X. An Analysis of the Documentation Requests

E. M. Keen

1. Introduction

Conclusions from the experiments performed with SMART have so far been made on the basis of the average results achieved for a set of search requests. Studies presented in sections III, V, VI, VII, VIII of this report have presented such results, and given in addition some elementary data about the individual requests, and examples of detailed system performance. The purpose of this study is to uncover still more of the variables and characteristics of the requests themselves in the context of the test environment and of the capabilities of the SMART system, in order to reveal basic problems and suggest improvements to the system. The 35 requests associated with the 82 document collection in documentation are used because the author's knowledge of the subject field aids the investigation.

2. Request Preparation

In section I, a brief description of the ADI documentation collection, requests, and relevance decisions is given. The collection consists of 82 documents, both in abstract and short text form. Requests were prepared by two non-users of the system, and the relevance decisions were also made by these persons by examining every document for potential relevance.

The request preparers were graduate students at Harvard University, one an engineer, and one an applied mathematician. Since neither person was familiar with the subject field of documentation, it was suggested that some familiarity with the subject would be gained by looking at the documents in the collection. The task of request preparation should then follow. Specifically, requests thought likely to be asked by workers in the field should be devised, but these requests should not build in any particular documents in the collection. No suggestions were made regarding request length.

An examination of all documents in the collection should follow, and every document should be judged not relevant or relevant. Neither person was familiar with the SMART system.

The full text of the 82 documents was supplied in the form of computer print-out, and none of the KWIC indexes or subject categories printed in the published volume were supplied. Both preparers worked independantly, one producing 17 requests and the other 18. The task appears to have been carried out as instructed, except that it is suspected that some requests may have been prompted by particular documents in the collection. This factor is not necessarily a weakness, since full relevance decisions were obtained, and thus the test is not based on the "source document" technique often criticized. Some comments on the task will be made when discussing "Unclear Requests"(part 2D) and "Relevance Decisions" (part 3).

3. Characteristics of the Requests

A) Length

Excluding non-subject words contained in the standard "common" word list used by SMART, the stem dictionary gives an average of 8.0 stems per request, and the thesaurus dictionary reduces this to 5.1 concept numbers per request. The eight most frequently used thesaurus concepts are given in Figure 1, together with the numbers of requests and frequency of use in the collection. More than half the 35 requests use one or more of these 8 concepts.

Comparison of request length between the two different preparers shows that person A constructed longer requests than person B; this matter is considered in part 5C.

B) Important Request Words

Since the requests are fairly short and many use words of high frequency in the collection, it would be expected that some requests would contain one or two quite important words that are vital to the request demand. For example, request A15 reads "How much do information retrieval and dissemination systems cost?", and request B4 "Automated information in the medical field". The words "cost" and "medical" are very important in the request statements, and render otherwise general requests much more specific. Since SMART gives weight to each request word in part on the basis of frequency of occurrence in the request and collection, these important words are liable to fail to receive the desired weight; this problem is taken up again in part 5D.

C) Multiple Need Requests.

Nearly all requests express the need as a single topic, but two requests demand documents on two topics. Request B1 asks for information on both coding and matching in machine systems, and request A1 requires information on "titles", meaning journal titles, organization names and presumably their abbreviations; and also "titles" meaning the subject state-

TERM (STEM)	THESAURUS CONCEPT NO.	NUMBER OF REQUESTS			FREQUENCY IN COLLECTION	
		TOTAL	*"A"	*"B"	STEM†	RANK†
Information	1	24	11	13	998	1st
Retrieval	5	14	12	2	255	14th
Automat	19	14	7	7	86	86th
Comput	3	7	2	5	326	8th
Dissemination	108	6	3	3	47	124th
Article	21	6	5	1	86	86th
Index	4	5	1	4	424	3rd
Scienc	10	5	4	1	300	9th

* Requests prepared by Persons A and B

† Frequency is based on word stems, therefore "Information" is the most frequent stem (Rank 1st) and is used 998 times.

Eight Thesaurus Concepts Most Frequently Used in the Requests Giving Frequency Data for Each Major Stem

Figure 1.

ment as applied to scientific papers. There is nothing inherently wrong about having multiple need requests, but for testing puproses such requests sometimes cause difficulties when binary relevance decisions are used. For example, of the three documents assessed as relevant to request A1, two clearly answer the first part of the request only, and one answers the second part of the request only; thus it is never possible, even for a perfect system, to establish a complete match between the request and the relevant documents. It is believed that where multiple needs are expressed separate search requests will give superior results.

D) Unclear Requests

Two requests in particular are unclear. In request A1 does the phrase "approximate titles" mean abbreviated titles? Does request A8 ask for documents in information retrieval as practiced in countries other than those speaking English, or, is it asking about information retrieval (practiced anywhere) of documents written in languages other than English? The full request statements may be examined in Appendix B.

Several other requests may be charged with perpetuating the unclear terminology that abounds in the field of documentation. Request B2, what does an "automated" information system include and exclude? Request B3 requires documents either describing the shortages that exist of information personnel, or some solutions to the problem such as the need to provide suitable training. Request B11 uses the words "index system", and since only one document on the cataloging of books is judged relevant, "index system" has been taken to be synonymous with book cataloging only.

Such requests would, in an operating situation, be clarified by interaction with the questioner; this advantage is, however, denied to the

SMART tests. It is surprising to note that the five requests quoted do perform quite well in the retrieval runs made, and of the total of 13 relevant documents involved only 3 receive consistently poor rank positions (below 15) on all search options.

E) Difficult Requests

Of considerable interest in the analysis of a system such as SMART is the identification of requests that may be quite reasonable in themselves, but that nevertheless create problems due to some system weakness. Six examples are given.

Request A2 contains the following negative statement: "...As opposed to references or entire articles themselves ...". SMART cannot recognize the significance of the negation, and a search will be made for the ideas as stated. Unless rules to recognize negative statements can be added to the system, users or request preparers must be advised to avoid negatives.

In request A8, "other languages" is a very important part of the request, but the idea of "languages other than English" is another negative statement which cannot be handled. Even if "other" were replaced by "foreign", correct matches with relevant documents would be difficult to achieve since a thesaurus concept that links "foreign" with all possible named languages or countries might work well for this particular request, but would at the same time provide an unhelpful grouping for other requests asking for one language in particular.

Request A10 contains the homonym "abstract", here used in the sense of "abstract mathematics" rather than the frequent collections use in the sense of a summary of a document. The use of phase recognition would cope with this problem, except that the phrase list in use does not contain the required phrase. A synonym problem also exists, because none of the relevant documents use the phrase "abstract mathematics". Only in one case was "abstract" used in a sense other than "document summary", namely, in "abstract trees". Interaction with the requestor seems necessary here, or a demand to list at least some example of "abstract mathematics".

Request All reveals the problem of ordinarily common words being used in a technical sense. Words such as "evaluation" and "need" are relegated to the common word list when the thesaurus is used, thus leaving a request specification only in terms of high-frequency words in the collection. The stem dictionary uses both words and gives a better performance result; however, since such words frequently occur in non-technical senses, two of the four relevant documents receive poor (lower than 15) rank positions. There seems to be no way of coping with such problems except to get requestors to supply alternative and less ambiguous words where possible.

Another example of this kind is in request A13, in which "criteria", "objective" and "evaluation" appear. In this case inclusion of these request words in the stem dictionary results in good rank positions for 5 of the 6 relevant documents. Where several such ambiguous words occur, the co-occurrence of all of them in a document in the incorrect sense is less likely; an improved type of phrase dictionary may overcome the problem.

