XX. ANALYSIS OF STUDENT REQUESTS

C. Harris

Introduction

A test of the retrieval performance of the SMART system was conducted in conjunction with a graduate course in Applied Mathematics given at Harvard during the Fall of 1964. The object of this test was to obtain a set of retrieval requests from a group of persons in no way connected with the SMART research project and thus without previous knowledge of the document collection, the dictionaries or the system. Previous tests and retrieval results have been described in this series, Information Storage and Retrieval, Report No. ISR-8, Secs. IV and X, and will be referred to as Staff results.

The processing options used in this experiment were the following:

(1) thesaurus derived concept classes, document vectors derived from document titles only;

(2) thesaurus derived concept classes with request altered by addition of terms "up" in the concept hierarchy;

(3) thesaurus derived concept classes, concept vectors weighted logically (that is, with weights of 0 or 1);

(4) concept classes derived from word stems on a one-to-one basis;

(5) thesaurus derived concept classes with phrase detection (simulated syntactic search);

(6) thesaurus derived concept classes with request altered by addition of terms "down" in the concept hierarchy;

(7) thesaurus derived concept classes (no further modification);

(8) thesaurus derived concept classes with phrase detection (statistical).

All preceding methods employ a cosine correlation scheme. All concept vectors are weighted numerically, with the exception of method 3 which is weighted logically. Statistical phrase detection is the detection of co-occurrences of phrase components within a sentence. Simulated syntactic search eliminates phrases whose components bear no syntactic relation. The analysis which follows deals with the phrasing of the queries, the comparison of the evaluation parameters and the effectiveness of the dictionaries.

1. Description of the Student Requests

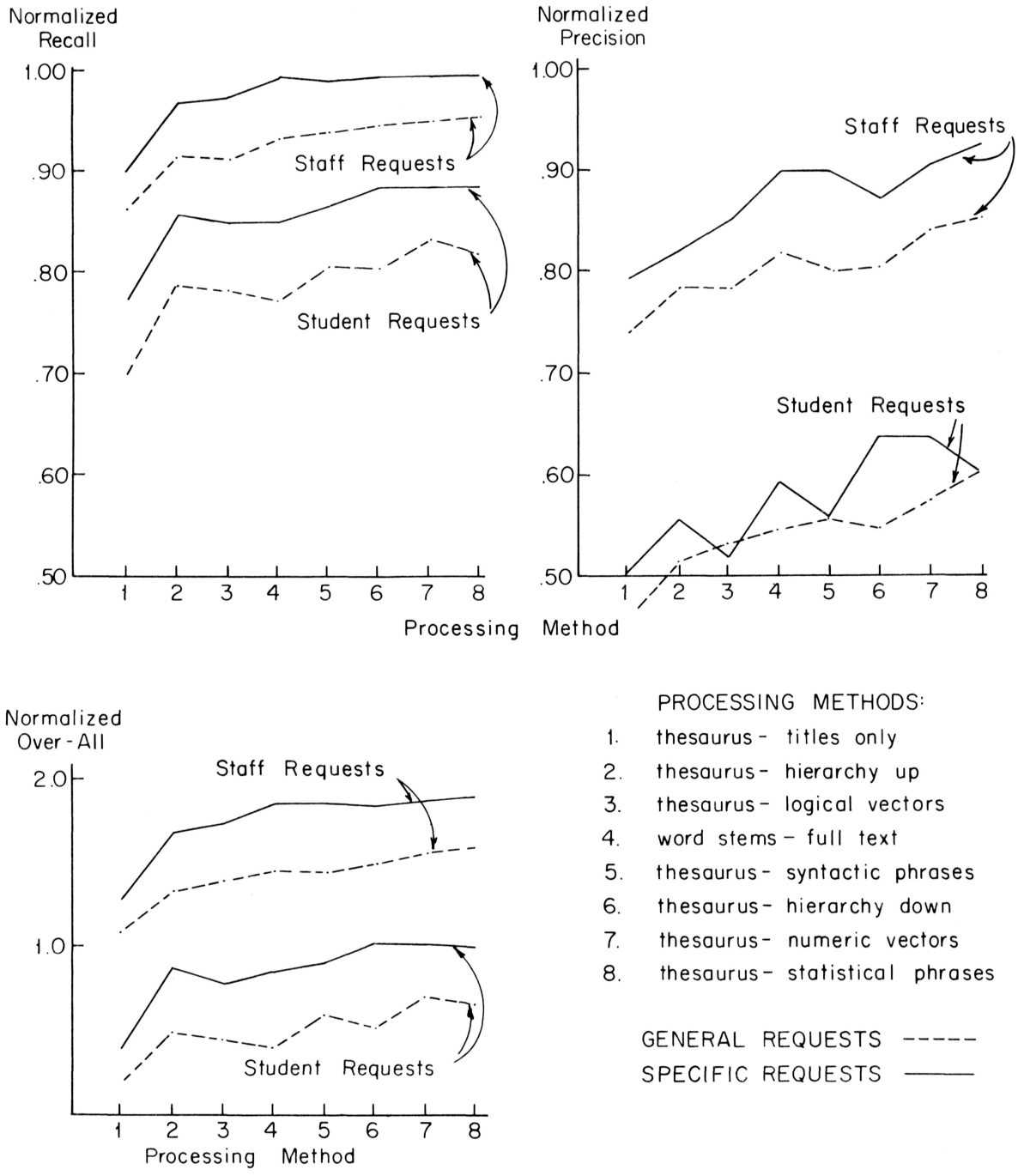
To keep the requests as unbiased as possible, a minimum of information about the SMART system and the document collection was given to the students. Each student was requested to submit two queries on a subject likely to be found in the computer literature. No instruction was given as to the language to be used in phrasing a request; furthermore, the document collection and SMART dictionaries were not made accessible at that time. The total number of requests submitted by the students was 42, ranging from one word to 78 words in

