I. Test Environment

E. M. Keen

1. Introduction

The SMART experiments described in this report are conducted in a controlled laboratory environment. Each experiment uses a fixed document collection together with a set of search requests and relevance decisions. Factors involved in the input, analysis and search procedures may then be varied, and results which are usually based on performance measures are obtained. This section describes and contrasts the several test environments in use, and introduces the test experiments that are described in subsequent sections of this report.

2. Document Collections and Search Requests

Three different document collections are currently in use, and a general description appears in [1]. For convenience, the main properties of the document collections and search requests are given in Fig. 1. The IRE-3 collection is an amalgamation of the 405 document IRE-1 and 375 document IRE-2 collections previously used for the first experiments with SMART. Documents in this collection cover most of the subjects in the area of computer science that were current during 1958-1962, and the abstracts were commercially prepared in order to provide a quarterly current literature information service, published in the "IRE Transactions on Electronic Computers".

The CRAN-1 collection is part of the collection used in the second Aslib Cranfield Project. The original documents are research reports and

journal articles of a technical rather than popular nature, centering on high speed aerodynamics, and including a small number of documents on internal aerodynamics (flow in compressors, for example) and atmospheric re-entry. Most of the abstracts used are those that were written by the authors of the papers; a small number of documents which had no author abstract use a commercially published abstract. No documents in the collection were published later than 1962, and most papers fall between 1954 and 1963.

The ADI collection comprises some of the "short papers" presented at the American Documentation Institute's Annual Meeting in 1963. Although most aspects of documentation are covered, the theme of the meeting was "Automation and Scientific Communication", and thus the collection emphasizes the research and mechanized rather than the operational and manual side of documentation.

Fig. 1 gives counts of the average lengths of the documents, Cran-1 being the longest, IRE-3 next, and ADI the shortest, except that in the case of ADI the full texts of the short papers are also available. The indexing that is available for the Cran-1 collection is about half the length of the abstracts. Fig. 1 also gives similar data for the search requests, showing that the IRE-3 requests are largest, followed by Cran-1 and then ADI.

The methods used for obtaining requests are briefly summarized in Fig. 1. The Cranfield requests were obtained from authors of research papers in aerodynamics and the requests cover the stated problems that the authors were investigating, which finally led to their research paper. Authors often supplied more than one request; 29 authors in all were

Characteristics			IRE-3	CRAN-1	ADI
DOCUMENT COLLECTIONS	Subject Area		Computer Science	Aero-dynamics	Documen-tation
	Number of documents in collection		780	200	82
	Average number of word occurrences (all words) per document	full text	-	-	1380
		abstract	88	165	59
		title	9	14	10
		indexing	-	33	-
	Average number of word occurrences (non common words deleted) per document	full text	-	-	710
		abstract	49	91	35
		title	5	11	7
		indexing	-	0	-
	Average number of distinct words per document, using suffix 's' dictionary	full text	-	-	369
		abstract	40	65	25
		title	5	9	6
		indexing	-	33	-
SEARCH REQUESTS	Number of search requests		34	42	35
	Average number of word occurrences (all words) per request		20*	17	14
	Average number of distinct words per request, suffix 's' dictionary		12*	8	8
	Request Preparation				
	a) Prepared by subject experts in course of their work			√ (42)	
	b) Prepared by staff members		√ (17)		
	c) Prepared by non-staff members with no knowledge of system and some familiarity with subject		√ (17)		√ (35)

* 17 requests prepared by staff members have average length of 24 words, and 14 suffix 's' concepts; 17 requests prepared by non-staff persons have average length of 16 words and 11 suffix 's' concepts.

Document Collection and Request Characteristics

Fig. 1

responsible for the 42 requests in use with the Cran-1 collection. The document collection consists not of the authors' own research papers, but of a number of the earlier papers that the author cited in the bibliography to his paper.

The first set of seventeen computer science requests were prepared for the IRE-1 collection by three project staff members. Two persons made up thirteen of the requests without any knowledge of how the system would perform in practice, but having extensive knowledge of the technique of operation of the system. Requests were devised to cover a cross-section of the major topics in the collection but were not "source document" requests in the sense that they were based on particular documents. A third person made up four of the requests using sets of classification headings that had been manually assigned to the IRE-1 collection. A second set of seventeen requests was prepared by one person hired for the task. This person was one of two persons who also prepared the 35 documentation requests. Requests were again not based on any one document in the collection. No guidance concerning the length of the requests was given — the hired persons tended to devise requests that were rarely longer than a sentence or two, whereas requests prepared by staff members were often longer.

Techniques used for obtaining relevance decisions are described in part 3. A much more detailed analysis of the characteristics of documents and requests is possible. Such an investigation is given in Section X, part 3, for the ADI documentation requests only. Direct performance comparisons between collections have not formed the purpose

of any major SMART experiment so far, but an attempt to make such a comparison is contained in part 6 of this section. The degree to which the documents and requests used in this laboratory environment may be regarded as typical of larger sized real-life situations is not known. What is certain, however, is that all the document collections are almost certainly contained in part if not in whole within actual collections being used, and there is nothing in the stated requests that suggests that they could not be posed in real-life situations.

Further collections and requests have been obtained for the purpose of making additional tests, as outlined in [2]. Fig. 2 supplied some tentative data on four new test environments that are currently under investigation.

3. Relevance Decisions

Data on the number of documents assessed as relevant is given in Fig. 3. The division into specific and general requests is made by dividing each request set into two equal or nearly equal sets according to the number of documents assessed as relevant. This method is therefore highly dependent on the characteristics of the test environment, but it permits a comparison of requests of differing generality, see part 6B.

The data in Fig. 4 shows the extent to which the requests cover the topic areas of the total collection. Between 55% and 88% of the documents in the collections are relevant to one or more of the requests; it may thus be assumed that most of the major collection topic areas are covered by one or more requests.

The techniques used for obtaining the relevance decisions are given

CHARACTERISTICS	CRAN-2	ISPRA/EURATOM	MEDLARS	TIME MAGAZINE
Subject area	Aerodynamics and Aircraft Structures	Documentation	Medicine	Section on world news
Number of documents	1,400	1,268	276	c600
Document length	Abstracts and indexing	Abstracts	Abstracts	Full text
Number of requests	225	48	18	c50

Characteristics of Four New Test Environments

Fig. 2

CHARACTERISTICS	IRE-3	CRAN-1	ADI
Number of requests, (all requests)	34	42	35
Number of relevant documents:			
a) grand total	592	198	170
b) range	2-65	1-12	1-33
c) mean, per request	17.4	4.7	4.9
d) median, per request	16	4	3
e) "generality number"*	22.2	23.6	59.2
Number of specific requests	17	21	17
Number of relevant documents:			
a) grand total	128	63	36
b) range	2-16	1-4	1-3
c) mean, per request	7.5	3.0	2.1
d) median, per request	7	3	2
e) "generality number"*	9.7	15.0	25.8
Number of general requests	17	21	18
Number of relevant documents:			
a) grand total	464	135	134
b) range	17-65	5-12	4-33
c) mean, per request	25.8	6.4	7.4
d) median, per request	20	6	6
e) "generality number"*	35.0	32.1	90.8

* Generality Number = $\frac{\text{Total Relevant Documents x 1000}}{\text{Total Documents in Collection}}$

Data on Documents Assessed as Relevant to the Requests

Fig. 3

Collection and Size	Number of Documents Relevant to one or more Requests	% of Total Collection Relevant to one or more Requests	NUMBER OF DOCUMENTS RELEVANT TO:						
			1 REQ.	2 REQ.	3 REQ.	4 REQ.	5 REQ.	6 REQ.	7 REQ.
IRE-3, 780	430	55.1%	287	120	21	2	-	-	-
CRAN-1, 200	153	76.5%	115	33	5	-	-	-	-
ADI, 82	72	87.8%	19	23	19	9	1	-	1

Number of Documents in the Collection Assessed as Relevant

Fig. 4

in Fig. 5. In virtually every case, the entire collections have been examined for relevance in relation to every search request. The only exceptions to this are four requests used in IRE-3 that were based on the classification headings. For these requests, those documents in the IRE-1 part of the collection originally classified under the given headings were taken to be relevant, and no other documents in the collection were examined.