A problem of synonym recognition is raised by request B13. The phrase "physical sciences" is really ambiguous and not very well chosen, since examination of the relevant documents reveals that it covers notions such as "materials", "chemistry", "engineering", "technology", "missiles and space technology" and "environmental engineering". The use of such wide ranging relations in a thesaurus concept would be reflected by a concept number with very many corresponding words; this would not serve all types of requests equally well, and would in any case require some recognition of phrases rather than single words.

These examples of difficult requests point out two areas in which future work is required. The first is the problem of ambiguity caused by natural language, which may partially be handled by sophisticated recognition procedures (to include negative statements, for example), but may in other cases only be handled by introducing constraints to the free statement of the request. The second problem is that of making synonym connections in cases where a generic term is used. A possible solution is the provision of more than one synonym dictionary, which would include one containing some quite large groupings of many words into few concepts to handle the difficult cases.

4. Relevance Decisions

Since both request preparers were alone responsible for the requests and relevance decisions produced, there arose no possibility for disagreement during the setting up of the test. Evidence suggests that a consistent and conscientious job was done, although it is very easy to argue that the judges were not competent, or that real information needs did not arise, and so on. Measurement of the accuracy with which this artificial procedure can simiulate real user requests, needs, and relevance decision awaits a carefully controlled comparative test. To submit the actual judgments made to a panel of judges for their opinion would undoubtedly reveal disagreements with the request preparers, and probably among the judges themselves; such a procedure would then serve no real purpose at the present time. A cursory look at the decisions has been taken, and some discrepancies are noted in order to illustrate the probable types of deficiency that may exist.

The discussion of unclear requests in part 4D is closely linked to the relevance decisions, since as J. O'Connor shows [1], relevance disagreements are often due to unclear request forms; furthermore, since many requests that are thought to be clear are not so in fact, one is led to different request interpretations and hence to different relevance decisions. Probably more examples than the five given in part 4D exist, but by the stringent criteria for clarity suggested by O'Connor, many real user requests would be regarded as unclear also.

Several of the requests deal with quite similar topics, and sometimes do not have as many relevant documents in common as the requests suggest. Examples are requests A15 and B8, B1 and B16, B9 and B11, and A5, B3 and B6. A clear error of judgment is seen for document 7, where the photo cam position method that is described for producing NASA's "Scientific and Technical Aerospace Reports" is thought relevant to request A7, which demands documents on systems for producing original papers by computer. The request preparer probably did not realize that NASA STAR is not a series of original reports. No specific examples have been found of documents that should have been recognized as relevant, except in those cases where two or more requests seem very similar, as noted.

5. Request Performance

A) General Performance Analysis Methods.

It was intended to divide the individual requests into three groups, namely:

a) Requests which perform badly on all processing options;

b) Requests which perform well on some options and badly on others;

c) Requests which perform well on all options.

Definitions of good and bad performance are arbitrary, but it is thought that good performance requires the rank position of a relevant document to be at least 15, and anything positioned lower than this is a poor result. Any requests which fall into groups a) and c) were thought to be particularly useful for analysis; in practice, however, all 38 requests fall into group b). Requests B6 and B14 perform well on nearly all options, but occasionally one of the relevant documents falls below rank position 10. There occurs a surprisingly large amount of change in the ranks of the relevant when options are tested; Figure 2 gives an example for one request and two relevant documents. In this request, all the options that are found on average to be the poorest, such as titles only, the use of cosine logical, and the "Hastie" Thesaurus give the best results.

Since the division into groups by performance achieved does not assist in the analysis, another method of analysis is suggested: this is to look for strong correlation between measurable request characteristics and the use of particular performance options. A summary of possible request characteristics is given in Figure 3, some of which have been described previously; these can now be used to look for direct correlation between characteristics and performance, as attempted in sections 5B, 5C, and 5D.

B) Variation in Generality, Length and Concept Frequency

Request generality refers to the number of documents in the collection that are relevant; using this principle, the request set may be divided into specific and general requests. With the 35 requests divided into sets of 17 and 18, request generality data is given in Figure 4 together with evaluation results of normalized recall and precision, comparing the stem and thesaurus dictionaries. As has been observed previously [2], the specific requests give

DOCT. LENGTH	CORRELATION FUNCTION \ DICTIONARY	STEM		THESAURUS-1		THESAURUS-2(HASTIE)	
		DOCT.82	DOCT.50	DOCT.82	DOCT.50	DOCT.82	DOCT.50
TEXT	Overlap Logical	60	43	59	36	40	65
TEXT	Cosine Logical	9	44	$\boxed{7}$	24	①	57
TEXT	Cosine Numeric	63	52	29	61	14	⑤
TITLE	Cosine Numeric	67	$\boxed{8}$	41	38	49	22

○ Relevant Document having best rank

□ Relevant Document having second best rank

◌ Request having best performance

⬚ Request having second best performance

Rank Positions of the Two Documents Relevant to Request A9,

Comparing 3 Dictionaries, Two Document Lengths, and Three Matching Functions.

Figure 2.

REQUEST CHARACTERISTICS	OBJECTIVE	SUBJECTIVE
1. Number of documents judged relevant in test collection, or request generality.	✓	
2. Length of request, using all words, subject words only, or dictionary concept numbers.	✓	
3. Clarity of request demand and quality of relevance decisions.		✓
4. Request processing in the system used, such as SMART. Factors such as quality of dictionaries used and frequency of use in the collection of the request words.	✓	✓
5. Test environment variables, such as the use of several personnel to prepare requests.		✓

Summary of request characteristics showing which are objective and which subjective from an evaluation viewpoint.

Figure 3.

DICTIONARY	EVALUATION MEASURE	ALL, 35 REQUESTS 1-33 RELEVANT, MEAN 4.9	SPECIFIC, 17 REQUESTS 1-3 RELEVANT MEAN 2.1	GENERAL, 18 REQUESTS 4-33 RELEVANT, MEAN 7.4
Stem	Normed. Recall Normed. Precision	.7601 .5326	.8018 .5358	.7209 .5295
Thesaurus 1	Normed. Recall Normed. Precision	.8016 .6069	.8482 .6126	.7576 .6016

Performance Results Comparing All the Requests, the Specific Requests only and the General Requests only, with Two Dictionaries.

Figure 4.

a better performance than the general ones, although normalized precision shows only a small difference. There is no correlation at all between request generality and dictionary in these results.

Request length results are given in Figure 5, with the requests again divided into two sets. The long requests perform better than the short ones, but do so by a much greater amount with the stem dictionary than the thesaurus dictionary, so that for the long requests normalized recall shows stem to be slightly superior to the thesaurus. This correlation suggests that the generally inferior stem dictionary may be adequate for long requests.

Requests may also be characterized by the frequency of use in the collection of the request concepts. Two methods of obtaining averages for the 35 requests are given in Figure 6, each method supplying an arithmetic mean and a median value. The average frequency per average request concept has been found to be the more satisfactory of the two, and requests are again divided into two sets by this principle in Figure 7. Requests having low frequencies per average concept are seen to perform best, with no real differences between stem and thesaurus dictionaries.

It is to be expected that these three characteristics of generality, length and concept frequency are strongly inter-connected, since specific requests are probably often long ones, and also probably have low average concept frequencies. A visual representation of the correspondence between the three characteristics is given in Figure 8; in Figure 9 it is shown that 19 of the 35 requests fall exactly into the two expected combinations of three characteristics each. These characteristics seem to be the only available objective means of stating whether requests are broad or narrow in a subject field sense; although perfect correspondence is not obtained, there is

DICTIONARY	EVALUATION MEASURE	ALL, 35 REQUESTS 4-18 CONCEPTS, MEAN 8.0	LONG, 17 REQUESTS 8-18 CONCEPTS, MEAN 10.8	SHORT, 18 REQUESTS 4-7 CONCEPTS, MEAN 5.4
Stem	Normed. Recall	.7601	.8363	.6883
	Normed. Precision	.5326	.6191	.4509
Thesaurus 1	Normed. Recall	.8016	.8352	.7699
	Normed. Precision	.6069	.6353	.5801

Performance Results Comparing All the Requests, the Long Requests Only and the Short Requests Only, with Two Dictionaries.