length. Although most requests were written in complete sentences, seven consisted of descriptive terms only.

After the requests had been prepared, the requestors were asked to study a listing of the document abstracts, and to supply a list of relevant documents to be used in obtaining recall and precision measures as described in Report No. ISR-8, Secs. IV and X. The 11 requests which could be satisfied by more than six documents were considered "general," in the terminology of the previous tests, (see ISR-8, Sec. IV-7) and the remainder were termed "specific." Of the specific requests, 7 were found by their originators to have no relevant documents in the test collection, leaving 24 specific requests which could be used in the test. It should be noted that the lists of relevant documents were provided by the requestors without consulting the computer retrieval lists. In several instances, students included with their relevance judgements the qualifying statement that they had found no truly relevant documents, but that the ones named were somewhat related. This well-intentioned stretching of relevance would tend to lead to poor retrieval performance, and should be kept in mind.

2. Retrieval Results

The Student requests were submitted to SMART and the results analyzed automatically. Average values of the normalized recall, precision and over-all measures for the general and specific requests are compared in Fig. 1 with the results of the Staff requests previously reported on. A number of results are at once apparent from Fig. 1:



Comparison of Averaged Measures for Staff and Student Requests

Figure 1

(a) the general trend of the results is similar for both Staff and Student requests;

(b) all measures obtained for Student requests are lower than for Staff requests;

(c) specific requests are consistently better than general requests, as had previously been reported;

(d) the use of hierarchy expansion in addition to the regular thesaurus is more effective "down" than "up;"

(e) the grouping of concepts by phrases, both syntactic and statistical, was less effective with the Student requests than with Staff requests.

The discrepancies between Staff and Student results have been examined in detail and certain valuable conclusions were reached. The generally lower normalized measures obtained in the later test are largely attributable to the test conditions, since relevance judgements and query statements were made without the originators being coached into the proper approach to SMART. However this provided in itself a worthwhile opportunity to judge the value of the system dictionaries. Whereas Staff relevance judgements were reached as a consensus, and were carefully pondered, the Student test relevance judgements received no such careful analysis. This in itself is enough to affect adversely the evaluation measures throughout the test.

The language of many student requests was also such that the SMART dictionaries would unfortunately, but unavoidably, assign much weight in the query concept vectors to rather insignificant ideas. As an example, consider the following request: "Tell me about numerical

