Interactive Search and Retrieval Methods Using Automatic Information Displays

M. E. Lesk and G. Salton

Abstract

Presently available information storage and retrieval systems do not produce retrieval results which will satisfy the information needs of all potential users. Interactive search methods using console displays and conversational computing methods promise to furnish retrieval results which are far superior to those achievable by conventional procedures. In the present study, various interactive search strategies are used in conjunction with the automatic SMART document retrieval system, and an attempt is made to evaluate the effectiveness of each method as part of a retrieval system. In particular, the usefulness of each method in retrieving wanted and rejecting unwanted items is discussed, as well as the cost of the user-system interaction in terms of additional user effort and computer time. It is found that for all but the most experienced users, automatic methods requiring little user effort are preferable to more sophisticated procedures which may produce somewhat better retrieval results at somewhat higher cost.

1. Introduction

Throughout the world, the design and operation of large-scale information systems has become of concern to an ever-increasing segment of the scientific and professional population. Furthermore, as the amount and complexity of the available information has continued to grow, the use

of mechanized or partly mechanized procedures for various information storage and retrieval tasks has also become more widespread. While a number of retrieval systems are already in operation in which the search operations needed to compare the incoming information requests with the stored items are performed automatically, no systematic study has ever been made of the use of man-machine interaction as a part of a mechanized text analysis and information processing system. Specifically, the recent development of high capacity random-access storage mechanisms and conversational input-output consoles should permit a rapid interchange of information between users and system. Such an interchange can then be used to produce improved search formulations, resulting in a more effective retrieval service.

The present report describes and evaluates the performance of a variety of such interactive search and retrieval procedures in which information supplied by the user population is taken into account in an attempt to achieve improved system responses. Several basic approaches to user-system interaction are possible. On the one hand, an attempt can be made to construct refined query formulations, using dictionary displays and similar methods, before any file search is actually attempted. On the other hand, an original query can be processed when it is first received, and a query reformulation attempted after the results of an initial search are actually available. These two procedures, termed pre-search and post-search, respectively, can in turn be executed in several different ways: either the system assumes most of the burden of the query reformulation through an automatic query alteration process, or, the users themselves can rephrase their queries using the available automatic displays. In the latter case,

the skill of the user population becomes a more important factor. The stored data most important in the pre-search methods might include synonym dictionaries and thesauruses, word frequency statistics, and lists of significant words; the post-search information, on the other hand, consists of the titles, abstracts, or texts of documents retrieved by a previous search process.

The investigation of the various interactive search and retrieval methods is carried out with the help of the automatic SMART document retrieval system [1,2]. The SMART system is a large computer-based retrieval system capable of performing a variety of different text analysis, search, and retrieval operations. Completely automatic text analysis and information searches are made using several different analysis methods and search strategies. Among the main text analysis procedures are synonym recognition, word disambiguation, phrase recognition, statistical term association, and hierarchical text expansion methods.

The effectiveness of the various analysis and search methods may be evaluated by using for this purpose the familiar recall and precision measures, representing respectively the proportion of relevant material actually retrieved, and the proportion of retrieved material actually relevant. Ideally, all relevant items should be retrieved for the user, while at the same time, all nonrelevant items should be rejected, thus leading to a system where both recall and precision are equal to 1. The performance effectiveness of an operating system can then be estimated by averaging recall and precision figures over many searches and comparing the results with the ideal situation where recall and precision are equal

to 1. The SMART system automatically generates for each search a set of recall-precision graphs first introduced by Cleverdon [3], and also includes procedures for performing computations of the statistical significance of the results. Evaluation data for a wide variety of automatic text processing, search and retrieval methods have previously been published [4].

In addition to the recall-precision data which reflect the capability of the system to deliver to the user the information he requests, it is also important in an interactive computing environment to take into account the amount of effort required from the user to obtain satisfactory results. Thus, the standard performance of fully-automatic search and retrieval operations must be compared against the improvements obtainable through interactive procedures at additional cost in user effort and computer time.

In the remainder of this study, the effectiveness of various types of interactive search methods is examined, including both pre-search and post-search methods, and semi- or fully-automatic query reformulation procedures. The results are compared using, in each case, the evaluation methods incorporated into the SMART system. Construction principles are then derived for future information services designed to use man-machine interaction during the search process.

2. Fully-Automatic Retrieval

In the SMART system, various fully-automatic language analysis procedures are used to normalize the text of incoming search requests and of stored documents. The normalized, reduced forms of the information items, consisting generally of weighted "concept" numbers, are then compared, and

the document representations which are most similar to the request representations are extracted from the file as answers to the queries. The language normalization procedures incorporated into the SMART system range from simple word stem matching methods to more sophisticated processes using stored synonym dictionaries and hierarchies, as well as statistical and syntactic analysis methods [1,2].

Three of the simplest language analysis methods, known, respectively, as word form, word stem, and thesaurus processes may be described as follows:

a) in the word form, or suffix 's', process, no word normalization in the proper sense is used at all, and the original words with only the final 's' removed (to confound, for example, "book" and "books") are compared directly;

b) in the word stem method, the original text words are reduced to word stems by a suffix cut-off process to confound words like "analyzer", "analysis", "analyzed", and so on, before the comparison between queries and documents;

c) in the thesaurus process, each word stem is looked up in a synonym dictionary, or thesaurus, where it is replaced by one or more so-called concept numbers, representing synonym classes; the concepts extracted from the thesaurus are then matched instead of the original word forms or word stems.

In all analysis methods, the terms are normally weighted, using word frequency and other criteria, before a comparison is made between stored documents and search requests.

An excerpt from a typical, manually constructed thesaurus is shown in Table 1. Three of the synonym classes defined by the thesaurus mapping are shown in the right-hand side of Table 1. Concept class 346, for example, contains words specifying objects which fly; category 345 lists words associated with weather. If a request were made, asking

"do planes fly when the weather is bad?"

the system would retrieve a document stating

"proper meteorological conditions are necessary for the successful piloting of aircraft",

since both document and query would be assigned the concepts 345 and 346.