Figure 5.

	MEAN	MEDIAN
Average Total Frequency (i.e. Sum of Concept Frequencies) Per Request	1,542	1,605
Average Concept Frequency Overall Concepts Per Request	220	220

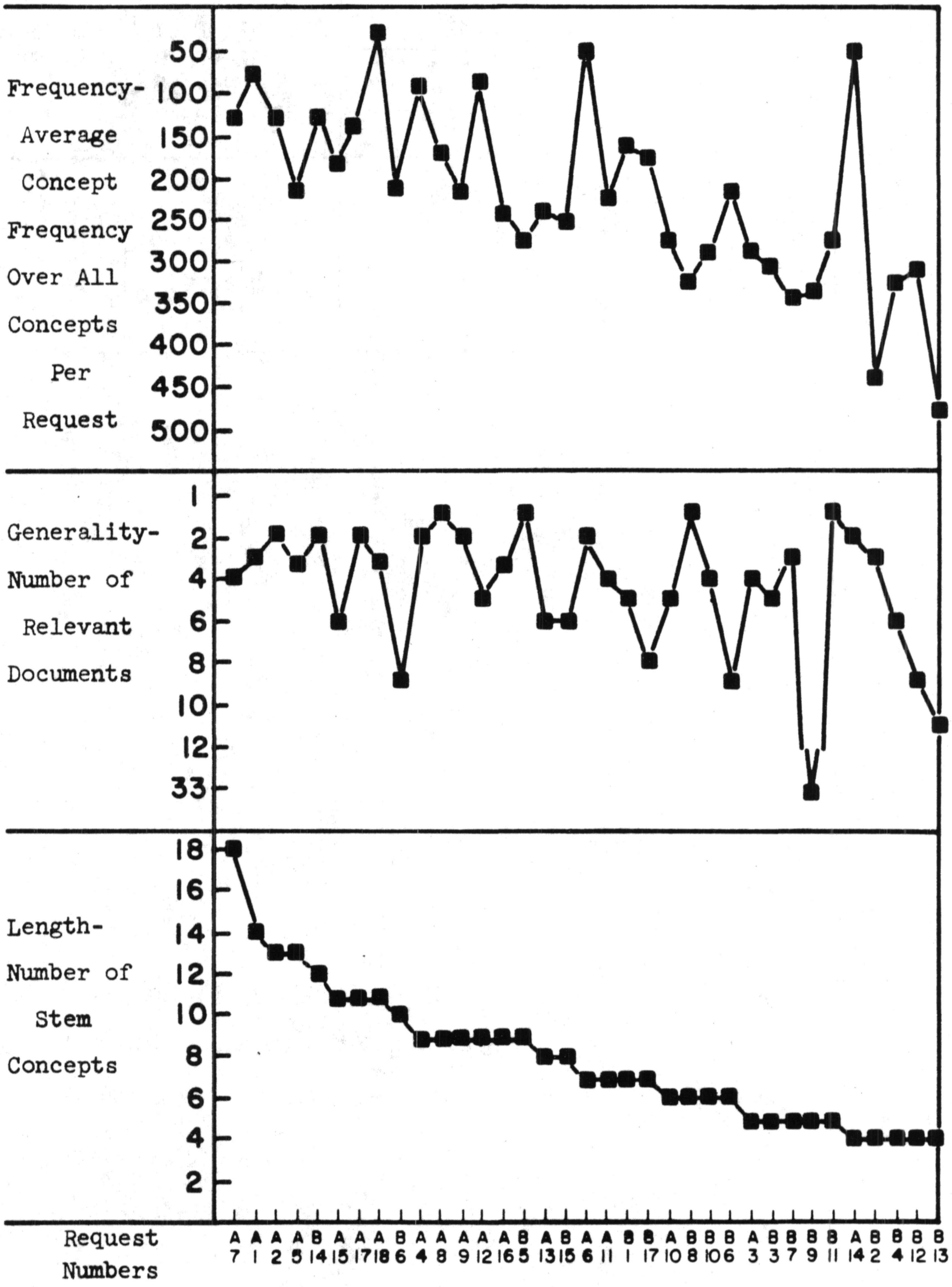
Data Showing Two Methods of Computing the Average Frequency of Use in the Collection of Request Concepts, using the Stem Dictionary.

Figure 6.

DICTIONARY	EVALUATION MEASURE	ALL, 35 REQUESTS FREQ. 26-483 MEAN 220	LOW FREQUENCY, 17 REQUESTS FREQ. 26-218 MEAN 130	HIGH FREQUENCY, 18 REQUESTS FREQ. 223-483 MEAN 305
Stem	Normed. Recall	.7601	.8154	.7080
	Normed. Precision	.5326	.5671	.5000
Thesaurus 1	Normed. Recall	.8016	.8527	.7533
	Normed. Precision	.6069	.6717	.5457

Performance Results Comparing All the Requests, the Requests Having Low Average Frequency Concepts and the Requests Having High Average Frequency Concepts, with Two Dictionaries

Figure 7.



Data for the 35 Individual Requests Showing Visually the Correspondence Between Length, Generality and Average Concept Frequency.

Fig. 8.

		LONG		SHORT		
		LOW FREQ	HIGH FREQ	LOW FREQ	HIGH FREQ	
SPECIFIC	LOW FREQ	9		2		11
	HIGH FREQ		2		4	6
GENERAL	LOW FREQ	4		2		6
	HIGH FREQ		2		10	12
		13	4	4	14	35

Data for the 35 Individual Requests Showing the Numbers of Requests that Fall into the Eight Possible Categories of Length, Generality and Average Concept Frequency

Figure 9.

possibly a strong correlation between these characteristics and real breadth of requests. The superiority of the specific, long and low frequency requests is again seen in performance figures for the 19 requests that exactly fall into the expected combinations as seen in Figure 10. The thesaurus dictionary is seen, as expected, to give the most improvement to the general short and high frequency requests. Further analysis of this type awaits suitable computer programs, since the hand analysis methods used are too time consuming.

C) Comparison of Requests of the Two Preparers

Since two persons were responsible for request preparation, any variation in the measurable characteristics of generality length and frequency already noted may be correlated with the different preparers. Figures 11, 12 and 13 show that a quite strong correlation does exist, since the requests from preparer "A" are on average more specific, longer and hence have lower mean frequencies than requests from preparer "B" (Figure 11). Figure 12 repeats the data of Figure 9, adding the request preparer distinction, and Figure 13 shows that if the eight sets of results in Figure 12 are divided into two sets of four each by the diagonal line in Figure 12, correspondence is quite marked and is probably statistically significant.