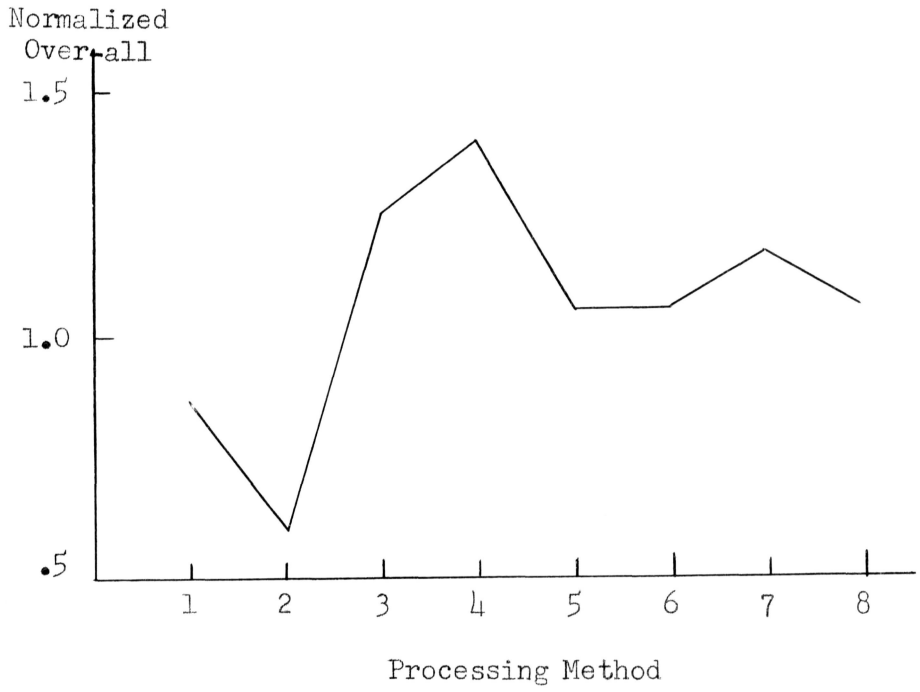
integration on computers." The three concepts of this query are 13CALC (numerical), 384TEG (integration) 110AUT (computer). Figure 2 gives the

Concept Identification	Terms	Frequency (No. of Documents)
13CALC	arithmetic, calculate, compute, evaluate, figure, interpolate, numerical, plot, reckon, recompute, value	145
384TEG	integral, integrate	13
110AUT	automaton, calculator, computer, data-processor, processor	219

Excerpt from the Thesaurus and Corresponding Frequencies of Occurrence from the Document Concordance

Figure 2

thesaurus lists of these concepts and their frequencies of occurrence in the document collection as determined by a concordance of concepts numbers in the whole collection. It is immediately apparent that in this instance the broad classifications of the thesaurus are a hindrance to retrieval rather than a help, and this explains why this particular request was best answered by use of the null dictionary (see ISR-8, Sec. VII-3) which is made up of single word stems obtained from the whole text of all documents. Figure 3 shows the normalized over-all measure for this request for all processing methods. Processing by word stems is method 4 in Fig. 1 and Fig. 3. Again, from the data of Fig. 2, it is apparent that no processing option will be able to improve much on the performance of the regular thesaurus

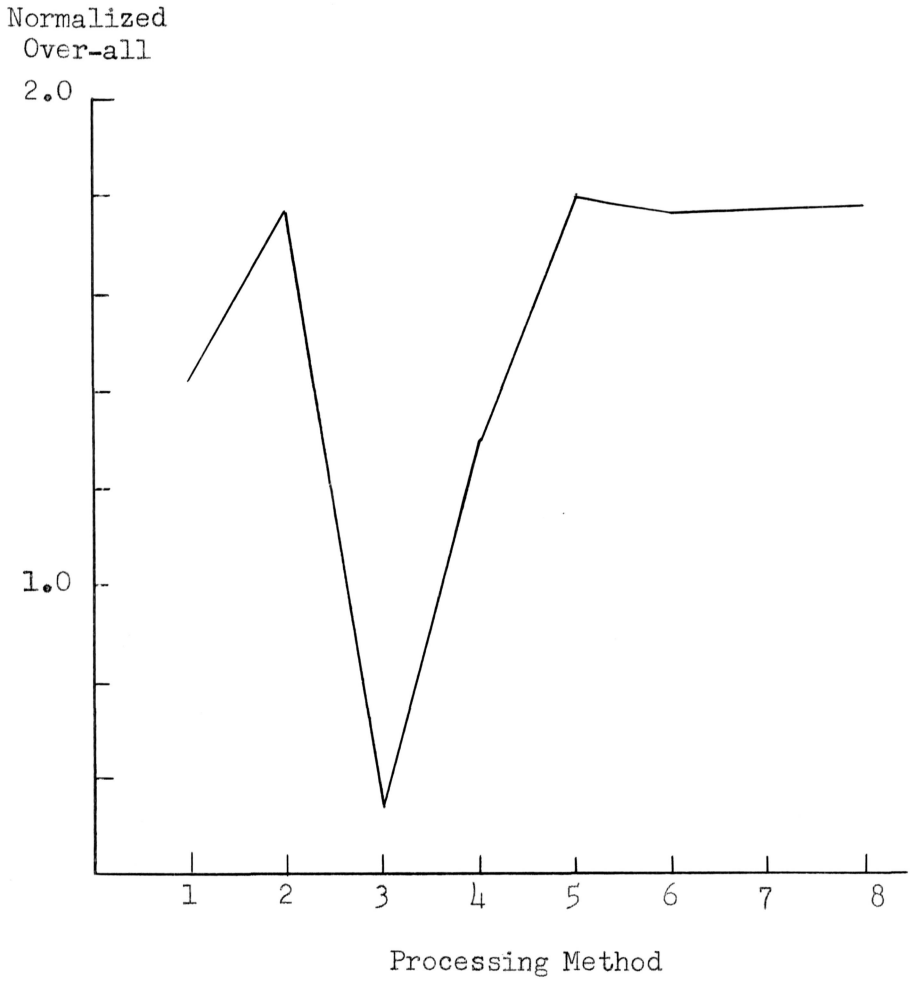


Normalized Over-all Measure for the Request "Tell me about Numerical Integration on Computers"