The handling of ambiguous words in the thesaurus is exemplified by the entry for "wind", which could be either the noun, referring to weather, or the verb, indicating a method of constructing loops or coils. The table shows that "wind" is in two categories, 345 and 233. 345, containing also "weather" and "atmosphere", represents the noun, and 233, which contains such words as "winding" "wire-wound", and "solenoid", represents the verbal meaning. Whenever "wind" appears, both 345 and 233 will be entered into the concept vector. Because the word is considered ambiguous, the weight will be divided between these two categories; each will receive half of the weight assigned to "wind".

It should be noted that the thesaurus entries may consist of word stems, so that "meteorolog" suffices to look up "meteorology" and "meteorological". If desired, however, suffixed forms of a word may be entered in

Alphabetic Order			Numeric Order	
Word	Concept Code	Syntax Code	Concept Code	Word
Wide	438	001 043 040	344	obstacle target
Will	32032	009 070 043 044 049	345	atmosphere meteorolog weather wind
Wind	345 233	070 043		
Winding	233	070 136 137		
Wipe	403	043 070		
Wire	232 105	070 043	346	aircraft airplane bomber craft helicopter missile plane
Wire-wound	233	001		

Thesaurus Excerpt
Table 1

Query Alteration Process	Explanation
Pre-Search	
1. Repeated Concepts	User chooses query terms to be repeated for emphasis
2. Thesaurus Display	User chooses terms obtained from thesaurus display to update query (with or without time restrictions)
3. Word Frequency	User looks at display of word frequency information before updating query
4. Source Document	User looks at display of source document before updating
Post-Search	
5. Title Display	User looks at titles of first five retrieved documents before updating
6. Abstract Display	User looks at abstracts of first five retrieved documents
7. Relevance Feedback	Query is updated automatically using relevance judgments supplied by user following an initial search
Combined Methods	
8. Abstract plus Thesaurus	User looks at pre- and post-search information

Typical Query Updating Methods

Table 2

the thesaurus; this has been done with "winding", since if only "wind" were in the dictionary, "winding" would also be treated as ambiguous, but the presence of "winding" in the thesaurus makes it possible to identify "winding" in the text with category 233 only.

The high concept number identifies "will" as a so-called common word, not to be used for content identification. The syntax codes shown with the thesaurus entries in Table 1 are not used in the simple automatic thesaurus process.

Since the fully-automatic thesaurus process based on concept number matching is often an effective analysis tool, more sophisticated language normalization methods may not normally be required in an operational retreival system.

3. User Interaction Through Pre-Search Methods

One of the main hopes in obtaining a retrieval performance which goes beyond that presently reached under normal operating conditions is to include the customer in the search process. In particular, fewer errors are likely to be made if the information obtained from the users is not restricted to the search request proper, but is supplemented by a variety of special user indications, or by evaluation data about the acceptability of items previously retrieved by the system in answer to the search requests. User-system interaction is now current for many computer application, often implemented by special input-output console devices, with the help of operating systems which enable the system to render more or less simultaneous service to a large class of users.

In an information retrieval environment, user interaction may take the form of simple dictionary display routines which can be used to present to the user selected dictionary excerpts as an aid in formulating the original search requests, or in reformulating queries which were originally inadequate [5,6]. Alternatively, more sophisticated methods may be used in which the reformulation of the search requests is automatically performed based on feedback information obtained from the user population [7,8].

The conceptually simpler methods are the pre-search procedures which are based on term and dictionary displays of previously stored information. In each case, a user would look at the displayed information and, based on the available data, would decide before any file search is actually attempted how his query could best be reformulated in order properly to reflect his information needs. The following types of pre-search information could be displayed for this purpose:

a) lists of terms included in the user's original search formulation together with word frequency information giving the frequency of use of each word in one or more of the stored document collections;

b) thesaurus excerpts corresponding to the terms included in the user's search formulation, and consisting, for each of the originally available terms, of a complete thesaurus class, including synonyms and other terms related to the original;

c) titles and abstracts of source documents, that is, of documents originally known to the user as relevant to his search query (the intent of the user is then to retrieve new documents similar to the source items).

The principal differences between fully-automatic retrieval and retrieval using pre-search interaction are summarized in the flowcharts of Figs. 1(a) and 1(b). The pre-search requires the generation of a computer display followed by a manual choice of terms on the part of the user during the query formulation process.

The display of word frequency information is designed to inform the user of the characteristics of the vocabulary which may be used to express his information needs. Thus, if a user notices that many of the terms included in his search request are general terms with a very high frequency of occurrence in the stored document collection (for example, terms such as "computer" and "automatic" in a collection on computer science), he may decide that it is wise to delete these terms from his query so as to prevent the generation of high query-document correlations for many nonrelevant documents. On the other hand, the user may decide to emphasize many highly specific, low-frequency terms by repeating them in the query formulation.