The previously examined subject request characteristics such as studies of unclear requests, requests having a multiple need, and requests containing identifiable important words (see part 5D) are almost equally divided among requests of the two preparers; thus, although the requests prepared by person "A" are expected to give the better performance, it is not correct to assume that "A" did a better quality job than "B". The six requests judged difficult for the system (Part 3E) comprise five "A" requests and one "B", but as has been noted only three of these requests

DICTIONARY	EVALUATION MEASURE	SPECIFIC, LONG AND LOW FREQUENCY, 9 REQUESTS	GENERAL, SHORT AND HIGH FREQUENCY, 10 REQUESTS
Stem	Normed. Recall	.8309	.5812
	Normed. Precision	.6085	.4072
Thesaurus 1	Normed. Recall	.8617	.7247
	Normed. Precision	.6702	.5870

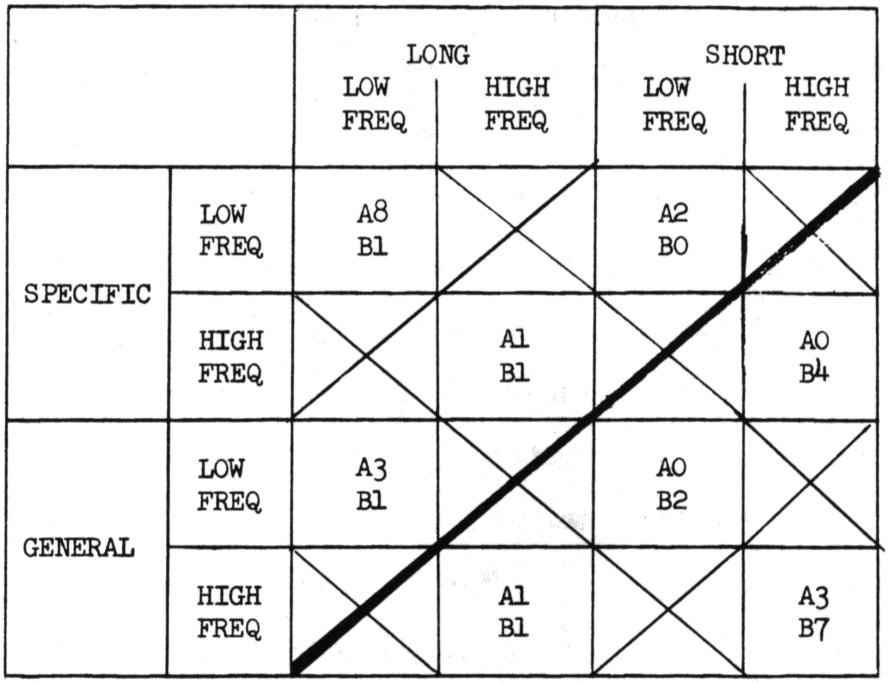
Performance Results Comparing Two Subsets of the Requests Having the Expected Correlations Between Length, Generality and Average Concept Frequency

Figure 10.

GENERALITY, LENGTH AND FREQUENCY	ALL 35 REQUESTS	PREPARER "A", 18 REQUESTS	PREPARER "B", 17 REQUESTS
Generality, Mean Relevant	4.9	3.3	6.5
Length, Mean Concepts	8.0	9.6	6.3
Frequency, Mean Frequency	220	157	287

Comparison of Length, Generality and Average Concept Frequency for all Requests, Requests Made by Preparer "A" and Requests Made by Preparer "B".

Figure 11.



		LONG		SHORT	
		LOW FREQ	HIGH FREQ	LOW FREQ	HIGH FREQ
SPECIFIC	LOW FREQ	A8 B1	☒	A2 B0	☒
	HIGH FREQ	☒	A1 B1	☒	A0 B4
GENERAL	LOW FREQ	A3 B1	☒	A0 B2	☒
	HIGH FREQ	☒	A1 B1	☒	A3 B7

Data for the 35 Individual Requests as Given in Figure 9 but Here Divided into the Requests of the Two Preparers.

Figure 12.

	PREPARER		
	"A"	"B"	
≧2/3 BROAD	14	3	17
≧2/3 NARROW	4	14	18
	18	17	35

Correlation Between Request Preparers and Request Type, Where "Broad" is the Sum of the 4 Categories to the Left of the Diagonal in Figure 12, and "Narrow" is the Sum of the Other 4.

Figure 13.

actually performed poorly (two "A" and one "B").

The performance results in Figure 14 compare the "A" and "B" requests for two dictionary runs each made on two different document lengths, text and abstracts. The expected superiority of the "A" requests is seen in the stem dictionary results, but with the thesaurus, the "B" requests perform slightly better. Since the "B" requests are quite inferior to "A" with the stem dictionary, this does leave a greater opportunity for improvement with the thesaurus; the initial inferiority on stem requires an explanation, however, as well as the fact that the thesaurus does not much improve the "A" requests. It is difficult to isolate any fundamental reasons for this result, because individual problems with both the stem and thesaurus dictionaries seem primarily to be the cause, as the following example shows.

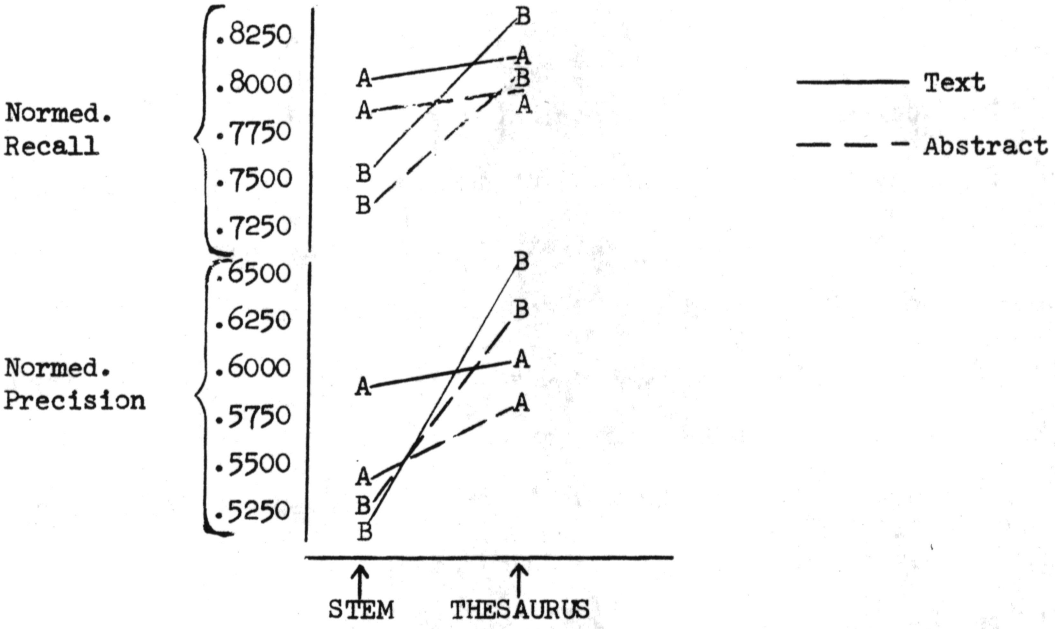
Request B10, with a normalized recall of 0.8205 with thesaurus and 0.3718 with stem, has four documents assessed as relevant, and the thesaurus produces improvements in rank positions of 22, 26, 32, and 60 compared with stem. Reasons for the superiority of thesaurus in this case are:

a) The thesaurus provides additional matching concepts between the request and all four relevant documents, including the important concept "computer". The stem dictionary fails to match this concept, because the suffix routine used does not conflate all word forms, and "computation" is separated from "compute" and "computer".

b) The thesaurus does not contain "system" but regards it as a common word to be ignored, and although the stem dictionary uses it and establishes matches with all four relevant documents, this high frequency word also establishes matches with many non-relevant documents.

c) The very important request concept "chemistry" is grouped with synonyms in the thesaurus which successfully increase the weight

INPUT AND DICTIONARY	EVALUATION MEASURE	"A" REQUESTS	"B" REQUESTS
Text, Stem	Normed. Recall Normed. Precision	.8028 .5917	.7516 .5209
Text, Thesaurus-1	Normed. Recall Normed. Precision	.8124 .6041	.8294 .6519
Abstract, Stem	Normed. Recall Normed. Precision	.7830 .5403	.7358 .5244
Abstract, Thesaurus-1	Normed. Recall Normed. Precision	.8000 .5826	.8033 .6327



Comparison of "A" and "B" Requests for Two Dictionaries and Two Document Lengths.