Figure 3

run (method 7): since there are no phrases in the request, phrase processing is of no help; since two of the concepts are broad, hierarchy expansion will only further confuse the retrieval of relevant documents; only processing by logical vectors is an improvement, since multiple occurrences of high-frequency terms within a document will no longer carry extra weight. In conclusion, the performance of this request might be improved by restricting the thesaurus classes, by recognition of the phrase "numerical integration" (which was not in the phrase library at the time the request was submitted) and by avoiding the word "computer," which is assigned to 219 documents as might be expected in a collection of computer abstracts.

By contrast, let us examine a request which performed much more satisfactorily. Figure 4 gives the normalized over-all measure for all



Normalized Over-all Measure for a Student Request
on Pattern Recognition

Figure 4

processing methods for the following query: "What is there written about pattern recognition in configurations made up of dots. That is, tell me about recognizing patterns where the figure is recognized

by difference in the density of the dots." The concept vector of this request is shown in Fig. 5. The vector represented in Fig. 5 includes statistical phrases weighted 3:1 as specified in the SMART phrase option for this test; 5 phrases were found.

2INPUT 4	13CALC 4	58PLAN 28	107DGN 12	129NUM 4
130MEA 4	184DEC 6	2100UT 4	246HRD 6	332SEE 216
340LET 12	393DIF 6	412FLD 6	478CEN 12	

NOTE: The number to the right of each concept identification is the weight assigned to it.

Numeric Vector for a Student Request on Pattern Recognition (Phrases Included)

Figure 5

From Fig. 5, it is seen that the regular thesaurus run with numeric vectors (method 7) is far superior to the run with word stems (method 4). This is an indication that the thesaurus classes of concepts in the query vector are properly constructed. In addition, the weights attached to the concept numbers, as shown in Fig. 5, tell us that the request is properly worded so that significant concepts carry a high weight, while ideas that are less useful do not. When the weight difference is removed, as in the logical vectors option (method 3), a sudden dip in the over-all measure takes place as expected.

It is also expected that a request that contains 5 statistical phrases should show a marked improvement when phrases are processed.

A look at Fig. 4 shows that this did not happen. To explain it, we must go to the actual rank order lists of documents for methods 7 and 8 which are reproduced in Fig. 6. From these lists we are able

REGULAR THESAURUS (method 7)			THESAURUS WITH PHRASES (method 8)		
Top 10 Documents	Relevant		Top 10 Documents	Relevant	
	Rank	Document		Rank	Document
351	1	351	353	2	82
350	2	350	82	3	351
353	5	82	351	4	350
348	6	163	350	7	163
82	18	1	225	9	1
163	32	205	314	34	205
335			163		
92			162		
48			1		
113			68		

Rank Order Lists with and without Phrases

Figure 6

to see how the rank of a particular document was affected by phrase processing. Document 205 remains approximately at the same rank because it contains no phrases; document 1 is brought up from r = 18 to r = 9; documents 82, 351, 350 and 163 maintain their high ranks except that first place has been lost, so that the precision value is much reduced. Upon examination, the irrelevant documents in the phrase run with ranks 1, 5, 6 and 8 prove to be also on the subject of pattern recognition, but not on the recognition of "dot

configurations." The thesaurus category for pattern recognition includes the recognition of speech patterns, and the phrases correctly identified and weighted in these four documents were misunderstood in answering the query. In conclusion, the retrieval success was not better because the thesaurus was not capable of distinguishing between specific types of pattern recognition; the phrase dictionary did not include phrases describing different kinds of patterns and equated all phrases that included the word "recognition."