In every case the request preparer made the relevance decision, and in a few cases, a consensus of opinion was used for cases of doubt for one or another of the seventeen staff prepared IRE-3 requests. Doubt in relevance decisions was usually settled by accepting the document as relevant. Dichotomous decisions only were made for the IRE-3 and ADI requests: a document was regarded either as relevant or non-relevant with no grades of relevance allowed. In the Cran-1 case, a scale of four degrees of relevance was used for the relevance judgments. In the experiments conducted so far with the SMART system, all four degrees of relevance were regarded as equally relevant. A small hand-calculated set of results taking into account these available relevance grades is presented in part 5 of this section.

Relevance decisions in the IRE-3 collection were always made by examining the document abstracts and never the full texts. This may be regarded as a weakness of this environment. A detailed examination of the relevance decisions for the ADI set is made in Section X, part 4. Whether the prepared requests and relevance decisions of the IRE-3 and ADI collections and even the author supplied data in the Cran-1 collection are typical of real-life situations is a disputed question. So far, no evidence has been produced to invalidate the methods used. Examination of relevance decisions on the three collections leaves the impression that the Cran-1 requests, which come closest

RELEVANCE DECISIONS	IRE-3	CRAN-1	ADI
Scale of five relevance grades, done by request preparers (subject experts), based mainly on full documents, by examination of entire collection.		√ (42)	
Dichotomous, done by request preparers (staff members), based on abstracts, by examination of entire collection.	√ (17)		
Dichotomous, done by request preparers (non-staff members), based on abstracts (IRE-3) and full text (ADI), by examination of entire collection.	√ (17)		√ (35)

Techniques Used for Obtaining Relevance Decisions

Fig. 5

to real user ones, contain more instances of relevance decisions that might be disputed than the other collections. It is suggested that real users tend to have less clearly defined requests in mind, and tend also to judge relevance by means of requirements that they fail explicitly to state in the request. The validity of prepared requests and relevance decisions for experimental testing is frequently challenged by opinion, but a controlled experiment that will show the differences (if any) for test purposes between prepared and real requests is still not at hand. Studies of agreement between different judges carrying out an identical relevance decision task have shown that poor agreement frequently results. But a more important question for experimental tests is whether differences in relevance decision actually alter comparative test results; that is, does option one perform better than option two both when person A does the relevance decision, and person B, and also when relevance decisions of both persons, or those common to both are used? A new documentation collection known as the ISPRA/Euratom collection is being used to test just this problem; test results will appear in a future report in this series.

4. Text Experiments

A) Experimental Procedures

The laboratory environments that have been described permit controlled tests by means of simulated searches. The operation of a retrieval system may be separated into three stages: input of the documents and requests to the system, procedures of content analysis applied to documents and requests, and the matching of the requests with the documents which is the output stage.

The test procedure that is followed requires that no more than one single describable change be made to these procedures at any one time, so that search results may be obtained each time one system component is altered. In this way a series of comparisons based on differences in document input may be made, and then perhaps a second series which compares different content analysis procedures. The primary use of the different test environments is to find out whether a conclusion drawn from an experiment in one environment also holds for another. Thus, if a given content analysis procedure is found to be very effective for the computer science collection, a parallel test of content analysis procedures can then be made with the aerodynamics and documentation collections.

Conclusions about the effectiveness of search results and system performance generally can be made from different viewpoints using several criteria [3]. For various reasons given in [3], the measurement of retrieval performance oriented towards user satisfaction predominates in the current SMART text experiments, and a discussion of methods and measures used is to be found in Section II of this report. A step recently added to the evaluation procedures is that of making statistical significance tests of the results, as described in references [4] and [1]. A further step in the evaluation, which is not formally built into the system requires a hand analysis of the search results involving an examination of individual requests rather than the use only of the averages for a set of requests. A fast-search analysis of specific instances of poor retrieval, for example, is necessary in order to make improvements to the system and to identify areas in which further

work is needed. An analysis of every instance of failure for every request in each experiment would be an impossibly large task; a judicious selection must therefore be made. Most of the sections in this report set out first to present the average results for a series of experiments, and then to make a fast-search performance analysis to uncover details and explanations for the search results obtained.

Since real user populations and currently growing collections are not available, it is correct to describe the experimental procedures used as "Simulated Search Methods: as does R. V. Katter in [5]. Katter criticizes such experimental techniques on several grounds: in particular, he says that mechanical type matching is unnecessary and cumbersome. Since the work reported by Katter does not tackle any problem other than human judgment reliability, his comments do not seem to apply to experimentation that deals with a total system, which are designed to evaluate performance from a user viewpoint. Search procedures used by SMART are not cumbersome, and simulated searches are believed to be necessary in order to provide useful relationships to reality.

B) Variables Tested

At the input stage, the use of natural language by SMART implies that there are not input variables to be tested, since full text processing of documents has not been attempted in many different subject areas. Different lengths of documents are therefore used, such as titles only, or abstracts. Some tests using variables of this type are covered in Section V.

Content analysis procedures in SMART are performed by using a series of dictionaries which differ in construction and effectiveness. The

following types of dictionaries have been tested in retrieval runs:

1. Suffix 's' only, in which request and document words are matched as they stand, with only the terminal 's' denoting plurals being removed. See Section VI.

2. Stems (Null dictionary), in which matching is based on word stems as identified by an automatic suffix removal procedure. See Section VI.

3. Thesaurus, where words (mainly stems) are grouped together on the basis of synonymy, or partial synonymy, using human judgment normally. See Section VII.

4. Statistical association (Concon), where synonyms or related words are identified automatically by using cooccurrence frequency of words in the collection. Apart from the control parameters which may be varied, no human judgment is used. See Section IX.

5. Hierarchies, where subject notions are arranged in a series of subordinate relations, such as genera and species, whole and part. Hierarchies tested so far use thesaurus groups, and texts include some of the many possible strategies of using hierarchies such as going "up" in the hierarchy to parents, or going "down" to sons. See Section VII.

6. Phrases, in which recognition of pairs and larger sets of words is achieved. Phrases are used in conjunction with thesaurus groups, and phrase recognition takes place when words from the required thesaurus groups occur within one sentence of the document or request. See Section VII.

7. Syntax, in which a syntactic analyzer is used to ensure acceptable grammatical relations between the component words of the phrases. The only retrieval results available have appeared in [6].

Although many versions of dictionaries of these types have been tested on the different collections with their differing subject areas, these seven general types describe all the kinds of content analysis procedures that have been tried at the time of this writing. Some of the descriptions applied to content analysis procedures by the Cranfield Project are introduced in part 4C for purposes of comparison. One further optional part of content analysis is the use of weighted rather than binary concept identifiers for the documents and requests; a description of this process appears in Section III.

The search stage requires some procedures for establishing a coefficient to reflect the match between requests and documents. This is then used in SMART to order the search output thus producing a ranked list arranged in decreasing correlation order. Such matching functions are discussed in Sections III and IV.

The main input, analysis and search variables are repeated, for convenience in Fig. 6. It can be seen that each experimental run must be described in terms of four variables: indications of document length and dictionary type are given with each search result, but use of the numeric vectors weighting scheme and the cosine matching function is always made unless otherwise indicated. Since several versions of some dictionaries are available and some additional variables not listed in Fig. 6 have also been investigated, many hundreds of runs can be made before all possible

Stages	Variables		Section
Input	Document Length		
		Text	V
		Abstracts	V
		Titles	V
		Indexing	V
Content Analysis	Dictionaries		
		Suffix 's'	VI
		Stem	VI
		Thesaurus	VII
		Hierarchy	VII
		Phrases	VII
		Syntax	-
		Concon	IX
	Weighted Concepts		
		Numeric Vectors	III
		Logical Vectors	III
Search	Matching Functions		
		Cosine Correlation	III
		Overlap Correlation	III
		Other Correlations	IV

The Input, Analysis and Search Variables Tested, including the Section Number in the present Report

Fig. 6

combinations of the variables are included. A selection from the total number of possible combinations has, in fact, been made, and over 70 sets of results have been obtained so far. In addition to the presentation of the results made in the various sections listed in Fig. 6, tables giving all the performance results appear in Appendix A.

C) Vocabularies and Index Language Devices

Because of the similarity of experimental procedures and continuing cooperation with the Cranfield Project, the relationship between the dictionaries used by SMART and the distinctions about vocabularies and index language devices drawn at Cranfield is briefly discussed.