Figure 14.

given to this concept compared with stem, producing increases of 1 to 18 1/2, 2 to 11 1/2, 3 to 25 and 4 to 11 1/2.

These three reasons for superiority of the thesaurus process are thought to be typical for other requests also. There is also a strong correlation of reasons a) and b) between the "A" and "B" requests, since as Figure 15 shows the concepts "computer" and "system" appear a total 16 times in the "B" requests, and only 6 times in the "A" requests, thus giving the "B" requests greater opportunity to benefit from the superior handling of these concepts in the thesaurus.

This treatment of the different sets of requests only scratches the surface of the problem and points mainly to some of the factors known to be involved.

D) The Recognition of Important Request Words

The presence of quite specific and important single request words and the problem of giving them a weight in proportion to their importance was noted in part 3B. In order to discover whether increases in the weight of such important words does improve retrieval performance, the 35 requests were examined (without knowledge of search results, relevant documents or concept frequency) to see whether such important concepts could easily be identified. Seventeen requests were found to possess such important concepts, and each of the concepts was tripled in weight. These decisions are recorded in Appendix B. This simulates a quite feasible requestor rule which would ask for any important concepts to be underlined in the request statement, and which the system could recognize and correspondingly increase in weight by some factor. Six requests in addition to the seventeen were also slightly modified, request A1 was divided into two as suggested by part 3C; in request

TERM (STEM)	REQUESTS OF PREPARER "A"	"B"
COMPUTER	2	5
STEM	4	11

Occurrence of Two Particular Words in the Requests.

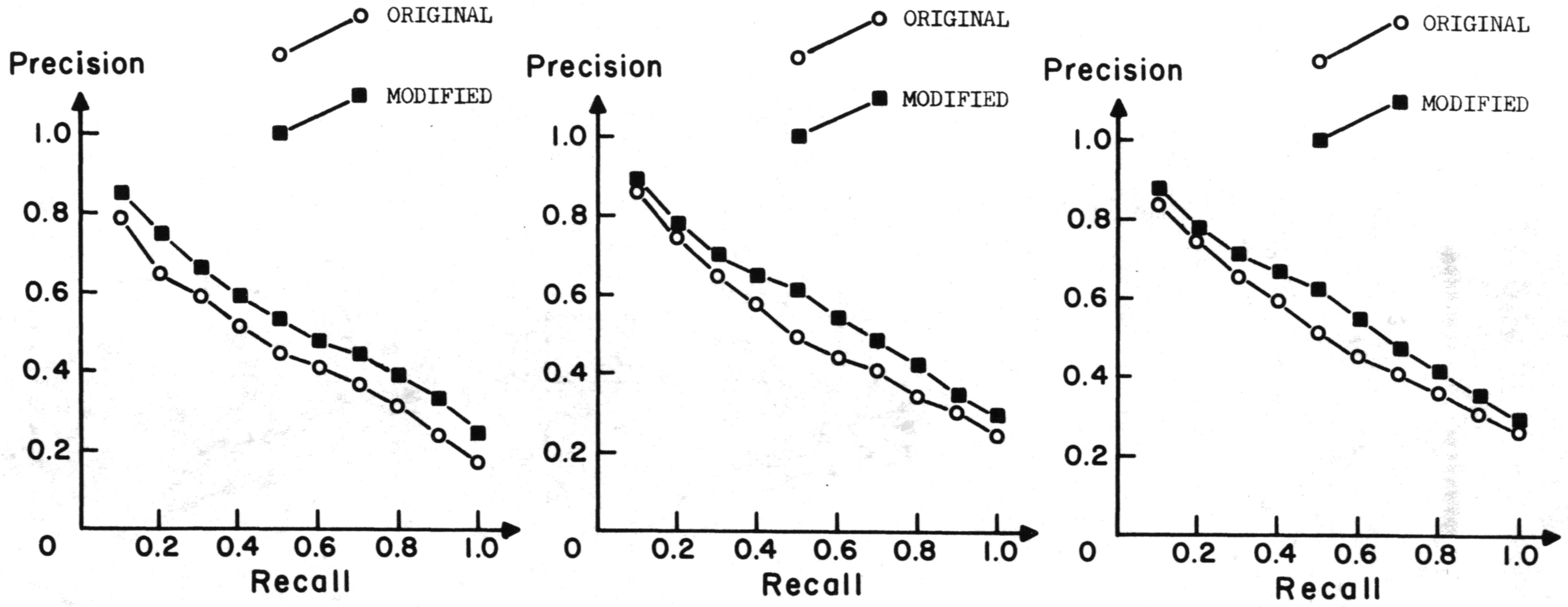
Figure 15.

A2 the negative statement was removed; in requests A8 and B11 the difficulties caused by common words used in a technical sense prompted selection of one or two synonyms for the given words; and in two requests keypunching errors which preserved hyphenated words were corrected. These six modifications are all thought to represent reasonable demands that would be made to users of an operational system.

These 24 requests are now processed together with the 12 requests for which no modification was made; they are described as "Hand Modified"; a total of 36 results because request A1 is split into two. Comparison of retrieval performance of the modified with the original unmodified requests is made for six retrieval runs in Figures 16 and 17. All precision recall curves for the hand modified requests show them to be superior over the whole performance range, with increases in precision at most recall values of more than 5%, and in the middle recall ranges of nearly 10%.

Using the Abstract thesaurus result for analysis, the six requests that were quite severely modified did not perform very well, only B11 was notably improved, and some of the others received a worse performance. Of the seventeen requests that had triply weighted important words, ten were improved, five has a worse performance, and two remained unaffected. Four of the ten that were improved are shown in Figure 18, and the two that were worsened by the greatest amounts are given in Figure 19, with rank positions for all the relevant and normalized measures.

It is of interest to note that at present these hand modifications do produce a superior result to the relevance feedback process described elsewhere [3]. Figure 20 includes a comparison, using an evaluation technique that differs from the plots in Figures 16 and 17 in order to achieve a fully user-oriented evaluation [4,5]. Further work on relevance feedback