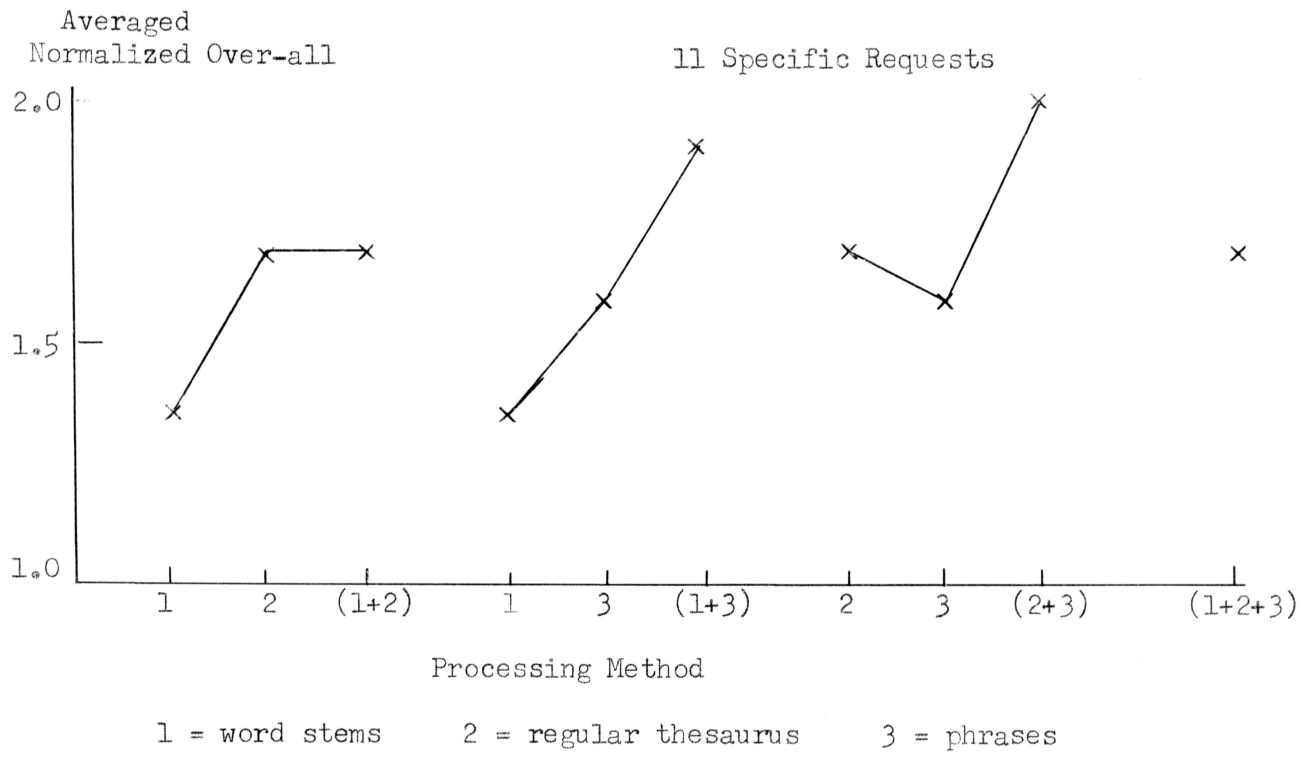
3. Combined Processing Methods

In Report No. ISR-8, Sec. IV-27, the use of combinations of processing methods was described. It was conjectured that performance characteristics might be improved by processing a search request by several methods and combining the respective outputs.

In the present test, twelve pairs and nine triple combinations of methods were calculated by SMART. However, in order to obtain meaningful analyses, it seemed necessary to re-examine each individual request in relation to the validity of its relevance judgements and to its performance on SMART with single processing methods, since there is a potential error in averaging results over requests that are not comparable. As an empirical means of selecting comparable requests, that is requests with reliable relevance judgements and reasonably clear wording, those requests were selected for study which had exhibited a relatively adequate performance with single methods.

This point of adequacy was arbitrarily set as the retrieval of more than one half the relevant documents in the first 15 for the specific requests, and more than one third the relevant documents in the first 15 for the general requests. The selection reduced the list of queries to 11 specific and 5 general, or approximately half of the original group.

The normalized over-all measure averaged over the selected requests is shown in Figs. 7 and 8 for combinations of the methods of word stems, thesaurus and thesaurus with phrases. In all cases, the



Normalized Over-all Measure for Various Combinations of Processing Methods Averaged over 11 Specific Requests for which More than Half the Relevant Documents were Retrieved