The dictionaries which have been described include the allowable content identifiers, and they become the language of the system, that is, the language used to represent stored documents and search requests. At Cranfield each dictionary constitutes a different index language, and further distinctions are made between, on the one hand, the different vocabularies of terms in which an index language may be operationally used (the index, lead-in, and code terms), and on the other hand, the fact that every index language is made up of one or more recall and precision devices which control the specificity of the index language. [7,8]. In SMART there is no distinction made or necessary between the possible different vocabularies, since code and index terms are always identical, and lead-in type terms are automatically a part of the dictionaries used.

The devices used in the construction of the SMART dictionaries can be identified according to their recall or precision effects, with the recall devices broadening class definition, and the precision devices narrowing

class definition. With the suffix 's' dictionary used as a starting point, the stems, thesaurus, hierarchy, or concon (statistical association) all constitute the recall devices, because in each case the suffix 's' content identifiers are replaced by concepts representing a whole grouping of words according to the principle used by the dictionary concerned. The use of phrases, syntax and the weighted concept identifiers (numeric vectors) are all examples of the use of precision devices, as well as the major device of coordination which is used in every SMART search, since all the matching functions make use in some way of the number of request terms that match with those in the documents.

Although the recall and precision devices are clearly used in the construction of the dictionaries as described, the use of the dictionaries in SMART does not always produce the expected effect. This is because the processing techniques possible with automatic systems can modify or even change the effect of these devices, and it is possible to use a dictionary which has been constructed on the principle of a recall device, in such a way that the result in the search becomes an increase in precision.

An example of this is provided by the work on statistical association at the Cambridge Language Research Unit (England), where in one test of their clumping procedure the clumps were seen to be acting purely as precision devices and not as recall devices at all [9,10]. This occurred simply because the clumps were used as a weighting device to reinforce certain of the concept matches that already existed without the clumps. Since in SMART concon, hierarchy and phrases are normally used to <u>add</u> concept numbers to the documents and requests,

these additional concept numbers sometimes have the effect of a recall device, and sometimes that of a precision device. Fig. 7 gives an example of the use of concon, where the first order matches of type A, and all second order matches, are recall devices because four additional matches between request and document are made possible by the use of concon. The one first order match of type B acts as a precision device because "cylinder" was already matched in request and document, but the concon pair "Unit → cylinder" gives added weight to "cylinder" in the document, and this in turn gives greater prominence to the match on "cylinder" in the final request/document matching correlation.

The case of the phrase dictionary is also complex. Phrases normally act as a precision device of coordination, when the phrases recognized are used to completely replace all occurrences of the component concepts which do not meet the phrase criteria individually. When phrase identifiers are merely added to existing identifiers, phrases may act as a precision device by virtue of the weighting effect already discussed. In a few circumstances, phrases also act as a recall device. The weighting effect acts as a precision device only when the concepts being added to the document increase the weight of some, and not all, of the identifiers already in the document; this is, in general, the case with SMART.

This discussion of devices shows that the clear-cut distinction between recall and precision devices is not easy to preserve, and its usefulness is probably now somewhat limited. This is particularly true because the effect of the various dictionaries is only detected in the performance measures if a certain type of cut-off is applied, namely, a cut-off that directly uses the

Associations Proved by Concon	Concepts in Request	Concepts in Document
Cylinder → Friction	Cylinder	Cylinder
Cylinder → Length	Known	Friction
Known → Kernel	Along	Length
Along → Blade	Etc.	Unit
Unit → Cylinder		Function
Function → Kernel		Velocity
Velocity → Blade		Etc.

(a) Section from concon pair association list

(b) Selection of concepts in request and document without concon

Request Expanded with Concon			Document Expanded with Concon			
(Cylinder)	→	Friction	⊥ Friction			} First order matches (type A*)
(Cylinder)	→	Length	⊥ Length			
		Cylinder	⊤ {Cylinder			} First order match (type B†)
			Cylinder	←	(Unit)	
(Known)	→	Kernel	⊥ Kernel	←	(Function)	} Second order matches *
(Along)	→	Blade	⊥ Blade	←	(Velocity)	
		Known	Unit			
		Along	Function			
		Etc.	Velocity			
			Etc.			

*Recall device effect

†Precision device effect

(c) Selection of the request and document concepts using concon, with five matches of three types generated

Illustration of Recall and Precision Device Effect of Concon

Fig. 7

number of matching terms or some constant threshold correlation coefficient. In an output graph, the effect of stem and thesaurus as a recall device can be seen when a threshold correlation coefficient of, say, 0.35 is applied to the search output. But such an effect cannot be detected in the complete precision versus recall curves that are normally used for evaluation. In particular, it is not correct to say that recall devices will cause the high recall end of the curve to be good, and precision devices will improve the high precision end. The only importance of the devices is that they become the means by which the specificity of the index language is altered; a dictionary that provides optimum specificity for a given test environment will exhibit a precision versus recall curve that is superior to all others probably over the whole performance range.

The optimum specificity of index language in the Cranfield project was found to be a stem type language. Such a result is given in a table of normalized recall ratios versus number of terms in language, (see Fig. 15, [11]). A plot of this type is included in Fig. 9, giving SMART results for three collections in addition to the Cranfield Project result. The Cranfield Project normalized recall curve is calculated differently from the SMART measure, so that no significance should be attached to the positions on the plot. The peak point of each curve shows the optimum dictionary; and whereas between 500 and 700 dictionary concepts produce optimum results for all three SMART collections, the Cranfield Project found that 2,500 was the optimum number. The interpretation of such plots needs further experimentation, since a count of the total concepts in a dictionary does not reflect the presence of a word in more than one concept, and the method employed is biased by collection size.

Dictionary	Total † Retrieved	Relevant Retrieved	Non-Relevant Retrieved	Recall*	Precision*
Suffix 's'	43	22	21	.111	.515
Stem	54	27	27	.136	.500
Thesaurus-3	227	67	160	.338	.295

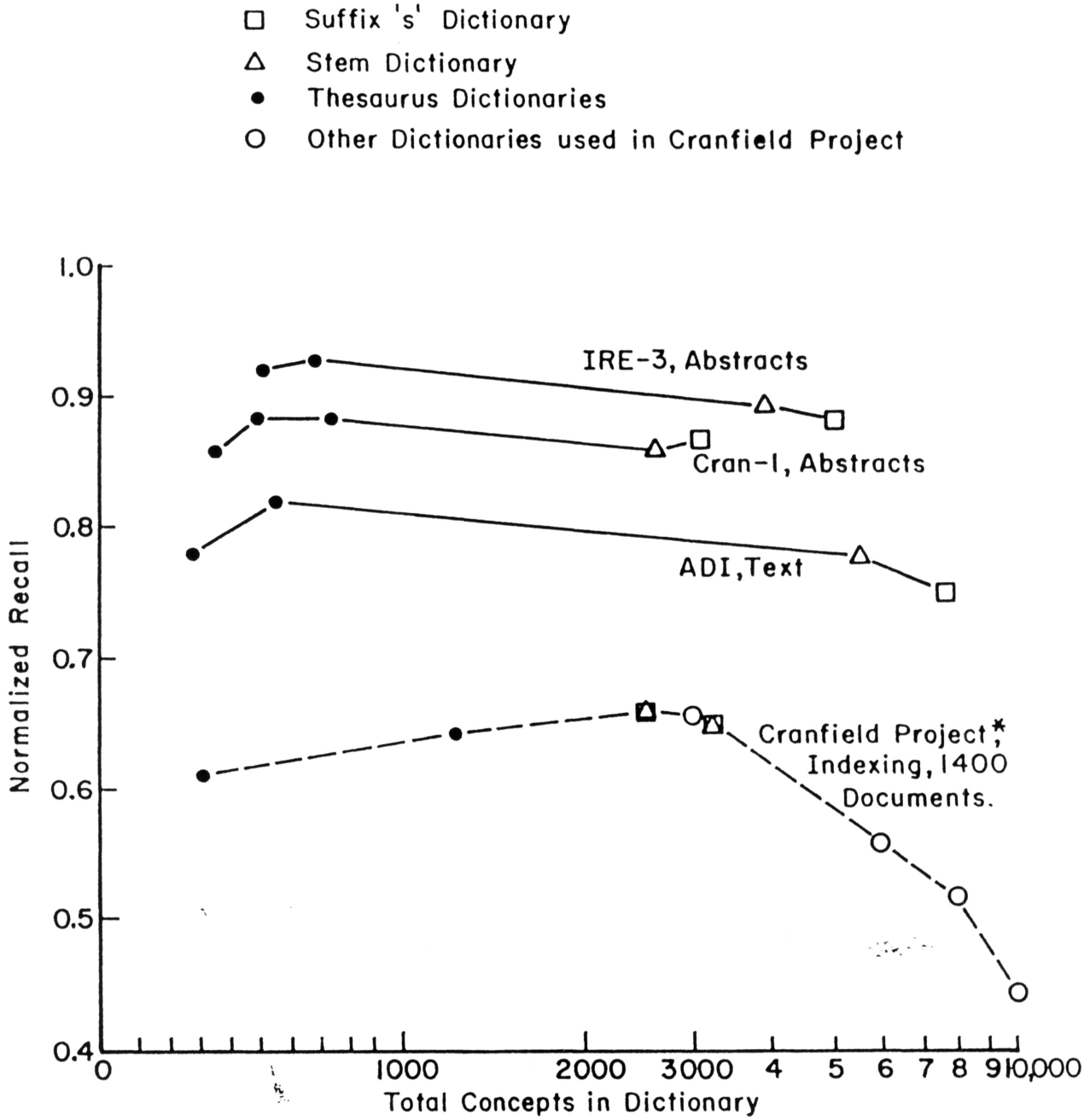
† Cut-off determined by Cosine Numeric Correlation $\geq$ 0.35