a) Full Text, Stem

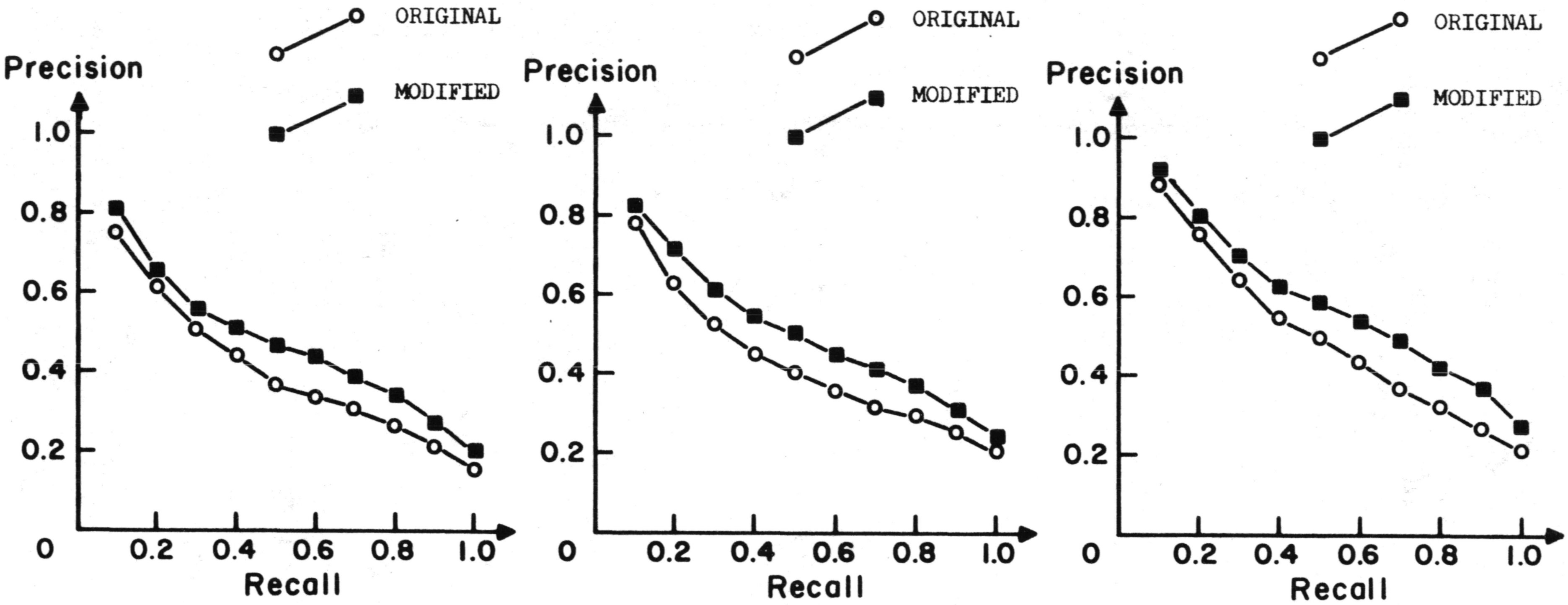
b) Full Text, Thesaurus

c) Full Text, Thesaurus
and Phrases

ADI Documentation Collection, 82 Documents, Averages over 35 Requests.

Performance Comparisons Between Requests as Received (Original) and Manual Modification by Important "Key" Term Weight Increases.

Fig. 16.



a) Abstracts, Suffix 's'

b) Abstracts, Stem

c) Abstracts, Thesaurus

ADI Documentation Collection, 82 Documents, Averages over 35 Requests

Performance Comparisons Between Requests as Received (Original) and Manual Modification by Important "Key" Term Weight Increases.

Fig. 17

REQUEST	RELEVANT DOCUMENT NUMBER	RANKS OF RELEVANT AND NORMALIZED MEASURES	
		"ORIGINAL"	"MODIFIED"
Request A9		33	1
(2 Relevant)	82	50	27
	50	NR .5000	NR .8438
		NP .1718	NP .6780
Request B2		5	1
	80	17	2
(3 Relevant)	65	42	42
	27	NR .7553	NR .8354
		NP .4392	NP .7683
Request B10		1	2
	09	2	1
	48	19	3
(4 Relevant)	69	27	4
	70	NR .8750	NR 1.0000
		NP .7387	NP 1.0000
Request B11		21	1
(1 Relevant)	36	NR .7531	NP 1.0000
		NP .3091	NP 1.0000

Results of Four Requests in Which Manual Modification by Important Term Weighting is better than the Original Unmodified Search.

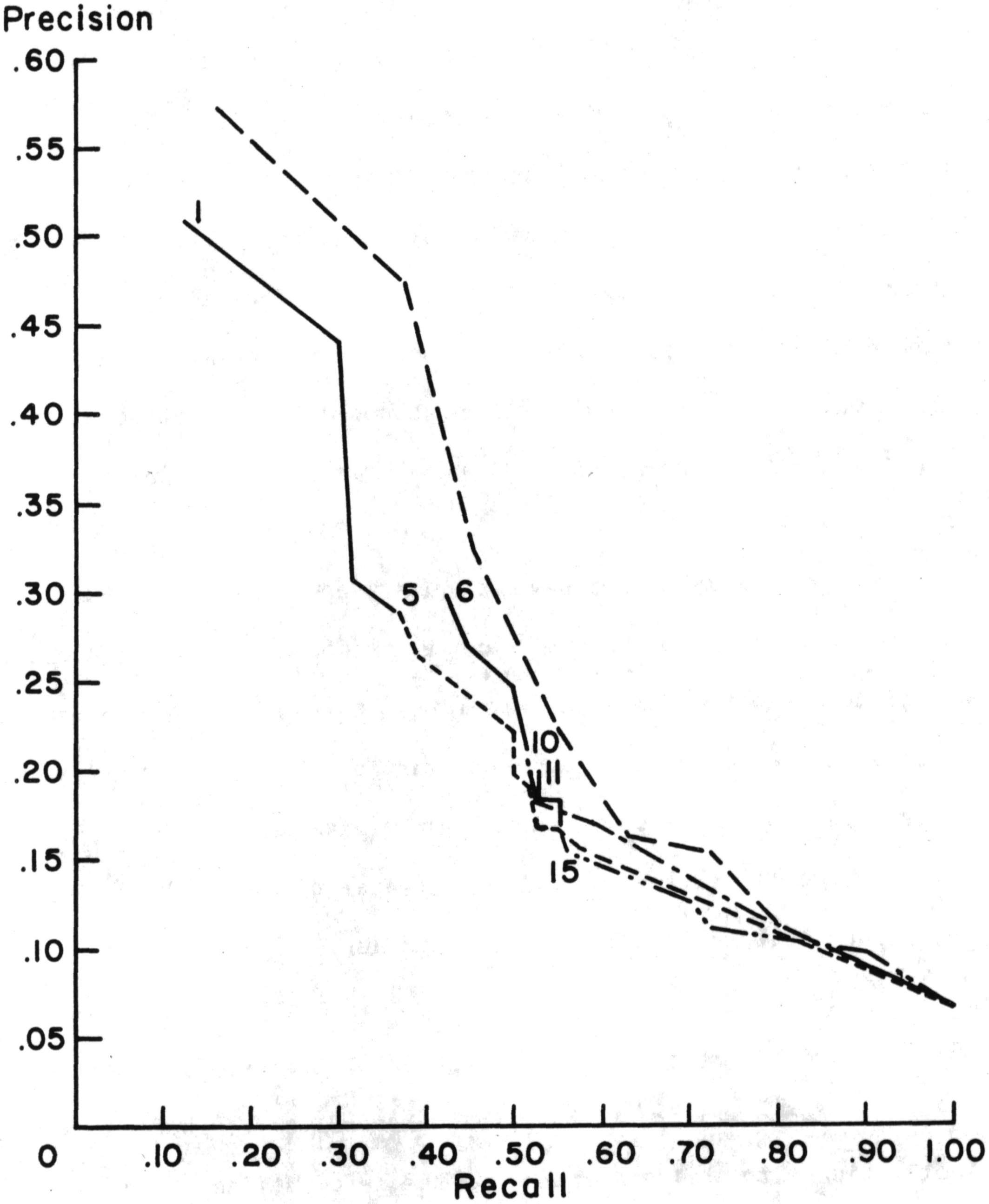
Figure 18.

REQUEST	RELEVANT DOCUMENT NUMBER	RANKS OF RELEVANT AND NORMALIZED MEASURES	
		"ORIGINAL"	"MODIFIED"
<u>Request A4</u>		1	1
(2 Relevant)	29	30	43
	63	NR .8250	NR .7437
		NP .6660	NP .6216
<u>Request A7</u>		1	2
	19	6	14
	07	8	16
(4 Relevant)	09	25	19
	40	NR .9038	NR .8686
		NP .7279	NP .5916

Results of Two Requests in Which Manual Modification by Important Term Weighting is worse than the Original Unmodified Search.