Figure 7

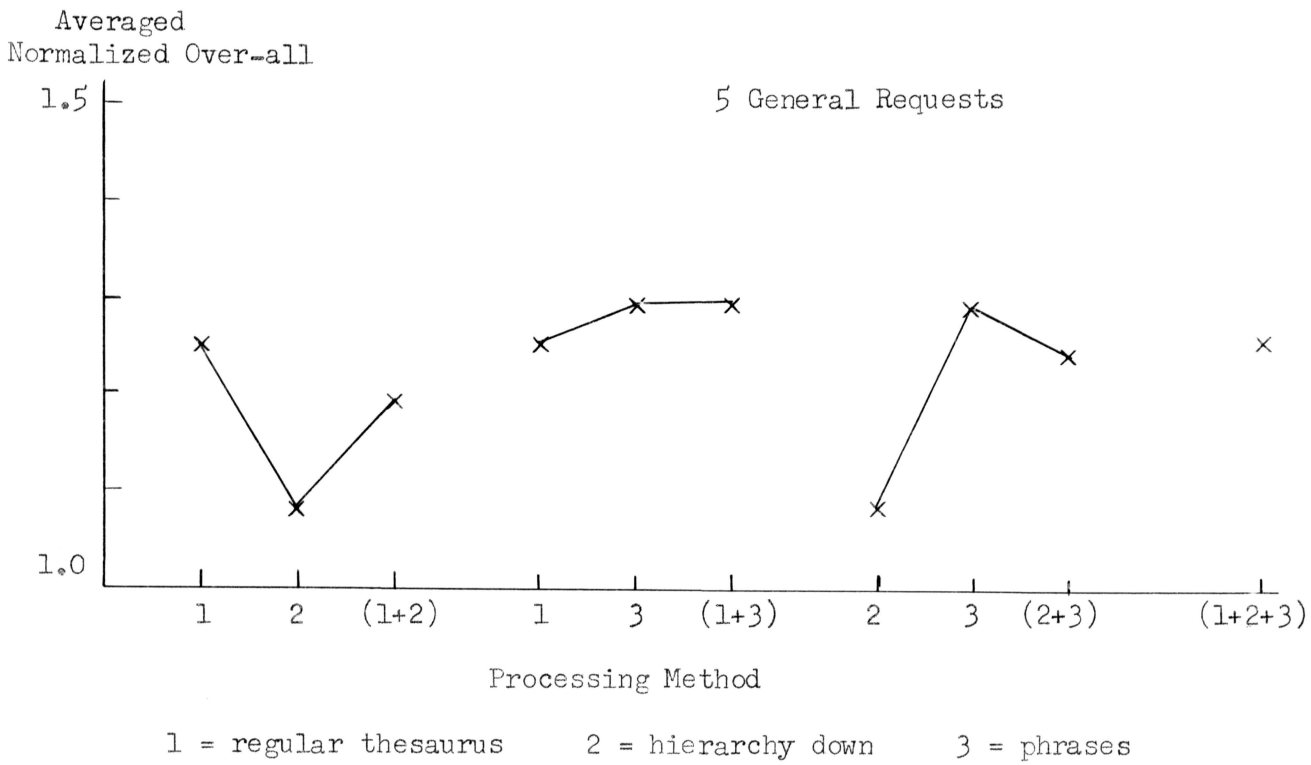
Normalized Over-all Measure for various Combinations of Processing Methods Averaged over 5 General Requests for which More than 1/3 the Relevant Documents were Retrieved

Figure 8

paired methods give a better performance than the component methods taken singly, but the triple method is not an improvement over the best pair. When we remember that the evaluation measures for the combined methods are calculated from a combined rank list obtained by merging of the component rank lists, we see that it is important which nonrelevant documents occur among the relevant documents on the single rank lists. The results of the merging of rank lists is

therefore not predictable. The best results are found when the two processing methods to be combined each retrieve the relevant documents with different ranks, so that the last ones on one list are the first on the other, but with the same intervening irrelevant documents. Retrieval of relevant documents with different ranks is most likely to happen when the processing methods are basically different.

As an example of an unsuccessful merging of methods, Fig. 9 compares combinations of the regular thesaurus, thesaurus with phrases



Normalized Over-all Measure for Various Combinations of Processing Methods Averaged over 5 General Requests for which More than 1/3 the Relevant Documents were Retrieved

Figure 9

and thesaurus with hierarchy processing. In this case, expansion down the hierarchy introduces so many extraneous documents that the rank order lists are diluted and any combination that includes the "hierarchy down" method is at once worse than the best single component.

