V. Document Length

E. M. Keen

1. Introduction

A major advantage in the design of automatic document retrieval systems is the ability to add new documents to the collection without the necessity for an individual manual content analysis. This is done by using the natural language text of the documents as input, together with automatic analysis procedures based on pre-stored dictionaries to achieve vocabulary normalization. Such an automatic procedure is not necessarily straightforward however, and various possible alternatives must be considered. This study will deal with the influence of document length as used in a SMART type system.

One of the main elements of a manual document analysis or indexing procedure that has been in use for many years is the process of term selection, whereby the indexer makes a choice of subject ideas from the document being indexed. This selection process always requires a difficult management decision because some of the users will benefit from highly exhaustive indexing (the selection of many subject ideas); on the other hand, factors such as cost and search time often limit the indexing process to one of low exhaustivity. As a first approximation, an automatic method using natural language text provides the answer to this problem, since the whole document text can now be used, without any pre-selection activity at all. Although use of full text is possible in theory, in practice, various limitations must be taken into account. For example, there exists the input problem,

namely the effort and cost associated with the transformation of whole document texts into machine readable form. Several possible solutions are suggested to this problem, such as the development of a universal print reader, or the use of some by-product of the typesetting stage.

Then there arises the problem of coding and searching documents which contain many mathematical equations, complex diagrams, or other essential non-textual material. Then again, for the user, the search response time is likely to be long when full text is stored even with small document collections, although faster search procedures may be possible in the future. Lastly, the use of full text may not serve all users well with regard to retrieval performance, since the requestor may be swamped with many documents that are strictly relevant but rather trivial in relation to the topic of the search request.

For these reasons, automated systems of the first-generation will need to consider selections of the document text, rather than the whole text. Many documents contain suitable selections of text made by the author of the document, such as the title itself, or probably better still, an abstract or summary of the paper. Like the product of manual indexing, an abstract or summary of a document is a précis of the document which distills the essential subject ideas into a few hundred words. The presence of bias or slant in both indexing and abstract preparation may not favor the use of natural language input, however, since in such a system there often exists no possibility of picking out only those topics of interest to the users of the system (as is possible in manual indexing). In addition, for some documents the natural language abstract may be a poorly written précis of the document. When an automatic system using abstracts is implemented it may be necessary to make up these deficiencies by manual effort; procedures are also

needed for handling documents which are not available with a suitable abstract. Some larger selections from the full text of documents consisting of more material than the abstract, yet less than full text, may be possible; for example, section headings and figure captions might be added to the abstract.

In the present study, several different selections of documents will be compared, the shortest being titles only, and the longest a collection of full text 'short' conference papers. Evaluation of these different document lengths will center on the retrieval performance achieved. Other evaluation criteria such as search time and input cost will be of considerable importance in operational environments, but in the experimental tests being performed on the SMART system no reasonable simulation test of these criteria can yet be made.

2. SMART Test Comparisons

Three series of comparisons of document length are presented. Firstly, the use of abstracts (including titles) is compared to the use of document titles alone. Results are presented for the three collections of documents being used for current experiments in the subject areas of computer science (IRE-3, 780 documents, 34 requests), aerodynamics (Cran-1, 200 documents, 42 requests), and documentation (ADI, 82 documents, 35 requests). Secondly, using the ADI Collection the abstracts are compared to the use of full text. In the main results, the text used includes the abstract, and both naturally include the title, so that three distinct document lengths are available for comparison. The ADI Text Collection consists of a set of short conference papers of average length 1,380 words; it is therefore not typical of scientific papers in general, and does not pose any problems due to non-textual material. The third comparison is made with the Cran-1 abstracts which are

compared to the manual indexing available for that collection. This comparison is made because the indexing takes up about half the length of the abstracts, and constitutes a valid comparison because of the unusual nature of the indexing, which is "... a base list of words, selected directly from the title and text of a document ... presented without any reference whatsoever to a control list for synonyms, related terms, etc." [1, page 41, see also pages 48, 52]. The controls used in indexing permitted the confounding of singular and plural word forms, as well as variant spellings, but the index terms were otherwise culled from the documents in natural language. The indexing used is then, in effect, another abstract of the documents, shorter in length than the author abstract, and produced by trained indexers. It is expected that the choice of subject ideas from the whole document by the indexers will be very similar on average to the choice of ideas made by the abstractors, although the area of overlap has not been determined.

Retrieval runs of the above comparisons are presented using the stem and thesaurus dictionaries and all results use the cosine correlation and numeric vectors, unless otherwise stated.

The comparative lengths of the documents in these comparisons are given in Figure 1. Although the lengths given in the figure are based on the concepts resulting from the documents being looked-up in the suffix 's' dictionary, relative lengths will remain the same using the stem and thesaurus dictionaries.

3. Effect of Changes in Document Length

In this part, the effect of changes in document length on the match between requests and documents is considered, followed by the expected differences in retrieval performance.

INPUT TEXT TYPE	SUBJECT FIELD OF COLLECTION	COLLECTION NAME	*AVERAGE (MEAN) CONCEPTS PER DOCUMENT USING SUFFIX 's' DICTIONARY
Abstract	Computer Science	IRE-3	40
Title	Computer Science	IRE-3	5
Abstract	Aerodynamics	CRAN-1	65
Indexing	"	CRAN-1	33
Title	"	CRAN-1	9
Full Text	Documentation	ADI	369
Abstract	"	ADI	25
Title	"	ADI	6

* Averages are based on 10% random sample.