* Micro-evaluation with total relevant = 198

Cran-1 collection, abstracts, averages over 42 requests

Evaluation using Correlation Cut-off Showing Recall
Device Effect of the Dictionaries

Fig. 8



*The "Normalized Recall" is here computed by a method different from that used with SMART, and the difference in position on the plot is of no significance.

Plot of Normalized Recall versus Number of Concept in the Dictionary, Comparing three SMART Results with that from the Cranfield Project.

Fig. 9.

For example, the Cranfield Project results are based on the 1,400 collection; with the 200 collection in use, the optimum dictionary contains approximately 1,300 concepts.

5. Relevance Grade Test Results

It was noted in part 3 that in the case of the Cran-1 collection, relevance decisions are available that reflect degrees of relevance as judged by the persons supplying the requests. Since all SMART tests made so far have not used these different relevance grades, a brief examination of the relevance grade is made here.

It seems reasonable to postulate that the four grades of relevance produce different types of difficulties in achieving a good retrieval performance. Specifically, the documents graded most highly relevant probably achieve high rank positions on the output list, and those documents graded as of very minor relevance may have low rank positions in the search output. One method of analysis that may show whether this does occur is illustrated for a single request in Fig. 10. The ranks of the seven relevant documents are given for the actual search result, using the Cran-1 collection and the suffix 's' dictionary. For each relevant document, a relevance grade score is given, with the most highly relevant documents scoring 4, the next most relevant 3, then 2 and finally 1. If the expected result is achieved, the relevant documents with a grade score of 4 will be marked higher than those of 3, and so on. To test this, two other theoretical results are recorded in Fig. 10, including one for which the relevance grade scores follow the postulated pattern (described as

ACTUAL SEARCH RESULTS				THEORETICAL BEST RELEVANCE GRADE SCORE			THEORETICAL WORST RELEVANCE GRADE SCORE		
Ranks of Relevant	Document Number	Relevance Grade Score	Recall	Document Number	Relevance Grade Score	Recall	Document Number	Relevance Grade Score	Recall
1	10A	2	.12	10C	4	.24	09I	1	.06
2	09I	1	.18	10+	3	.41	983	2	.18
9	10+	3	.35	10B	3	.59	10A	2	.29
21	10B	3	.53	10A	2	.71	10D	2	.41
52	10D	2	.65	10D	2	.82	10+	3	.59
128	983	2	.76	983	2	.94	10B	3	.76
128	10C	4	1.00	09I	1	1.00	10C	4	1.00

Cran-1 Collection Abstracts, Suffix 's' dictionary

Request Q230, 7 relevant

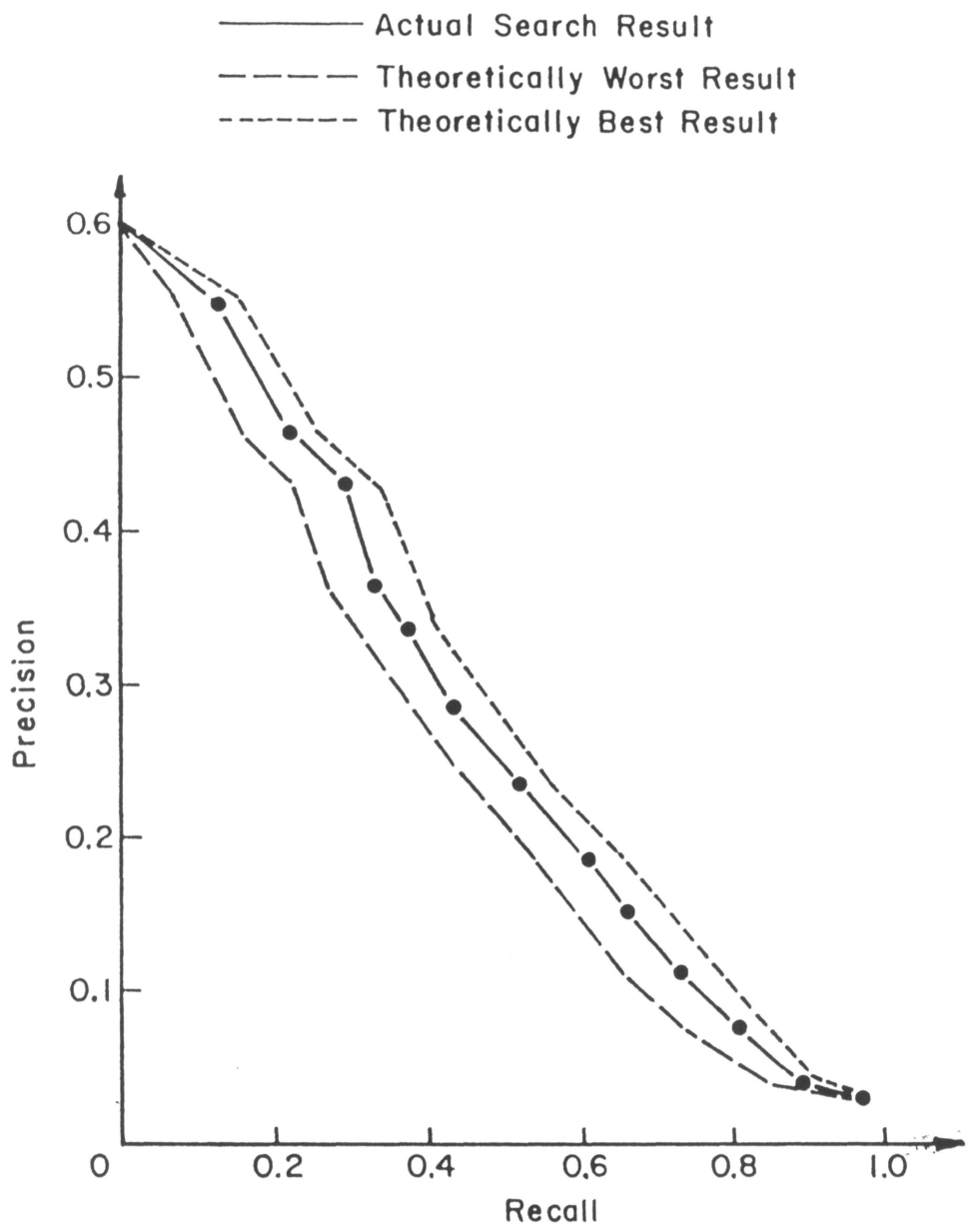
Analysis of Relevance Grade Results for a Single Search Request

Fig. 10

"theoretical best"), and another where they are completely the opposite of the expected pattern (described as "theoretical worst"). It is important to notice that the rank positions occupied by the seven relevant documents are not altered, but that some of the relevant documents exchange their position so as to obtain the desired relevance grade orders.