4. Revision of Requests and Dictionaries

It is apparent from the preceding analyses that revision is necessary both in the dictionaries and in the form of some of the requests. On the whole, the results of this test were similar to those obtained for the set of Staff requests, but the generally lower performance of Student requests suggests several possible improvements:

(a) the broadest thesaurus classes need to be studied and refined to prevent ambiguity;

(b) the phrase dictionary needs to be greatly expanded if full advantage is to be taken of the statistical phrase option;

(c) the hierarchy structure must be limited to groupings of concepts which are unambiguous and should preferably not contain concepts with a high frequency of occurrence in the document collection;

(d) requests should be written in an unambiguous way and not contain high frequency concepts unless they are specific to the query;

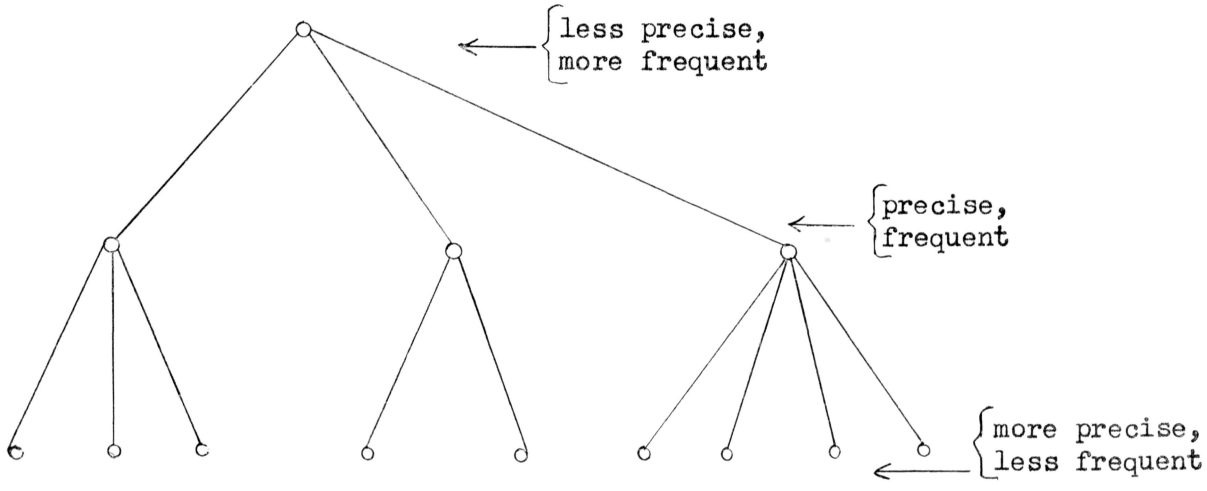
(e) familiarity with the SMART dictionaries is desirable.

An extensive revision of the SMART dictionaries has been made along the lines indicated above. All concept classes in Thesaurus 2

have been reviewed with the help of a document concordance and a frequency listing of word stems. The broader, high-frequency concepts have been split up, and Thesaurus 3 now consists of 736 concept classes, as against 511 concepts in Thesaurus 2. However, the total number of word stems has not been changed.

The phrase dictionary has been increased from 98 to 373 phrases covering a wider variety of combinations of word stems. In particular, numerous phrases have been included which specify the usage of common terms such as "function," "time," "system," "control;" (typical phrases are "manual control," "traffic control," "accuracy control," "control system," "feedback control," "flight control").

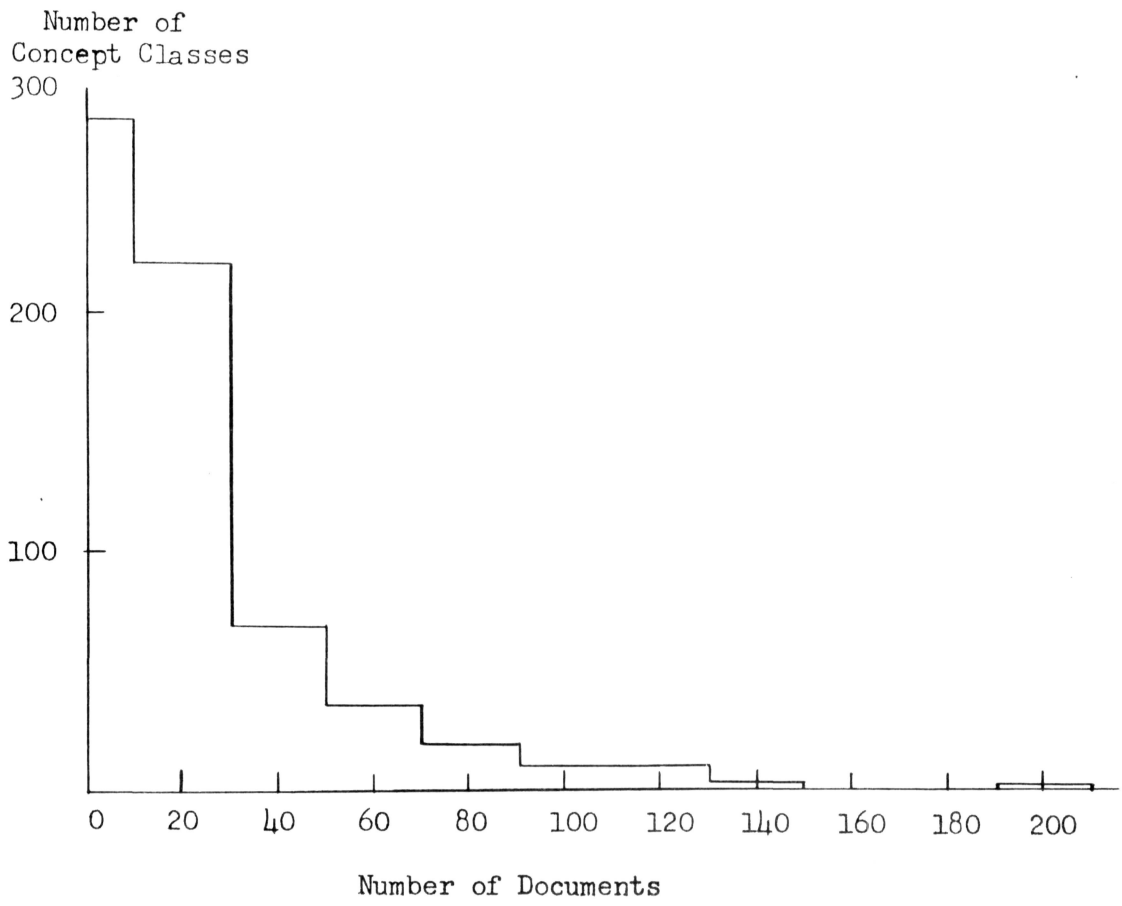
Because of the extensive changes in the thesaurus and phrase dictionary, the hierarchy has been completely re-written. Figure 10 shows the format which the hierarchy attempts to follow.