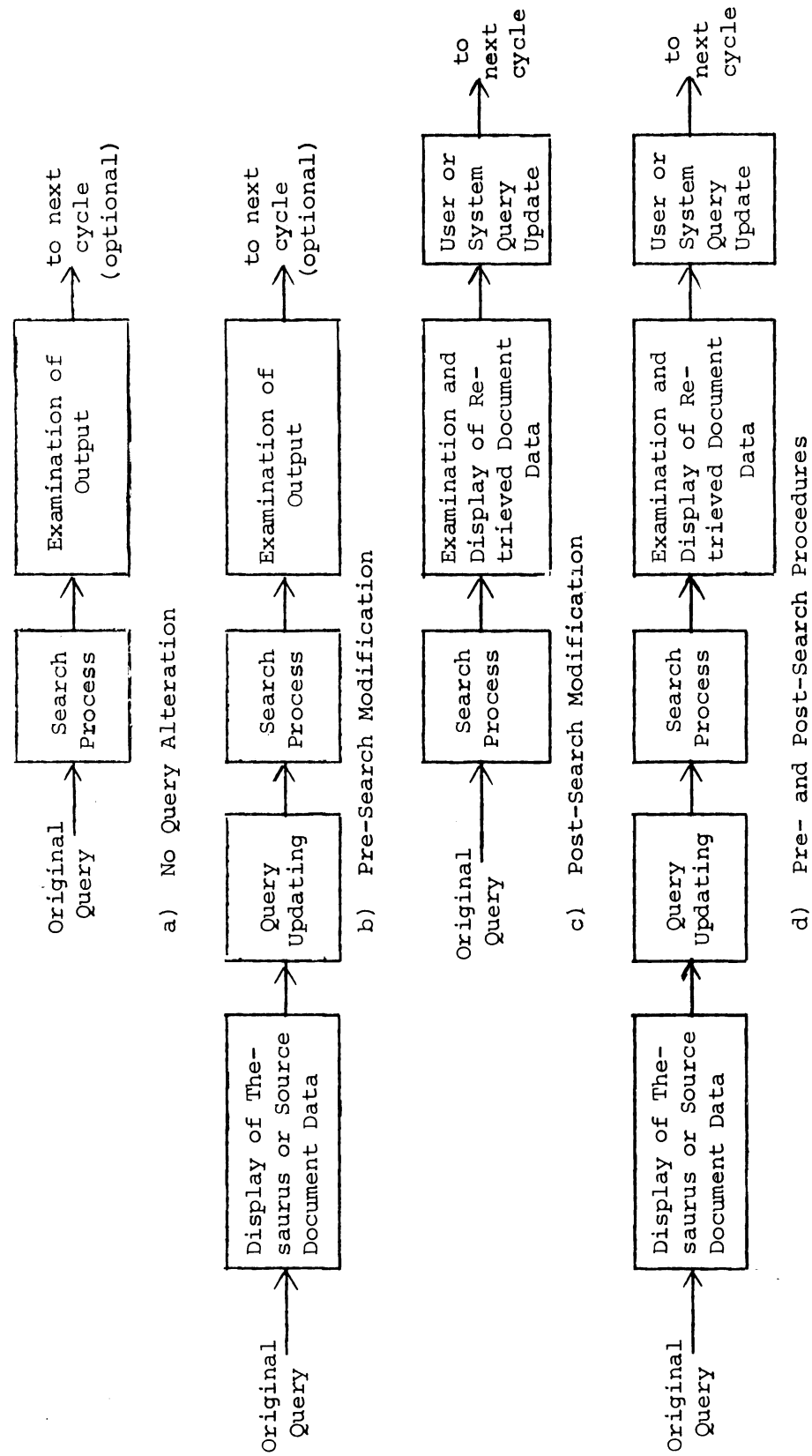
A thesaurus display can be used for manual query updating by requesting a printout of the complete thesaurus classes corresponding to each term included in the original query. Consider, as an example, a query dealing with the "contraction of satellite orbits", and assume that the user signifies that he is interested in the "satellite" class. The computer might then type out terms such as

> Discoverer, Sputnik, Vanguard, Cosmos, Moon, rocket, trajectory, countdown, drag, telemetry, etc.

After studying the display, the user might decide that his original query formulation had been insufficiently specific, and the query might then be altered by addition of the terms "Discoverer, Sputnik, Vanguard, Cosmos, drag, and telemetry". The other displayed terms would, however, be rejected as not being germane to the search topic. A second expansion might begin by typing in the term "drag", and then considering the new display of terms related to "drag".

Thesaurus displays are also occasionally useful for the removal from the query formulation of incorrectly used and ambiguous terms. For example, a user interested in information retrieval who identifies his search topic as "IR" might discover that the thesaurus display produces a list of synonyms in the area of "infra-red spectroscopy". As a result, the term "IR" would, of course, be removed from the search formulation.

The use of thesaurus displays for manual query updating provides an opportunity for a selective choice of synonym and related terms. That is, the user can choose some terms to be added to the original query, and others to replace already existing ones in an attempt to improve search precision. On the other hand, the automatic thesaurus process operates less selectively and provides synonym recognition by the standard process of automatically replacing the word stems originally included in the search requests and documents by the corresponding concept class numbers extracted from the thesaurus. The automatic thesaurus process is thus designed to normalize query and document statements by generalizing the respective formulations rather than by making them more specific. Such a process may be expected to improve recall, since more relevant documents could now



Iterative Search Procedures

Fig. 1

match the query statements and could thus be retrieved in answer to the respective search requests.

Obviously, the manual query updating methods using thesaurus displays places a considerable burden on the user, since he is forced to consider a large number of alternative possibilities before eventually making a move. Moreover, the choice must be made before a search has actually been performed, at a time when he cannot know as yet how well the machine will perform with any potential query formulation.

A comparison of the effectiveness of manual and automatic thesaurus procedures is contained in section 5 together with the other evaluation output.

4. User Interaction Through Post-Search Methods

The post-search methods are those applicable after an initial search has first been performed. In such a case, one or more documents will already be available, including in particular those items which were initially judged to be most similar to the search requests. These items can now be used in a manner analogous to that previously utilized for the thesaurus displays. Specifically, the titles, or abstracts, of the first few retrieved documents can be examined, and document terms which appear to reflect the wanted subject area can be added to the query statement, while ambiguous and unwanted terms can be removed.

Consider, for example, the previously cited query dealing with the "contraction of satellite orbits", and assume that the first two retrieved items are entitled "Discoverer satellite and South Pacific splash down",

and "The moon and the tides". A user could now proceed to add "Discoverer satellite" to the original query, but could avoid the addition of "South Pacific".