With these three stipulated rank orders for each request in the set, three average precision versus recall curves can be drawn, using methods described in Section II, part 5 in particular. Results for the set used as an example are given in Fig. 11. The plot shows that in this instance the ordering by relevance grade seems to be almost random, since the curve of "actual search result" falls mid-way between the theoretically best and worst. Further retrieval runs could be tried, but it is not believed that great differences will be seen when compared with these simulated situations. Calculation of the curves based on relevance grades for a thesaurus run has shown that the difference in merit between that thesaurus dictionary and suffix 's' using the relevance grade scores to obtain recall is virtually identical to the merit between the two runs when no relevance grades are allowed. This means that, with these two dictionaries at any rate, it is apparent that one dictionary is not more effective than another in retrieving relevant documents of particular relevance grades.

These results are in accord with similar tests made on the same data in the Cranfield Project [12, page 215]. The conclusion that there is no strong correlation between degree of relevance and ease in retrieval is probably due to the difficulty of making the relevance grade judgments in the first place.



Cran-1 Collection, Abstracts, Suffix 's' Dictionary,
Micro Averages over 42 Requests, Pseudo-Cranfield Cut-off.

Performance Curves based on Relevance Grades

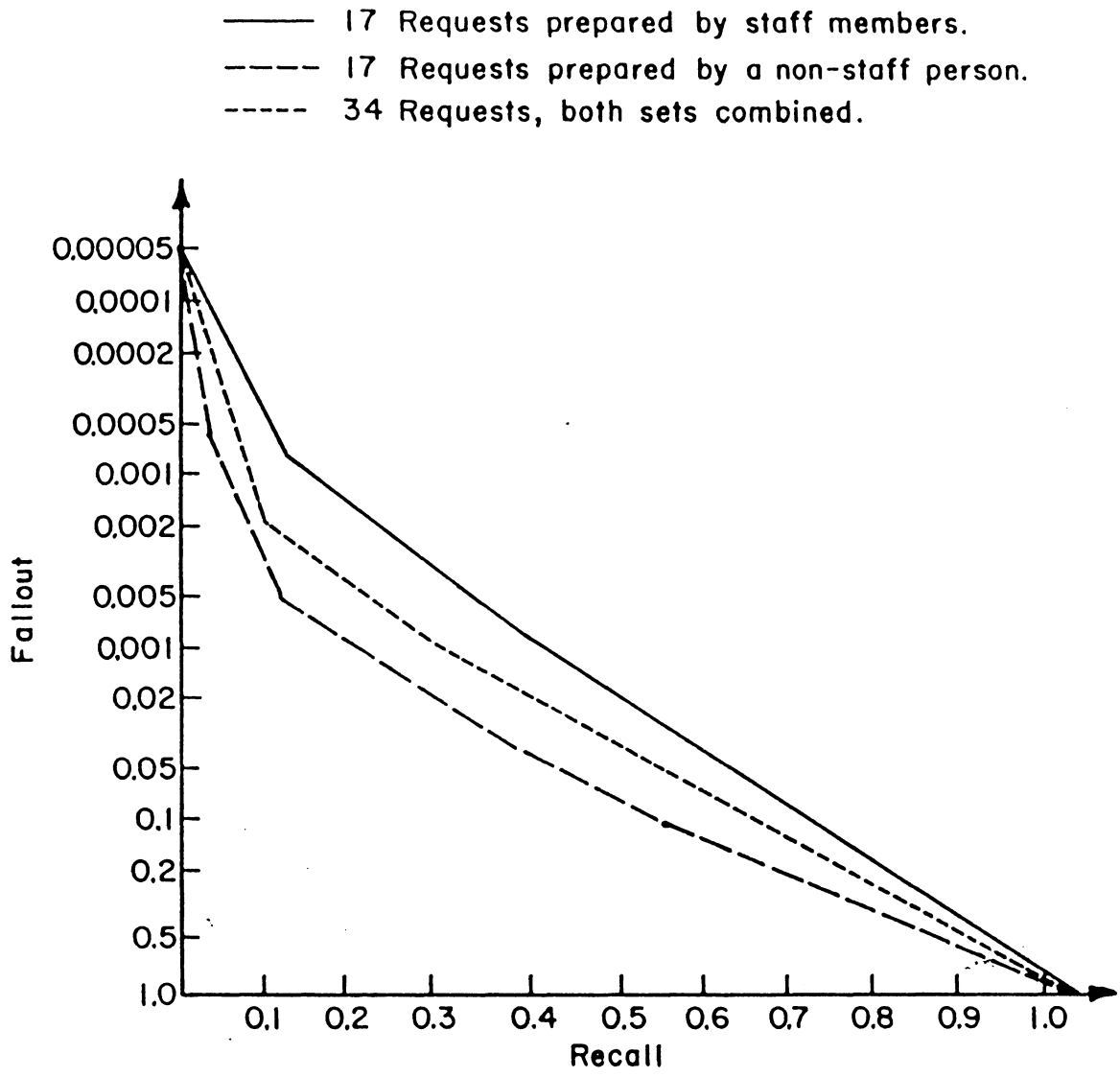
Fig. 11.

6. Request and Collection Comparisons

Some of the test environment characteristics presented in parts 2 and 3 and Figs. 1, 3, and 5, may be evaluated using search performance results. Three types of comparisons are made in the next few paragraphs.

A) Request Preparation

The set of 34 requests available for the IRE-3 computer science collection is made up from two sets of 17, each set being distinguished mainly be the persons preparing the requests and making the relevance decisions. A comparison of retrieval performance of the request set prepared by staff members with the set prepared by a non-staff person is made in Fig. 12. Fallout is used because average generality on the staff set is 27.5, with a mean of 21.6 relevant per request, and on the non-staff set, generality is 16.8, with a mean of 13.2 relevant per request. This suggests that the non-staff requests are more specific than the staff requests, and re-examination does show that the staff requests are a little longer (see footnote in Fig. 1). Three of the staff requests are found to be very similar to three of the non-staff ones. A comparison of these three pairs is therefore given in Fig. 13. Relevance decision agreement is quite strong for two of the pairs, but in every case the staff request is the longer and exhibits a quite superior retrieval performance. The variables to be considered in examining this type of request preparation and relevance decisions are known to be numerous, and it is not surprising that these subjective tasks have a large effect on the performance outcome. A paper by John O'Connor [13] and work done at System Development Corporation [14] provide further knowledge of these variables which can be used in future



IRE-3, Abstracts, Stem Dictionary, Micro Averages over request sets indicated, Pseudo-Cranfield Cut-off.

Performance Comparisons of Staff and Non-staff Prepared Requests

Fig. 12

Requests	LENGTH, STEM CONCEPTS		Total Relevant Documents	Documents Relevant to Both S and N Request	*Normalized Recall	*Normalized Precision
	Number of Unique	Sum of Weights				
S "Thin Films"	20	29	17	8	.9760	.8783
N Q001	10	10	20		.7637	.4864
S "Pattern Recg"	15	17	11	8	.9405	.8648
N Q004	7	7	8		.8909	.7260
S "Differentl EQ"	22	24	37	17	.9773	.8638
N Q016	17	19	20		.9522	.7911

S = Staff Prepared Requests

N = Non-Staff Prepared Requests

* Results using Abstracts, Stem Dictionary, Cosine Numeric

Comparison of Staff and Non-Staff Prepared Requests
on IRE-3, using three Pairs of Similar Requests