Figure 19.

	1 5	Initial Search
Relevance Feedback	6 10	First Iteration, Five New Documents Examined
	11 15	Second Iteration, Five New Documents Examined



ADI Abstracts Thesaurus, "Pseudo-Cranfield" Cut-off, Macro Evaluation Over 35 Requests.

Comparison of Performance Effectiveness of Relevance Feedback with Manual Modification by Important Term Weight Increases

Fig. 20

is expected to result in improvements, but for high precision requirements hand modified requests may always be superior (at the cost of increased user effort).

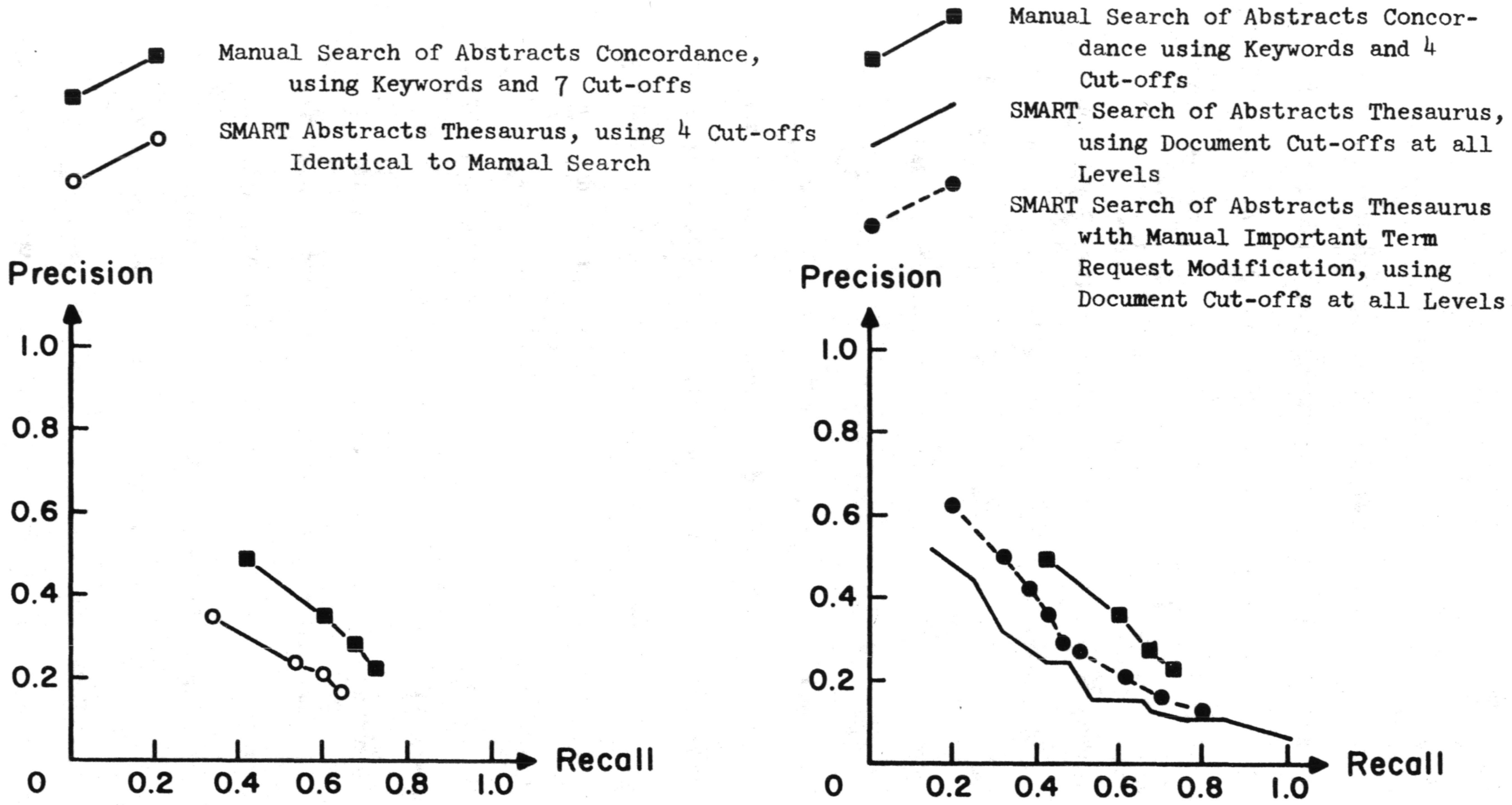
The treatment of important request words might be made more drastic, for example by the use of an "essential" word rule, which would only present to the searcher documents that contain the noted important words. This strategy could in fact be achieved by assigning very high weights to the important concepts (weights of several hundred would be needed in text runs); alternatively, modifications could be made to the search algorithm. It is almost certain that this procedure would imply that some relevant documents would never be found, although large increases in precision might be possible.

For example, for seven requests containing important concepts chosen at random, only 9% of the relevant items would be lost, and although actual precision results cannot be calculated, of the 86 non-relevant documents that were given rank positions above 16 in the output, 32 would be excluded by this rule. Other requests subsequently examined occasionally produce a much greater recall ceiling, and also a greater precision improvement, so that this procedure is worth further experimentation.

6. Performance Effectiveness and Search Procedures

A comparison of the retrieval results obtained in the documentation collection with the performance of the aerodynamics and computer science collection shows the documentation results to be quite inferior. Data concerning this fact were given in Section I, where it was seen that with a stem dictionary in use, at 0.80 recall, the amount of non-relevant examined as a proportion of the total non-relevant in the collection is 0.45 with documentation, 0.23 with aerodynamics and 0.17 with computer science. At the other end of the scale, examining the output until 0.03 of the non-relevant is encountered gives recall values of .16 with documentation, .43 with aerodynamics and .51 with computer science. Factors in the text environments differ between the collections, and matters such as the quality of terminology in the subject language, as well as the testing of techniques used for collection gathering, request preparation and relevance decisions all contribute to the differences observed in unknown proportions.

It seems likely that the imprecise terminology encountered in documenation which appears in both the documents and requests is a major cause of the poor performance, and in order to overcome these problems extra human intellect may be needed in the system. It may not be possible to build synonym dictionaries that will entirely provide for this, but good dictionaries together with a good choice of search strategy is likely to improve performance considerably. Some proof of the value of carefully chosen search words is given in Figure 21, where a hand search of the KWIC type concordance to the abstracts is compared with a SMART abstracts Thesaurus result. The hand searcher chose up to five keywords for each request and was allowed to use any words that might be considered useful as suggested by the request statement. A comparison with SMART in Figure 21 a) is made after fitting the SMART results to the hand searches by making cut-offs in the SMART ranked output in such a way that the number of documents retrieved for each request are identical to the hand searched results, Figure 21 b) shows two fuller SMART curves obtained by making a series of cut-offs after one document, two documents, and so on, up to the last document in the collection. It is not surprising that the hand searches work better, since the free choice of



(a)

(b)

ADI Abstracts, Macro Evaluation over 35 Requests.

Comparisons of a Hand Search in the KWIC Concordance with

SMART Searches Using Abstracts Thesaurus

Fig. 21.

synonyms made on an individual request basis should work better than the obligating use of the pre-constructed set of synonyms contained in the thesaurus. The hand searches do not use any feedback to obtain a fair comparison, since the search keywords were chosen before any reference was made to the KWIC concordance. The result of SMART using the hand modified "important" concept with increased weights is included in Figure 21 b); the curve now lies much closer to the hand result. Naturally a hand system permitting coordinate keyword searches would extend the hand curve to higher precision values, and choices of more than five keywords per request would enable higher recall ratios to be reached.