The document feedback expansion may be even more difficult to carry out than the dictionary display procedure, since the user is forced to make sophisticated decisions using relatively large text excerpts. Thus, whereas the dictionary display procedure can often be performed in less than a minute per query, approximately four minutes are required on the average for the use of five typical document abstracts. Furthermore, the document expansion process also entails a higher cost in machine time and storage space than the dictionary display, since document abstracts in natural language form constitute a much greater bulk than dictionary excerpts. In addition, an initial retrieval run must first be made before document feedback can be used. On the other hand, a stored dictionary need, of course, not be available for the document feedback method.

Another post-search method is designed particularly for those users who do not wish to assume the burden of query reformulation themselves. For such users, an automatic relevance feedback method is available which requires only a minimum of interaction with the user, since most of the burden is placed on internally stored routines [7,8,9,10]. Specifically, an initial search is first performed for each request received, and a small amount of output consisting of some of the highest scoring documents, is presented to the user. Some of the retrieved output is then examined by the user who identifies each document as being either relevant (R) or not relevant (N) to his purpose. These relevance judgments are

later returned to the system, and used automatically to adjust the initial search request in such a way that query terms, or concepts, present in the relevant documents are promoted (by increasing their weight), whereas terms occurring in the documents designated as nonrelevant are similarly demoted.

If the terms from the relevant items are added to the search requests, while terms from nonrelevant items are subtracted, the first query updating operation can be represented by the equation:

$$q_1 = q_o + \sum_i \underline{r}_i - \sum_i \underline{s}_i \quad ,$$

where q_o is the original query formulation, q_1 is the updated query, $\underline{r}_i$ is the set of terms identifying the i^{th} document specified as relevant by the user, and $\underline{s}_i$ is the set of terms identifying the i^{th} nonrelevant document. This process produces an altered search request which may be expected to exhibit greater similarity with the relevant document subset, and greater dissimilarity with the nonrelevant set.

The altered request can next be submitted to the system, and a second search can be performed using the new request formulation. If the system performs as expected, additional relevant material may then be retrieved, or, in any case, the relevant items may produce a greater similarity with the altered request than with the original. The newly retrieved items can again be examined by the user, and new relevance assessments can be used to obtain a second reformulation of the request. This process can be continued over several iterations, until such time as the user is satisfied with the results obtained. Since the method makes very few demands

on the user, the automatic relevance feedback process may be expected to be preferred by users unfamiliar with the system operations. On the other hand, the process is not likely to be effective if the user is unable to identify for the system at least one document which is clearly relevant to his needs.

The post-search methods as well as the combined methods making use of pre- as well as post-search information are illustrated in the bottom half of Fig. 1. A summarization of all the query updating methods is given in Table 2.

5. Evaluation Results and Discussion

The experimental results included in this section are based on the manipulation of a collection of 200 abstracts of documents in aerodynamics, together with 42 search requests proposed by scientists active in aerodynamics. Complete relevance judgments, prepared by these same scientists, were available which identify for each query the set of relevant documents. The aerodynamics collection has previously been used for test purposes by the Aslib-Cranfield project [3] and by the SMART system [4].

The thesaurus used for both the manual and automatic query expansion operations contains 3230 word stems and 736 thesaurus classes. This thesaurus was constructed by SMART staff members using text concordances, word frequency lists, standard dictionaries and reference works, and word lists obtained earlier from the Cranfield project. An attempt was made to time the query expansion operations by restricting the use of the thesaurus display to either one minute, two minutes, or more than two minutes. While

the output of Table 3 shows that increasingly more terms can be added to the queries as more time becomes available for the updating operations, the differences in retrieval effectiveness are small, and the evaluation output shown represents the output obtained for a display time of two minutes.

The main results are presented first in terms of recall-precision graphs, and then in terms of cost and user effort.

A) Recall-Precision Results

The evaluation output is presented in Figs. 2 to 7 using the standard recall-precision graphs, averaged in each case over the 42 queries used with the collection of 200 documents. The curves are, as usual, monotonically decreasing, reflecting the fact that as more relevant items are retrieved (as recall goes up), more irrelevant items are also retrieved (causing the precision to go down). Increasingly more effective retrieval performance is reflected by recall-precision curves close to the upper right-hand corner of the graph where both recall and precision take on ideal values of 1. Next to the graphs, some of the numeric values are presented in terms of recall-precision tables, giving the average precision values at certain selected recall values.

Significance values, computed by a standard t-test, are also included in the output figures, representing in each case the probability that the performance values for two specified processing methods are in fact derived from the same distribution. Thus, if the computed probability value is high, the two methods are assumed to be statistically indistinguishable; on the other hand, if the probability value is low - say 0.05 or less - the

likelihood that the evaluation results could have been derived from the same data set is very small, and the differences in performance can then be assumed to be statistically significant.

The following principal conclusions can be drawn from the output of Figs. 2-7:

a) automatic thesaurus vs. pre-search using thesaurus display (Fig.2):

the automatic thesaurus expansion and the manual expansion using pre-search thesaurus display both produce an improvement in performance over the word stem matching process. Overall, the automatic thesaurus (which requires no user intervention) is superior. At high precision, however, the greater selectivity of the words chosen by the manual process produces better results. The superiority of the automatic thesaurus at medium and high recall is attributed to the previously mentioned difficulty of selecting appropriate terms from the thesaurus display.

b) automatic thesaurus vs. pre-search using source document display (Fig. 3):

the source document display produces a precision improvement of up to ten percent over and above the automatic thesaurus process; however, the table appearing with Fig. 3 shows that the improvement is not statistically significant. The relatively modest increase in performance may be due to the fact that the source documents and queries used in the experiment originated with the same authors, so that the source documents contain many of the terms already included in the query statements; also, some of the source documents appear only marginally relevant to the actual queries; both of the interactive pre-search methods turn out therefore to be not substantially

superior to the fully-automatic thesaurus method (except at high precision).