Fig. 13

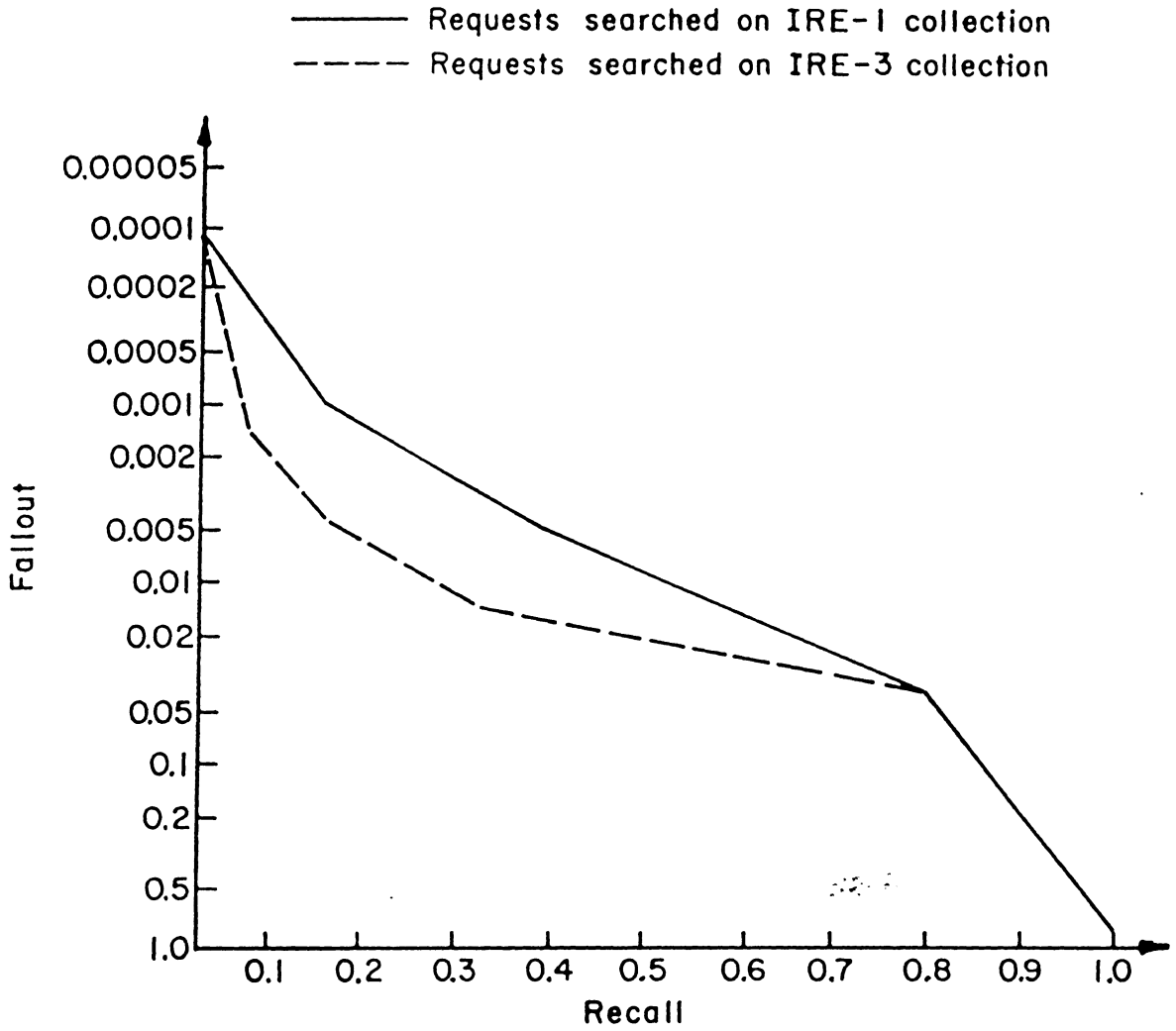
experimental tests. An examination in more detail of the documentation requests appears in Section X, parts 3 and 4.

One further performance comparison appears in Fig. 14. Here the 17 staff-prepared requests are searched on the two different collections, the only variation being that the relevance decisions for the IRE-2 collection were made much later in time than those on the IRE-1 collection. The mean relevant per request in the IRE-1 collection is 10.9, and in the IRE-2 collection 10.6, implying only a small change in generality (27.0 on IRE-1, and 28.3 on IRE-2). The small difference in performance observed must be due in part to the fact that relevance decisions by the same individual are not entirely consistent over periods of time, and also because the IRE-2 collection may be more hostile to good retrieval (there may be more marginally relevant or falsely matched documents).

B) Specific and General Requests

Data given in Fig. 3 divides up the request sets into specific and general according to the numbers of relevant documents in the collection. A performance comparison of specific with general requests raises some quite complex evaluation problems which are discussed in Section II. Because no complete solution to these problems has yet been found, a reasonably simple presentation will be given.

Fig. 15 is a simplified representation of nine comparisons made between four sets of specific and general requests. The first request set is from the IRE-1 collection using the 17 staff-prepared requests, since this result has appeared previously [15]; the three other request sets are the IRE-3, Cran-1 and ADI sets which are now used for test purposes.



IRE Collections, Abstracts, Stem Dictionary, Micro Averages over 17 Requests, Pseudo-Cranfield Cut-off

Performance Comparison of the 17 Staff Preferred Requests Search on the two Document Collections.

Fig. 14

Collection	Average Rank of Last Relevant (Fig. 17)	Recall at Given Document Cut-off (e.g. Fig. 16)	Fallout Versus Recall Curve, Pseudo Cut-off	Normalized Recall	Normalized Precision	Average Rank of First & Second Relevant (Fig. 18)	Fallout at Given Document Cut-off (e.g. Fig. 16)	Precision at Given Document Cut-off (e.g. Fig. 16)	Precision Versus Recall Curve, Pseudo Cut-off
IRE-1	S	S	S	S	S	G	G	G	S
IRE-3	S	S	S	G	S	G	G	G	G
CRAN-1	S	S	S	S	S	G	G	G	G
ADI	S	S	S	S	S	G	G	G	G

S = Specific Requests Perform best

G = General Requests Perform best

All results are based on Abstracts with Stem Dictionary

Summary of Results of Nine Ways of Comparing Specific and General Requests

Fig. 15

The comparisons in columns 4 and 5 use the normalized evaluation measures, and in columns 3 and 9 the usual curves of fallout versus recall, and precision versus recall are used. The entries in the table show, for example, that for the ADI comparison the fallout versus recall curve produces a superior performance for the specific requests, while the precision versus recall curve shows the superiority to be with the general requests. Any slight crossing of the curves is ignored in this table.

The entries in columns 2, 7 and 8, are explained by the example given in Fig. 16. Using the IRE-3 results, the detailed performance results of the specific and general requests are compared at five document cut-off levels. It can be seen that although the general requests retrieve a greater number of relevant at each cut-off point compared with the specific requests, a lower recall ratio is achieved each time since with the general requests there are many more relevant to find. Also, the general requests are seen to achieve better fallout and precision ratios at each cut-off. Returning to columns 3 and 7 in Fig. 15, with only one exception the precision versus recall curve shows the general requests to be best, and the fallout versus recall curves all favor the specific requests. The exception to this, noted in the case of IRE-1, may be explained by the fact that these 17 staff-prepared requests perform very much better than any other set for any of the collections, and the useful length of these requests seems to offset the generality effect which favors the general requests in the set.

This description reveals the difficulties involved in making this type of test comparison. As is suggested in Section II, part 7 or 8, user-oriented evaluation seems to be performed best by recognizing two

Performance	SPECIAL REQUESTS Cut-off, n Documents					GENERAL REQUESTS Cut-off, n Documents				
	n = 1	2	3	4	5	n = 1	2	3	4	5
Total Retrieved	17	34	51	68	85	17	34	51	68	85
Relevant Retrieved	12	21	27	32	38	14	24	37	48	61
Non-Relevant Retrieved	5	13	24	36	47	3	10	14	20	24
Recall	.09	.16	.21	.25	.30	.03	.05	.08	.10	.13
Fallout	.00038	.00098	.00181	.00272	.00355	.00023	.00078	.00109	.00155	.00186
Precision	.71	.62	.53	.47	.45	.82	.71	.73	.71	.72

IRE-3 Collection, Abstracts, Stem Dictionary, Micro Averages over 17 Specific and 17 General Requests.

Comparison of Specific and General Requests in the
IRE-3 Collection at the First Five Document Cut-offs

Fig. 16

extreme types of user need, those of high precision and those of high recall. The high recall comparison may be carried out by comparing the average rank position taken up by the last relevant document, and Fig. 17 shows that in all tests the specific requests are very clearly superior. The high precision comparison may be carried out by computing an average rank position for the first two relevant documents, and Fig. 18 shows that here the general requests give a superior performance. These results use the stem dictionary. Since the thesaurus dictionaries normally produce a superior performance, some change in merit between the specific and general requests might result for the thesaurus runs. The same data is therefore repeated for the thesaurus runs in Figs. 19 and 20, where it is again seen that for a high recall need the specific requests are best, and for a high precision need the general requests are best.