This result does not condemn the automatic indexing procedures, because hand searches are less easy to conduct in the sort of situation in which SMART would operate, such as a large file of individually long document surrogates. It is clear also that in an operational use of SMART; search strategies employing several dictionaries in a variety of possible ways could be used, and for users willing to employ some intellect to strongly interact with the system, quite large performance gains may be expected. Ways in which a system might operate are: the use of several dictionaries successively, until the required performance is reached; the use of several dictionaries with "merged" output results [6]; use of a manual or automatic method of making an accurate pre-search best dictionary choice, yet to be developed; use of dictionary display methods to allow users willing to strongly interact to delete or add synonyms; and the use of relevance feedback methods to iterate searches and improve performance.

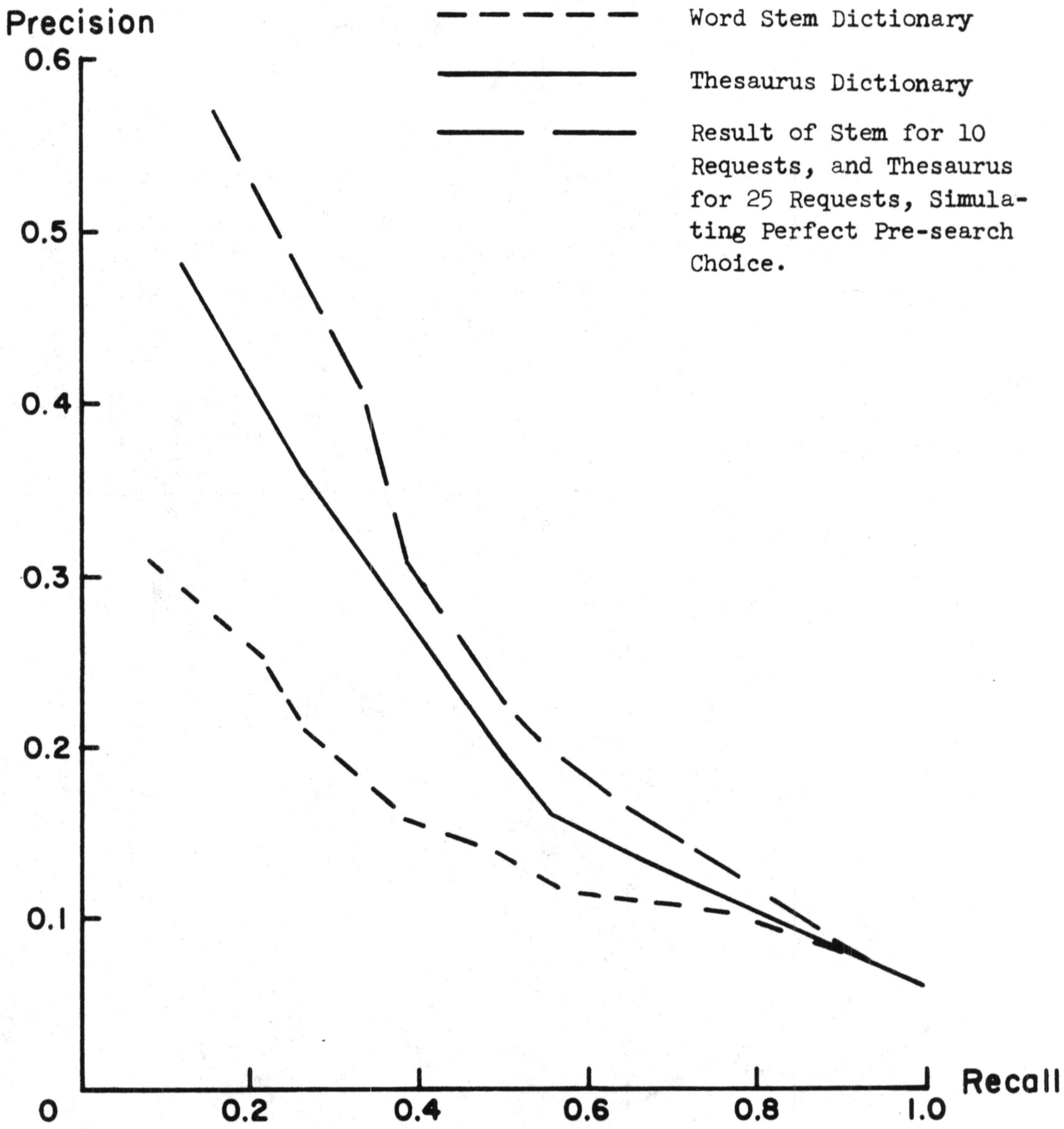
Of these suggestions the idea of making a pre-search dictionary choice has been explored but with no success so far. If, for example, long requests work better with the stem dictionary, and short requests need the thesaurus, then a simple automatic choice could be made by the system. Using the Abstract Stem versus Abstract Thesaurus result, Figure 22 shows that more of the factors of generality, length or concept frequency are correlated with either dictionary, and even the requests made up by the two preparers do not markedly prefer particular dictionaries. Other criteria may be discovered to aid such a pre-search choice, and if a perfect choice were achieved the result would be as given in Figure 23, where the curve based on choice of the best dictionary is seen to be better than use of either dictionary exclusively.

The possibility of achieving a satisfactory automatic subject recognition is considered in an extensive analysis performed by J. O'Connor [7]. It seems certain that some loss of performance due to inability to correctly recognize and match with ideas asked for in requests will be experienced unless very sophisticated procedures can be developed. However, failure in matching occurs also in manual systems due both to errors and inability to cope with the tasks of manual indexing and vocabulary control; it is thus by no means certain that automatic systems will in practice prove inferior.

	GENERALITY		LENGTH		FREQUENCY		PREPARER	
	SPECIFIC	GENERAL	LONG	SHORT	LOW FREQ.	HIGH FREQ.	"A"	"B"
Stem, 10 Requests	4	6	6	4	4	6	6	4
Thesaurus, 25 Requests	13	12	11	14	13	12	12	13

Attempt to Use Factors of Generality, Length and Frequency to Make a Pre-search Dictionary Choice Between Stem and Thesaurus, Abstract Results, Showing Numbers of the 35 Requests that Fall into Each Category.

Figure 22.



ADI Abstracts, "Pseudo-Cranfield" Cut-off, Macro Evaluation of 35 Requests.

Comparison of Perfect Pre-search Choice of Two Dictionaries with each Dictionary Used Exclusively.

Fig. 23.

References

[1] J. O'Connor, Relevance Disagreements and Unclear Request Forms, Draft Report, 1966.

[2] G. Salton, The Evaluation of Automatic Retrieval Procedures — Selected Test Results using the SMART System, American Documentation, Vol. 16, July 1965.

[3] J. J. Rocchio and G. Salton, Information Search Optimization and Iterative Retrieval Techniques. In Proc. of the 1965 Fall Joint Computer Conference, Washington D.C., Spartan Books, 1965.

[4] E. M. Keen, User Controlled Search Strategies. To be published in Proc. of the Fourth Annual National Colloquium on Information Retrieval, 1967.

[5] H. A. Hall and N. H. Weiderman, The Evaluation Problem in Relevance Feedback Systems, Report ISR-12 to the National Science Foundation, Section XII, Computer Science Department, Cornell University, June 1967.

[6] J. Rocchio, Combinations of Analysis Methods — The Merged Output Results, Report ISR-9 to the National Science Foundation, Section XIX, Harvard Computation Laboratory, August 1965.

[7] J. O'Connor, Automatic Subject Recognition in Scientific Papers: An Empirical Study. Journal of the ACM, Vol. 12, October 1965.