c) post-search procedures using displays of titles or abstracts of previously retrieved items (Figs. 4 and 5):

title and abstract post-search displays are superior to both of the pre-search displays, as shown in Figs. 4 and 5. Improvement with title display is limited to the high precision regions, since the titles are so short that words not in the query are rarely included. The query alterations due to title display are therefore limited to deletion of unnecessary concepts, improving mostly the precision. Abstract displays produce both precision and recall improvements, at the cost of greatly increased work on the user's part. The amount of text examined during an abstract display process is about 1000 words, from which five to ten may be selected for query expansion.

d) automatic thesaurus vs. post-search updating using abstract display and relevance feedback (Fig. 6):

both the manual post-search method with abstract display and the automatic relevance feedback process are superior to the standard word stem process; the abstract display is best in the very high precision ranges. The performance differences between the two post-search methods are not significant at high precision, although the improvements obtained with both methods over the standard word stem process are significant. The relevance feedback output included in Fig. 6 is obtained by retrieving, in each case, 5 documents at a time, asking the user to identify any relevant items, and adding the corresponding terms to the search request.

e) combined pre-search dictionary and post-search abstract display (Fig. 7):

Fig. 7 shows that a combination of abstracts and thesaurus displays offers an overall improvement of about twenty percent over the standard word stem process, and of ten to fifteen percent over the thesaurus process; in both cases, the improvement is statistically significant. When word frequency information is added to the display, a further improvement results for the word stem procedure, since the user can now ensure that all parts of the query are properly weighted. The output of Fig. 7 is approximately equivalent to the automatic relevance feedback process (Fig. 6); however, the combined pre- and post-search process requires much more user effort and experience than the relevance feedback method before it can operate successfully.

B) Overall Evaluation

The performance of the various interactive procedures is summarized in Table 4. The first column reflects computer demands; the second, user effort; and the last two reflect search effectiveness in terms of recall and precision improvements over and above the normal word stem matching method.

Since the post-search methods require two separate file searching operations - one prior to the interactive process and one following it - the computer demands are comparatively higher for post-search than for the other methods. Thus, when search time may be expected to be considerable - for example, for very large collection sizes - the pre-search procedures may become mandatory.

From the user's viewpoint, the less information is displayed, the easier will normally be the interactive process. Thus the relevance feed-

Query Type	Average Number of Significant Terms per Query
Original Query	8.3
Terms added in 1 minute	3.6
in 2 minutes	2.0
later	1.0

Variation in Query Length

Table 3

Processing Method	Demands on Computer	Demands on User	Precision Improvement over Word Stem Match	
			Low Recall	High Recall
A) Fully Automatic				
word stem match	normal	none	-	-
automatic thesaurus	normal	none	+4%	+6%
B) Pre-Search Interaction				
thesaurus display	normal +	medium-high	+6%	+4%
source document display	normal +	medium	+8%	+5%
C) Post-Search Interaction				
title display	high	medium	+13%	+2%
abstract display	high	very high	+17%	+7%
relevance feedback	high	low	+10%	+7%

Performance Summary

Table 4

back procedure is simplest, since the user must merely identify one or another document as either relevant or nonrelevant; the pre-search thesaurus displays, and the post-search abstract displays are hardest, since complicated decisions are required to update the search requests.

Turning now to the performance parameters, it is seen in Table 4 that, everything else being equal, the post-search methods are more powerful than the pre-search procedures. (Unfortunately, those are also the methods which put the highest demands on the computer). One obvious reason why the post-search methods operate more reliably is that a computer search has already been performed before the user is asked to update the query. Thus, the query alteration process is undertaken with prior knowledge of how well the original query has performed. The post-search alteration can then be used to initiate small changes for queries requiring only little improvement, and more massive changes for the others. For the pre-search methods, no such prior information is available.

Of the post-search methods, the best performance is obtained with abstract display; however, this method also makes the greatest demands on the user. The relevance feedback method is superior and much preferable from the user's viewpoint.

To summarize the performance and cost indications, the following search strategy would appear to be useful under most circumstances:

a) normally, use standard automatic thesaurus method without user interaction;

b) if improvement is needed and search time is not excessive, use relevance feedback;

c) if search time is at a premium, use pre-search source document or thesaurus display;

d) on the other hand, if high retrieval performance is mandatory, try post-search abstract display.

The difficulties of the manual query updating methods may be illustrated by the example of query 317, reproduced in Table 5. The original word stem retrieval run produces the two relevant documents in rank positions 4 and 10. From the thesaurus display, the following words were selected: "elastic", "resilient", "unstiffened", "modulus", "aeroelastic", "laminar-boundary-layer". This promotes the two relevant documents to rank positions 2 and 5; however, the automatic thesaurus run yields rank positions 2 and 4, without any user interaction. When the post-search displays are used, the results are similar. Title display is not very effective for this particular query, yielding only an indication that "theoretical" should be increased in weight, which raises the rank positions of the relevant from 4 and 10 (in the original word stem run) to 4 and 9. Abstract display is more fruitful, adding "elastic" and "resilient" as well. This increases the ranks of the relevant documents to 1 and 6. However, the same query, now processed through the automatic thesaurus (abstract display and automatic thesaurus run) yields perfect performance, as does the automatic thesaurus run with relevance feedback.