It is quite likely that in an operational situation, users wanting high recall would tend to pose specific requests, and users wanting high precision would tend to pose general requests. But there could certainly be exceptions to this, and the suggested correlation might not exist at all. The most disturbing part of this finding is that the specific requests, which were thought to be the better ones for retrieval purposes, do not perform very well for high precision users, although with the thesaurus dictionaries in use the gap between specific and general requests on ADI and Cran-1 (Fig. 20) is narrower than with the stem dictionaries (Fig. 18). Further work in this area requires better procedures for distinguishing specific and general requests, since the use of request generality in a small test collection is not intended to produce any fundamental division that

COLLECTION AND DICTIONARY	SPECIFIC REQUESTS		GENERAL REQUESTS	
	Average Rank of Last Relevant	Number of Requests	Average Rank of Last Relevant	Number of Requests
IRE-1, Abstract, Stem	33.1	9	141.5	8
IRE-3, Abstract, Stem	272.7	17	395.2	17
CRAN-1, Abstract, Stem	48.8	21	95.3	21
ADI, Abstract, Stem	25.4	17	52.3	18

Comparison of Specific and General Requests using the Average Rank Position of the Last Relevant Document to represent High Recall Need, (Stem Dictionary)

Fig. 17

COLLECTION AND DICTIONARY	SPECIFIC REQUESTS			GENERAL REQUESTS		
	Average Rank of First Relevant	Second Relevant	Number of Requests	Average Rank of First Relevant	Second Relevant	Number of Requests
IRE-1, Abstract, Stem	1.2	3.9	9	1.0	2.4	8
IRE-3, Abstract, Stem	7.4*	19.9*	17	1.4	3.4	17
CRAN-1, Abstract, Stem	11.1†	18.0†	21	4.6	8.2	21
ADI, Abstract, Stem	11.1	20.2	17	5.4	12.6	18

* Deleting one very bad request gives average ranks 2.0 and 14.4

† Deleting one very bad request gives average ranks 6.0 and 13.0

[Where requests have no second relevant, that request is not included in the average results]

Comparison of Specific and General Requests using the Average Rank Position of the First and Second Relevant Documents to represent High Precision Need, (Stem Dictionary)

Fig. 18

COLLECTION AND DICTIONARY	SPECIFIC REQUESTS		GENERAL REQUESTS	
	Average Rank of Last Relevant	Number of Requests	Average Rank of Last Relevant	Number of Requests
IRE-1, Abstract, Thesaurus-2	18.9	9	96.9	8
IRE-3, Abstract, Thesaurus-3	210.0	17	293.5	17
CRAN-1, Abstract, Thesaurus-3	45.2	21	85.8	21
ADI, Abstract, Thesaurus-1	22.1	17	46.1	18

Comparison of Specific and General Requests using the Average Rank Position of the Last Relevant Document to represent High Recall Need, (Thesaurus Dictionary)

Fig. 19

COLLECTION AND DICTIONARY	SPECIFIC REQUESTS			GENERAL REQUESTS		
	Average Rank of First Relevant	Second Relevant	Number of Requests	Average Rank of First Relevant	Second Relevant	Number of Requests
IRE-1, Abstract, Thesaurus-2	1.1	3.4	9	1.1	2.1	8
IRE-3, Abstract, Thesaurus-3	9.4	20.0	17	1.1	2.4	17
CRAN-1, Abstract, Thesaurus-3	5.1	10.6	21	3.9	8.8	21
ADI, Abstract, Thesaurus-1	6.5	18.2	17	2.3	7.9	18

Comparison of Specific and General Requests using the Average Rank Position of the First and Second Relevant Documents to represent High Precision Need, (Thesaurus Dictionary)

Fig. 20

that is valid outside the particular test. Other parameters such as request length and request concept frequency are used in the study in Section X.

C) Collection Comparisons

The data which describe the test environments in Figs. 1, 3, 4, and 5 reveals many points at which the environments differ, such as collection and request sizes, collection and request average lengths, request generality, request preparation and relevance decisions, and so on. It is recognized that at present, it is not possible to sufficiently control these variables so that comparisons between collections can be made under the assumption that the effects of these variables have been adequately controlled. Suitable control of these and other so far unrecognized variables would permit comparisons between collections of documents in different subject areas. This might be of interest since the terminology of different subject areas might be regarded as lying on a continuum ranging from "hard" or "firm" subject areas to "soft" or "mushy" as suggested by Cleverdon [16]. This may be a valid hypothesis, since in data retrieval situations in some areas of chemistry, the firm language permits simultaneous high recall with high precision performances, whereas in other areas such as parts of the social sciences the imprecise language often produces very much poorer precision recall curves. Alternatively, it may be the case that subject fields contain sub-areas of soft and firm terminology: in aerodynamics, for example, descriptions of wing shapes and aspect ratios seem to be fairly unambiguous, whereas treatment of gas and fluid flow phenomena seems to abound with ambiguities.

Information is given in Fig. 21 comparing nine collections on the basis of word occurrences. A standard list of 204 common words is used in each case to isolate the total non-common words and total unique non-common words. It may be noted that in seven of the collections, the proportion of non-common to total word occurrences is between 55.3% and 56.5%; even the two ADI collections are not far outside this range. The proportion of unique (or distinct) non-common words to total non-common word occurrences varies both with document length and collection size. For example, if the collections are divided into the six having 82-405 documents, and the three having 780-1400 documents, the unique-to-total proportion (c/b) varies directly with average document length within the two groups. The one small exception is the Medlars collection, but the abundance of technical names in medicine may be the cause. Although further analysis could be done, the data in Fig. 21 suggests that the common factor of English text provides strong uniformity in the statistics given irrespective of subject area. This does not directly confirm or reject the subject language precision ideas, since ambiguity is not reflected in any of the statistics given.

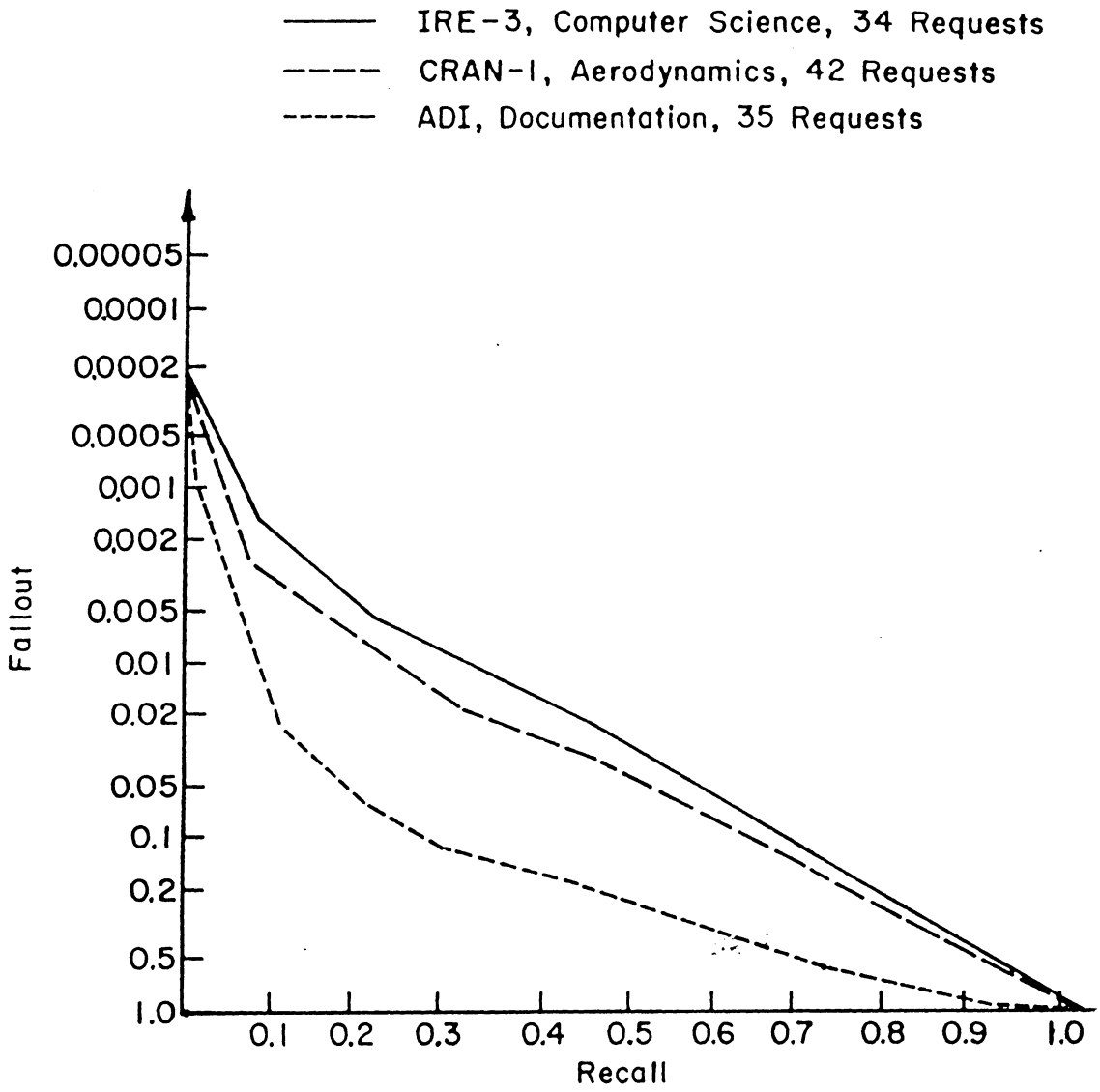
A retrieval performance plot comparing results from three collections is given in Fig. 22. The type of dictionary used is the automatic stem procedure, since use of thesaurus dictionaries would introduce the additional element of varying human skills in thesaurus construction. Furthermore, many variables exist due to request preparation and relevance decisions between the collections; the extent to which these variables affect the result is not known. It can be suggested however, that the superiority of the computer science collection and the inferiority of the