Hierarchy Format

Figure 10

In establishing a three-level hierarchy, the aim is to restrict the distance between related nodes, and thereby increase the likelihood that hierarchy processing will produce matching concepts. High-frequency concepts are, in general, untied to the hierarchy except in those instances where a filial connection is unambiguous. In addition, an extensive system of cross-references links various sections of the hierarchy. These cross-references can be called by the hierarchical expansion option REFER. Hierarchy expansions "up" (STEM) and "down" (LEAF) within concept groupings are the ones which have been tested in the current tests.



Frequency Distribution of Concept Classes in the Document Collection

Figure 11

Concept No.	Frequency (No. of Documents)	Terms
601	230	computer,DACOMP (data processor),data-processor, electronic-computer
496	135	system
615	108	operate
608	104	program
603	101	digital
32	98	dictate,impose,necessary, necessit,need,quali,require, requisite,restrict
614	88	apply,applic
605	87	calculate,compute
202	85	complete,entire,full,general, total,universal,whole
92	84	DIGCMP (digital computer)
600	82	machine
215	80	acquire,detect,find,get,note, obtain,recover,refer
121	79	memory,store
623	79	circuit,network
609	76	design
602	74	problem
68	71	apparatus,device,equip, hardware,instrument
604	70	analog
119	70	artificial,automat,mechan
618	69	function
147	67	handle,process
227	67	charge,electric,electron

Most Frequent Concept Classes According to Thesaurus 3

Figure 12

INSTRUCTIONS TO REQUESTORS

1. Requests to SMART may consist of full grammatical sentences or of series of terms.

2. Most functional words (such as prepositions, adverbs, conjunctions, auxiliary verbs) and many common descriptive terms are ignored by SMART and can safely be included. However, idioms which contain ambiguous words may be misunderstood. For example, "in order to," "in addition," "a number of" have the meanings of "order," "addition" and "number."

3. It is important to select words which are unambiguous and which specifically define the desired subject. It is especially desirable to use terms which are specific to the subject, rather than the broad terms which apply to a wide range of computer technology.

4. Additional weight is given by SMART to repeated terms. Therefore it is useful to write more than one sentence and to repeat the significant phrases. Caution: the repetition of broad terms, such as "computer, system, operate, structure, design, calculate..." which are frequent in the document collection, should be avoided.

5. Affirmative sentences should be used since SMART ignores negative words.

6. Synonyms of the more esoteric technical terms could be used, or the SMART Thesaurus should be consulted, if one is available, to insure that all significant meanings will be found.

Sample Instructions for Submitting Requests to SMART

Figure 13

In phrasing requests to SMART, it is desirable to avoid high-frequency terms unless they are thought to be distinctly useful in retrieval of related documents. The document concordance for Thesaurus 3 is shown in Fig. 11, and the concept classes which occur in more than 65 documents are listed in Fig. 12. Sample instructions for submitting requests to SMART are given in Fig. 13.