To achieve perfect performance using only manual updating methods and word stem matches, it is necessary to utilize a combined thesaurus display, abstract display, and word-frequency information,

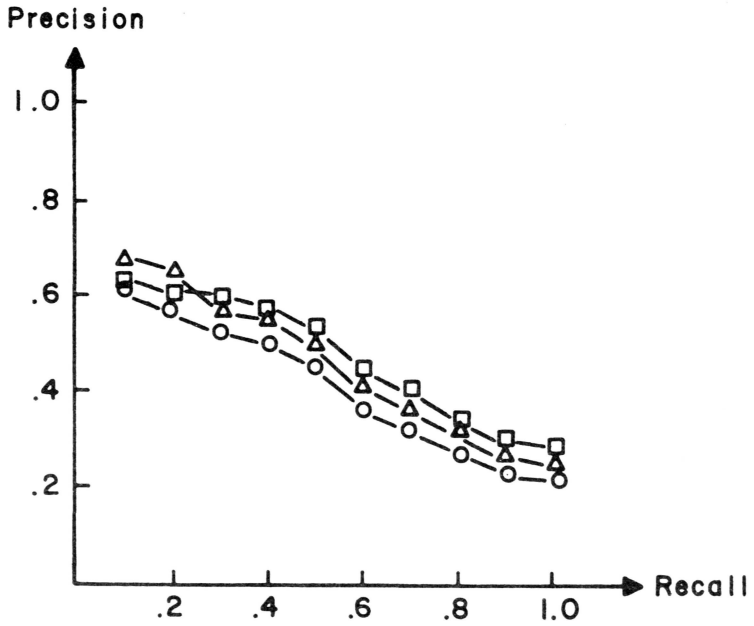
which yields the following rather complex set of changes: delete "anyone" and "investigate"; increase the weight on "theoretical" and "flexibility" by a factor of two; add with weight of one the words "analytic", "resilient", "calculate", "unstiffened", "aeroelastic", and "laminar-boundary"; add with weight of two "flexure"; and add with weight of three "elastic". These changes produce a word stem run with perfect performance, but at far greater time and trouble than the automatic thesaurus with abstract display run. The exact adjustment of the term weights is normally performed more accurately and more easily by the automatic thesaurus. The manual methods are thus best reserved for users with the skill and interest to consult lengthy displays and to make complex decisions.

A meaningful cost analysis is difficult to make without the use of an operational time sharing system to perform the experiments. Table 6 contains an estimated cost summary based on running times for the IBM 7094. Machine and user costs are assumed to be $75.00 and $10.00 per hour, respectively. Scanning time is 5 milliseconds per document, and additional central processor time is ignored. Table 6 shows that the post-search methods are clearly the most expensive (they also are the most effective), with relevance feedback relatively cheaper than abstract displays. In general, the automatic procedures appear economically and operationally better suited to the retrieval operations than the manual methods. Since the cost of human time may be expected to continue to increase relative to the cost of machine time, the automatic procedures may grow even more attractive in the future.

o original queries (word stem)

Δ dictionary display (word stem)

□ original queries (thesaurus)



a) Recall–Precision

Recall	Precision			T-test Significance	
	o	Δ	□	Δ / o	□ / o
0.1	.634	.691	.669	.580	.798
0.3	.534	.594	.605	.203	.339
0.5	.462	.510	.541	.199	.168
0.7	.343	.376	.411	.138	.181
0.9	.253	.292	.314	.061	.252

b) Recall–Precision Tables And Statistical Significance Output

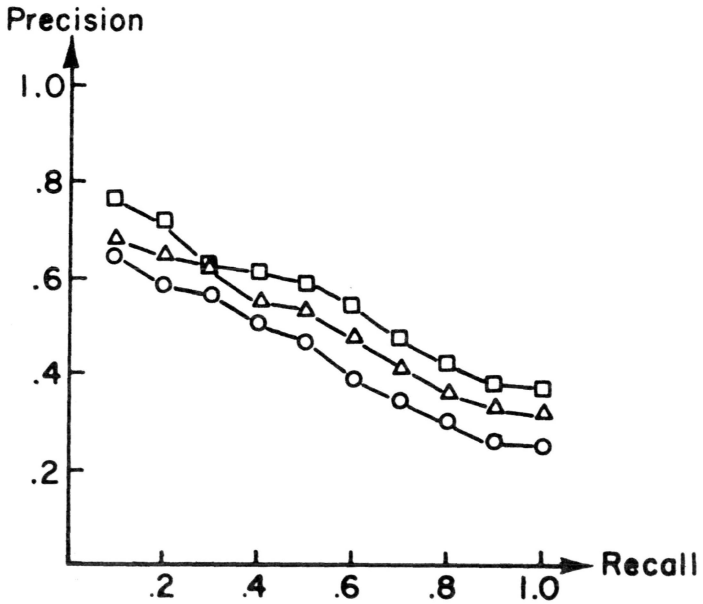
Effectiveness Of Dictionary Display Compared With Stored Thesaurus

Fig. 2

○ original queries (word stem)

△ original queries (thesaurus)

□ source document display (thesaurus)



a) Recall-Precision Graph

Recall	Precision ○—○	Precision △—△	Precision □—□	T-test Significance △ / ○	T-test Significance □ / △
0.1	.634	.669	.748	.798	.327
0.3	.534	.605	.603	.339	.869
0.5	.462	.541	.585	.168	.523
0.7	.343	.411	.470	.181	.767
0.9	.292	.314	.362	.252	.770

b) Recall-Precision Tables and Statistical Significance Output

Effectiveness of Source Document Display

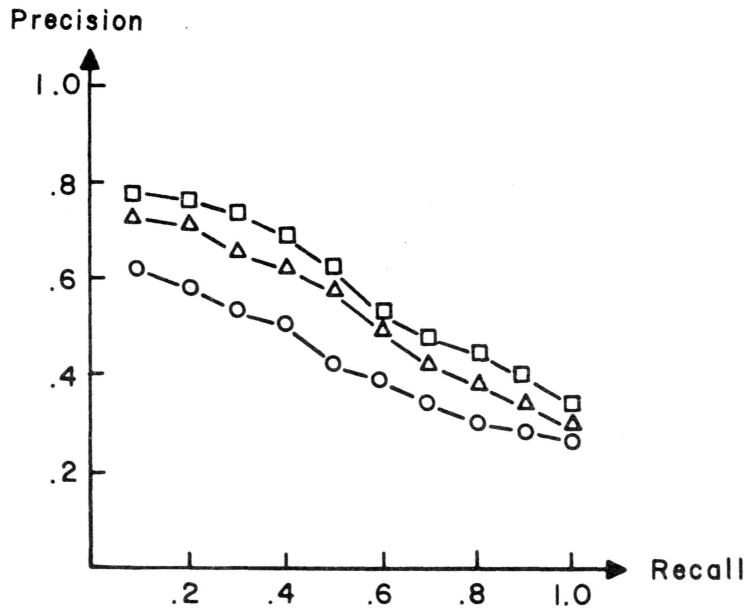
(averages over 200 documents and 42 queries)