Collection	Average Document Length Average of (b)	Collection Size	Total Word Occur. (a)	Total Non-Common Word Occur. (b)	b/a	Total Unique Non-Common Words (c)	c/b
ADI Abs.	35	82	4,872	2861	58.7%	1321	46.2%
IRE-1	44	405	31,663	17,729	56.0%	4041	22.8%
IRE-3	49	780	68,947	38,572	55.9%	5477*	14.2%
IRE-2	56	375	37,284	20,843	55.9%	3751	18.0%
ISPRA	58	1268	131,491	73,410	55.8%	7980	10.9%
Medlars	80	276	38,958	22,023	56.5%	5331	24.2%
CRAN-2	91*	1400	231,294*	127,813*	55.3%	8887*	7.0%
CRAN-1	91	200	33,042	18,259	55.3%	3123	17.1%
ADI Txt	710	82	113,130	58,190	51.4%	7925	13.6%

* Estimated

Comparison of Word Occurrence Statistics of the English Text in Nine Collections

Fig. 21



Results based on Abstracts, Stem Dictionary, Micro Averages over Request Sets, Pseudo-Cranfield Cut-off.

Performance Comparison of Three Collections in Different Subject Areas

Fig. 22

documentation collection does follow the expected pattern if precision in subject terminology is of importance. Comparison with Fig. 12 reveals that request preparation is a large variable, and use of the 17 non-staff prepared requests would be expected to result in a curve lower than that for the Cran-1 aerodynamics results.

Another technique of collection comparison uses the average rank technique as used in part 6B for comparing the specific and general requests. Fig. 23 gives results based on the stem dictionaries, and Fig. 24 gives results based on the thesaurus dictionaries. The average rank positions of the first and second relevant documents reflect the viewpoint of a user needing high precision. Ignoring differences in collection size, the Cran-1 aeordynamics collection gives a good result using the thesaurus and on ADI the first relevant receives the best average rank. Use of the percentage figure to take into account changes in collection size restores the expected merit. Figs. 23 and 24 also record the average rank positions of the last relevant document to reflect the viewpoint of the high recall user. The average rank is directly affected and ordered by collection size, but the percentage figure shows that IRE-3 and Cran-1 perform a little better than ADI.

COLLECTION AND DICTIONARY	Number of		FIRST RELEVANT		SECOND RELEVANT		LAST RELEVANT	
	Documents	Requests	Average Rank	% of Collection	Average Rank	% of Collection	Average Rank	% of Collection
:RE-3, Abstracts, Stem	780	34	4.4	0.6%	11.6	1.5%	334.0	42.8%
:RAN-1, Abstracts, Stem	200	42	7.8	3.9%	13.0	6.5%	72.0	36.0%
ADI, Abstracts, Stem	82	35	8.1	9.9%	15.7	19.2%	39.2	47.8%

Comparisons of the Three Collections using Average Rank Positions of the First, Second and Last Relevant Documents, Stem Dictionary

Fig. 23

COLLECTION AND DICTIONARY	Number of		FIRST RELEVANT		SECOND RELEVANT		LAST RELEVANT	
	Documents	Requests	Average Rank	% of Collection	Average Rank	% of Collection	Average Rank	% of Collection
IRE-3, Abstracts, Thesaurus-3	780	34	5.2	0.7%	11.2	1.4%	251.8	32.3%
CRAN-1, Abstracts, Thesaurus-3	200	42	4.5	2.3%	9.7	4.9%	65.5	32.8%
ADI, Abstracts, Thesaurus-1	82	35	4.3	5.2%	12.2	14.9%	34.5	42.1%

Comparison of the Three Collections using Average Rank Positions of the First, Second and Last Relevant Documents, Thesaurus Dictionary

Fig. 24

References

[1] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Information Storage and Retrieval, Report ISR-12, to the National Science Foundation, Section III, Department of Computer Science, Cornell University, August 1967.

[2] G. Salton, The SMART Project - Status Report and Plans, Information Storage and Retrieval, Report ISR-12, to the National Science Foundation, Section I, Department of Computer Science, Cornell University, August 1967.

[3] G. Salton, The Evaluation of Computer-based Information Retrieval Systems, Proceedings 1965 International FID Congress, Spartan Books, Washington, 1966.

[4] M. E. Lesk, The Significance Programs for Testing the Evaluation Output, Information Storage and Retrieval, Report ISR-12, to the National Science Foundation, Section II, Department of Computer Science, Cornell University, August 1967.

[5] R. V. Katter, Experimental Investigations of a Method for Analyzing Document Representation, Progress Report, TM 3090, System Development Corporation, August 1966.

[6] J. J. Rocchio and G. Salton, Information Search Optimizatioi and Interactive Retrieval Techniques, Proceedings 1965 Fall Joint Computer Conference, Washington, 1965.

[7] C. W. Cleverdon and J. Mills, The Testing of Index Language Devices, Aslib Proceedings, Vol. 15, No. 4, April 1963.

[8] C. Cleverdon, J. Mills and M. Keen, Factors Determining the Performance of Indexing Systems, Volume 1, Design, Aslib Cranfield Research Project, Cranfield, 1966.

[9] K. S. Jones and D. Jackson, Some Experiments in the use of Automatically Obtained Term Clusters for Retrieval, 1967 F.I.D.-I.F.I.P Conference, Rome, June 1967.

[10] K. S. Jones and D. Jackson, The Use of the Theory of Clumps for Information Retrieval, Report on the OSTI Supported Project at the Cambridge Language Research Unit, Cambridge, England, June 1967.

References (contd.)

[11] C. Cleverdon, The Cranfield Tests on Index Language Devices, Aslib Proceedings, Vol. 19, No. 6, June 1967.

[12] C. Cleverdon and M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 2, Test Results, Aslib Cranfield Research Project, Cranfield, 1966.

[13] J. O'Connor, Relevance Disagreements and Unclear Request Forms, American Documentation, Vol. 18, No. 3, July 1967.

[14] C. A. Cuadra, et al., Experimental Studies of Relevance Judgments: Final Report, Vol. 1, Project Summary, TM - 3520|001|00, Vol. 2, Description of Individual Studies, TM - 3520|002|00, System Development Corporation, June 1967.

[15] G. Salton, The Evaluation of Automatic Retrieval Procedures - Selected Test Results Using the SMART System, American Documentation, Vol. 16, No. 3, July 1965 (Also ISR-8, Section IV).

[16] C. W. Cleverdon, The Testing and Evaluation of the Operating Efficiency of the Intellectual Stages of Information Retrieval Systems, Proceedings of the International Study Conference on Classification Research, Elsinore, 1964.