (two searches in all), and of 10 to 13 percent for two feedback iterations (three searches). The same data for the SMART thesaurus indicate that the small improvements over MEDLARS obtainable in one search operation turn into a large advantage of 18 to 25 percent after one feedback operation, and of almost 30 percent after two feedback operations. It is interesting to note in this connection that the recall performance of the SMART thesaurus after the second feedback operation (0.4029) lies within 9.88 percent of the maximum recall achievable in the present test (0.4471), while the precision (0.7438) is within 25 percent of the ideal precision (1.0000).

Clearly, the feedback procedures produce the same large increases in retrieval effectiveness predicted by previous tests of the SMART operations. Furthermore, when used with simple term extraction from document abstracts or text, the feedback operations turn a deficit caused by the lack of sophistication in the language analysis into a clear advantage over the conventional, manual indexing methods.

6. Conclusions

The following main conclusions appear in order as a result of the extended SMART- MEDLARS comparison:

a) The strong points of the automatic retrieval system appear to be the vector matching techniques which furnish

Analysis Method	Cutoff Determining Number Retrieved	Recall	Percent Difference from MEDLARS	Precision	Percent Difference from MEDLARS
MEDLARS (Controlled terms)	Boolean Search	.3117		.6110	
SMART automatic discriminator dictionary	Correlation				
0 - initial search	.2109	.2872	- 8%	.5879	- 4%
1 - iteration feedback	.2934	.3677	+15%	.7178	+17%
2 - iteration feedback	.3250	.3801	+22%	.7427	+21%
SMART thesaurus	Correlation				
0 - initial search	.3720	.3232	+ 4%	.6106	0%
1 - iteration feedback	.4880	.3915	+25%	.7427	+18%
2 - iteration feedback	.5125	.4029	+29%	.7438	+22%

SMART Automatic Dictionary and Thesaurus with Feedback

Versus MEDLARS Controlled Terms

(SMART correlation set to retrieve a total of 127 documents)

Table 10

ranked document output, the automatic construction methods for word control lists, and the feedback operations;

b) the simple word stem extraction process using document abstracts and query texts is only 15 to 20 percent less effective than the best available manual indexing based on controlled vocabularies;

c) automatic language normalization procedures can be used to build dictionaries, and thesauruses, whose operations produce output results equivalent to the standard manual indexing;

d) the SMART relevance feedback procedures produce large improvements in retrieval effectiveness;

e) the Boolean search techniques which appear to have been developed for use with earlier punched-card technologies are clearly inferior to vector matching techniques producing ranked output in decreasing query-document similarity order;

f) no technical justification appears to exist for maintaining controlled manual indexing in operational retrieval environments.

. Acknowledgement

Dr. Joseph Leiter and Mr. Constantine J. Gillespie of the National Library of Medicine, Dr. M. E. Lesk of Bell Laboratories, and Professor F.W. Lancaster of the University of Illinois helped with the test design. Mr. Robert G. Crawford and Mrs. Barbara Galaska of the SMART staff designed respectively the automatic discriminator dictionary and the SMART thesaurus. The writer gratefully acknowledges the assistance of these individuals.

References

[1] G. Salton, Automatic Text Analysis, Science, Vol. 168, No. 3929, 17 April 1970, p. 335 - 343.

[2] H. Coblans, Words and Documents, Aslib Proceedings, Vol. 23, No. 7, July 1971, p. 337 - 350.

[3] J. R. Sharp, Where do we go from here, Aslib Proceedings, Vol. 23, No. 1, January 1971, p. 33 - 46.

[4] J. H. Schneider, Selective Discrimination and Indexing of Scientific Information, Science, Vol. 173, 23 July 1971, p. 300 - 308.

[5] D. F. Hersey, W. R. Foster, E. W. Stalder, and W. T. Carlson, Free Text Word Retrieval and Scientist Indexing: Performance, Profiles and Costs, to be published Journal of Documentation.

[6] F. W. Lancaster, Evaluation of the MEDLARS Demand Search Service, National Library of Medicine, Bethesda, Md., January 1968.

[7] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Book Co., New York, 1968.

[8] G. Salton, The SMART Retrieval System — Experiments in Automatic Text Processing, Prentice Hall Inc., Englewood Cliffs, N.J., 1971.

[9] G. Salton, A Comparison between Manual and Automatic Indexing Methods, American Documentation, Vol. 20, No. 1, January 1969.

[10] R. M. Hayes, Review of "A Comparison between automatic and manual indexing systems", Computing Reviews, Vol. 10, No. 6, June 1969, p. 274.

[11] M. E. Lesk and G. Salton, Relevance Assessments and Retrieval System Evaluation, Information Storage and Retrieval, Vol. 4, 1969, p. 343 - 359.

[12] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1968, p. 8 - 36.

[13] K. Bonwit and J. Aste-Tonsman, Negative Dictionaries, Scientific Report No. ISR-18 to the National Science Foundation and the National Library of Medicine, Section 6, Cornell University, Department of Computer Science, October 1970.

[14] G. Salton, Experiments in Automatic Thesaurus Construction for Information Retrieval, IFIP Congress-71, Ljubljana, August 1971, to be published by North Holland Publishing Company.

[15] K. Sparck Jones and D.M. Jackson, Current Approaches to Classification and Clump Findings, Computer Journal, Vol 10, No. 1, 1967, p. 29 - 37.

[16] C. C. Gotlieb and S. Kumar, Semantic Clustering of Index Terms, Journal of the ACM, Vol. 15, No. 4, October 1968, p. 493 - 513.

[17] C. McAllister, review of Automatic Processing of Foreign Language Documents, Computing Reviews, Vol. 12, No. 5, May 1971, p. 224.

[18] G. Salton, Search and Retrieval Experiments in Real-Time Information Retrieval, IFIP Congress 68, No. Holland Publishing Co. Amsterdam 1969, p. 211 - 220.

[19] M. E. Lesk and G. Salton, Interactive Search and Retrieval Methods using Automatic Information Displays, Proc. AFIPS Spring Joint Computer Conference, AFIPS Press, Montvale, N.Y., 1969, p. 435 - 446.

[20] G. Salton, The Performance of Interactive Information Retrieval, Information Processing Letters, Vol. 1, No. 2, July 1971, p. 35 - 41.