Fig. 3

○ original queries (word stem)

△ title display (word stem)

□ abstract display (word stem)



a) Recall – Precision Graph

Recall	Precision ○	Precision △	Precision □	T-test Significance △/○	T-test Significance □/○
0.1	.634	.767	.799	.002	.003
0.3	.534	.627	.714	.002	.001
0.5	.462	.564	.627	.007	.001
0.7	.343	.377	.423	.014	.001
0.9	.253	.275	.328	.035	.001

b) Recall – Precision Tables And Statistical Significance Output

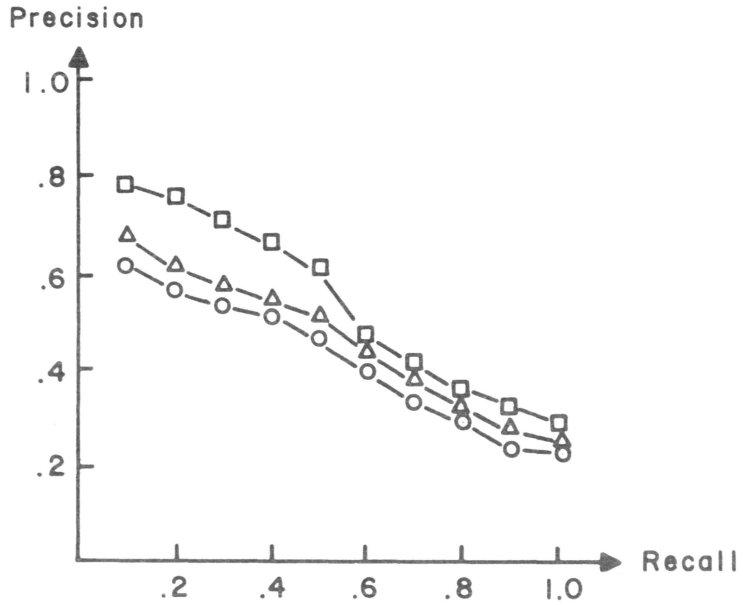
Comparison Of Title And Abstract Display

Fig. 4

o original queries (word stem)

△ dictionary display (word stem) 2 minutes

□ abstract display (word stem)



a) Recall – Precision Graph

Recall	Precision			T-test Significance	
	o	△	□	△ / o	□ / o
0.1	.634	.691	.799	.580	.003
0.3	.534	.594	.714	.203	.001
0.5	.462	.510	.627	.199	.001
0.7	.343	.376	.423	.138	.001
0.9	.253	.292	.328	.061	.001

b) Recall – Precision Tables And Statistical Significance Output

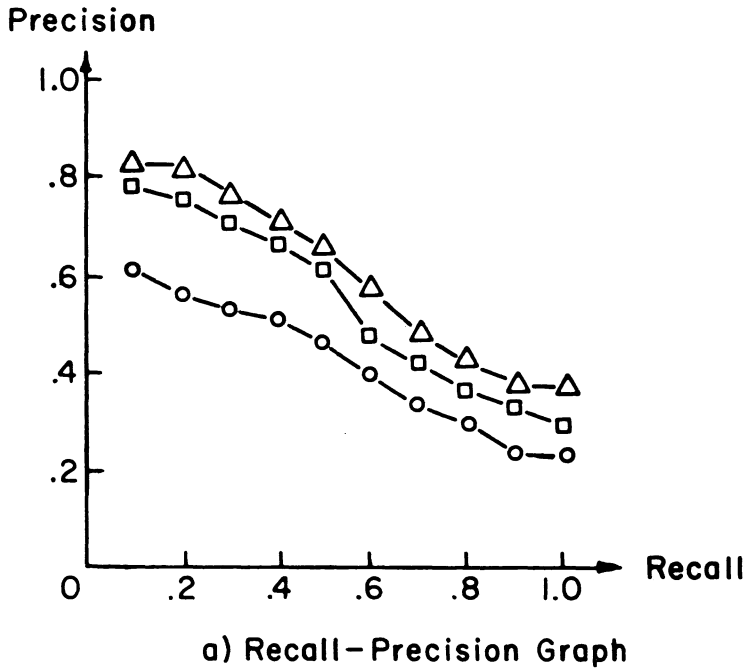
Comparison Of Dictionary And Text Display

Fig. 5

O original queries (word stem)

Δ relevance feedback (word stem)
one iteration - increment only

□ abstract display (word stem)



a) Recall-Precision Graph

Recall	Precision O	Precision Δ	Precision □	T-test Significance □ / O
0.1	.634	.809	.799	.003
0.3	.534	.784	.714	.001
0.5	.462	.704	.627	.001
0.7	.343	.510	.423	.001
0.9	.253	.388	.328	.001

b) Recall-Precision Tables and Statistical Significance Output

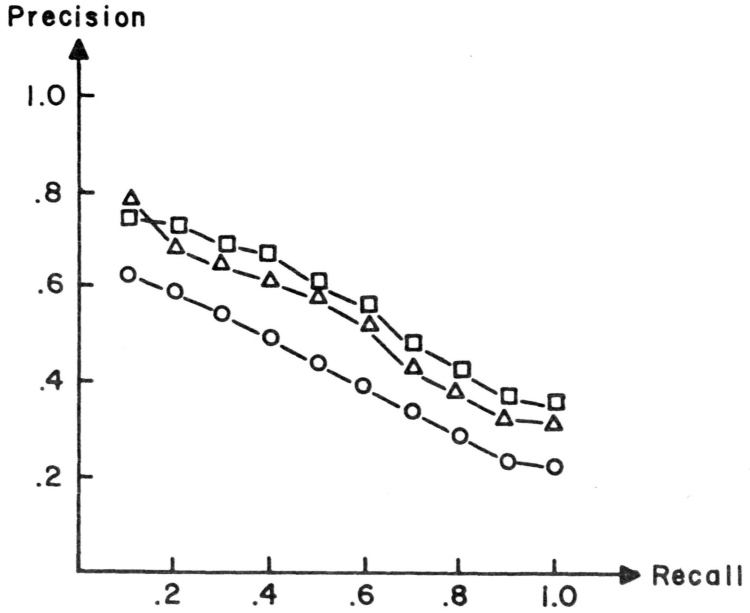
Comparison Of Abstract Display and Relevance Feedback

Fig. 6

o original queries (word stem)

Δ dictionary and abstract display (word stem)

□ all displays and frequency information (word stem)



a) Recall-Precision Graph

Recall	Precision			T-test Significance	
	o	Δ	□	Δ / o	□ / o
0.1	.634	.794	.787	.009	.009
0.3	.534	.668	.695	.007	.002
0.5	.462	.595	.631	.014	.006
0.7	.343	.445	.469	.001	.001
0.9	.253	.349	.361	.002	.002

b) Recall-Precision Tables And Statistical Significance Output

Comparison Of Combined Methods (Word Stem Process)

Fig. 7

Query 317: Has anyone investigated theoretically whether surface flexibility can stabilize a laminar boundary layer?
(Two Relevant Documents)

Processing Method	Terms "Added" or Deleted	Ranks of Relevant Documents
A) Fully Automatic		
word stem match	——	4,10
automatic thesaurus	——	2,4
B) Improved Searches		
word stem plus thesaurus display (pre-search)	"unstiffened", "modulus", "elastic", "resilient", "aeroelastic"	2,5
word stem plus title display (post-search)	"theoretical"	4,9
word stem plus abstract display (post-search)	"elastic", "resilient", "theoretical"	1,6
C) Perfect Searches		
automatic thesaurus plus relevance feedback	——	1,2
word stem plus abstract display plus automatic thesaurus	"elastic", "resilient"	1,2
thesaurus display plus abstract display plus word frequency display	anyone, investigate, "theoretical", "flexibility", "analytic", "resilient", "calculate", "unstiffened", "aeroelastic", "laminar-boundary", "flexure", "elastic"	1,2

Typical Manual Query Updating

Table 5

Processing Method	Estimated Cost per Query	
	50,000 documents	100,000 documents
A) Fully Automatic		
word stem match	$ 5.00	$10.00
automatic thesaurus	$ 5.00	$10.00
B) Interactive Pre-Search		
thesaurus display	$ 6.00	$11.00
source document display	$ 5.50	$10.50
C) Interactive Post-Search		
title display	$10.50	$20.50
abstract display	$13.00	$23.00
relevance feedback	$10.50	$20.50
D) Partial Search		
cluster searches (one-tenth of collection)	$ 0.50	$ 1.00
cluster search plus relevance feedback	$ 6.00	$11.50
cluster search plus abstract display	$ 8.50	$14.00

Assumptions: machine cost $75.00/hour
document scan 5msec/doc
central processing cost 0
human time $10.00/hour

Estimated Cost Figures

Table 6

The bottom part of Table 6 shows that processing cost goes down drastically if partial searches of the collection are performed, rather than full searches. Such partial "cluster" searches are implemented with the SMART system; however, the cluster searches cannot be used if a recall performance higher than about 50 percent is required [11].

6. Conclusion

The best overall process for precision purposes is the abstract display used in conjunction with a word stem matching procedure. For recall purposes, a combination of abstract display with thesaurus word normalization appears best. The automatic relevance feedback approximates the abstract display method while requiring much less user effort. Considering the complexity of the abstract display system, a sensible set of recommendations for high performance real-time retrieval would be the following:

a) for highest precision, use title display and word stem matching;

b) for highest recall with normal users, use the automatic thesaurus followed by automatic relevance feedback; with experienced and patient users, use abstract display and dictionary display plus frequency information;

c) for maximum cost reduction at lower performance, use partial searches of the document collection.

These rules provide a graded set of feedback methods, ranging from automatic procedures which make only minimal demands on the user and are suitable for novices (automatic thesaurus expansion, relevance feedback),

to methods permitting sophisticated user-system interaction which combine the best features of manual and automatic query adjustment (thesaurus and abstract display). One may expect that a suitable mix of user feedback procedures can be found to produce optimal retrieval under many different conditions over many types of user classes.

References

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

[2] G. Salton, et al., Scientific Reports on the SMART System to the National Science Foundation, Nos. ISR-11, ISR-12, ISR-13, Department of Computer Science, Cornell University, Ithaca, New York, June 1966, June 1967, and January 1968.

[3] C. W. Cleverdon, Jack Mills, and E. M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1, Design, Aslib-Cranfield Research Project, Cranfield College of Aeronautics, 1966.

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

[5] R. M. Curtice and V. Rosenberg, Optimizing Retrieval Results with Man-machine Interaction, Center for the Information Sciences Report, Lehigh University, Bethlehem, Pa., 1965.

[6] H. Borko, Utilization of On-Line Interactive Displays, in Information Systems Science and Technology, D. Walker, editor, Thompson Book Co., Washington, D. C., 1967.

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

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

[9] G. Salton, Search and Retrieval Experiments in Real-Time Information Retrieval, Proceedings IFIP Congress - 68, Edinburgh, August 1968.

References
(contd)

[10] E. Ide, User Interaction with an Automated Information Retrieval System, Scientific Report No. ISR-12 to the National Science Foundation, Section VIII, Department of Computer Science, Cornell University, June 1967.

[11] R. T. Grauer and M. Messier, "An Evaluation of Rocchio's Clustering Algorithm", Scientific Report No. ISR-12 to the National Science Foundation, Section VI, Cornell University, Department of Computer Science, June 1967.