Average lengths of documents using titles, abstracts, indexing and full text as used in the SMART experiments on three document collections.

Figure 1.

It may be expected that, commencing with documents short in length, any increase in length will increase the number of concepts that match between the requests and documents. In the type of test environment used by SMART, namely a simulated real-life situation using requests and relevance judgments that are inevitably subjective in nature, it is quite rare for any short length documents to completely match with all the request concepts. In cases where a complete match does occur, it is naturally not necessary to increase the document length to improve the request/document match, except that in the numeric vectors scheme, the matching concepts are often increased in the longer documents.

The effect of the use of the cosine correlation with numeric vectors is complex, because this matching scheme includes the length of both the request and document, as well as the matching concepts in the algorithm, as follows:

$$\text{Cosine Correlation Coefficient} = \frac{Mw}{\sqrt{Rw \times Dw}}$$

where Mw = The concepts that Match between a Request and a Document, using the sums of products of the weights assigned to the matching concepts;

Rw = The total concepts in the Request, using the sums of the squares of the weights assigned to the concepts;

Dw = The total concepts in the Document, using the sums of the squares of the weights assigned to the concepts.

The resulting coefficient is obtained for each request in relation to every document in the collection, so that the output of the search may be an ordered list of documents. In tests investigating document length, all other variables such as the request set, the document collection, the word dictionary

and the matching algorithms are held unaltered while, say, abstracts and titles are compared. Considering the cosine correlation coefficient for just one document in relation to one request, it is clear that a change from titles to abstracts will not affect Rw in the equation. Factor Dw will increase directly with an increase in document length however. Factor Mw will either increase or remain constant, depending on whether the use of the abstract compared with title only achieves a match with more of the request concepts, and/or increases the weights of the concepts that already match on titles. The resulting difference in correlation coefficient between the title and abstract input cannot be predicted: if the abstract provides more matching concepts (Mw), and does not increase document length (Dw) too drastically, the abstract result will give a higher correlation coefficient than the title. If the abstract provides no additional matching concepts or increased weights, then the correlation with abstracts will be less than that on titles.

An example of what happens in one particular case is given in Figure 2. Details of the request and relevant document are given, as well as portions of the document as looked-up in a thesaurus dictionary using first the title only, then the whole abstract, then the full text. Document length sharply increases to 109 concepts with full text over 12 in the abstract and five in the title. The match between the request and document starts at two out of the six possible concepts with titles; the use of abstracts increases the weight of these two matching concepts, and full text increases the matching concepts to all six, as well as improving weights. However, the cosine correlation coefficients show that in this example the increases in document length exert more influence in the coefficient than the increases in matching concepts, so that the correlation coefficient drops from 0.3651 to 0.3608 with abstracts, and further still to 0.2034 with text.

Request QA8 Describe information retrieval and indexing in other languages. What bearing does it have on the science in general?

Request looked-up in Thesaurus Dictionary

Dictionary Concept Number	Weight	Words in Thesaurus Group
1	1	Information
4	1	Index
5	1	Retrieve, information-retrieval, IR, recall, recover.
10	1	Science, scientific
134	1	General, comprehens-, total, universal
147	1	Language, lingu-

Document 61 (Relevant) is entitled: An experiment in automatic indexing of French documents.

Example of the change taking place in the correlation coefficient (cosine) between a request and a relevant document when three document lengths, titles, abstracts, full text, are used. (ADI Collection, Thesaurus Dictionary)

Figure 2

Document 61 looked-up in Thesaurus Dictionary

	Concept	Weight	Concept	Weight	Concept	Weight
Title only (5 concepts)	4	1	19	1	147	1
	8	1	126	1		
Additional Concepts and Weights added by Abstract (12 concepts)	4	3	48	1	115	1
	8	2	51	1	126	2
	19	2	57	1	147	2
	42	1	72	1	286	1
A sample of the additional concepts and weights added by full text (109 concepts)	1	1	5	6	8	28
	3	1	6	1	10	1
	4	24	7	4	12	1
	.	.	.	.	.	.
	.	.	.	.	.	.
	134	2	147	6	286	17

Cosine Correlation

1. Title only $\cos = \frac{2}{\sqrt{30}} = 0.3651$

2. Abstract $\cos = \frac{5}{\sqrt{192}} = 0.3608$

3. Full Text $\cos = \frac{40}{\sqrt{38,674}} = 0.2034$

Figure 2
(continued)

It is important to realize that this discussion of the effect of changes in document length on the correlation coefficient applies to correlations with either relevant or non-relevant documents, and a change in correlation resulting from a change in rank gives no indication of the change in retrieval performance. Retrieval Performance is always a trade-off between relevant and non-relevant documents, and increases in document length can just as easily worsen performance as improve it, since longer documents produce greater opportunity for incorrect matches with non-relevant documents. An illustration of the effect of change in document length on a non-relevant document is given in Figure 3, where document 41 is presented for both title and abstract searches in relation to request QA8, used earlier in Figure 2. Again the cosine correlation shows a decrease with increase in document length, but in this case since the change from titles to abstracts does not even improve the weights in the two matching concepts, a severe drop in correlation takes place. Retrieval performance for this request is given in Figure 4, considers only documents 61 (relevant, see Figure 2) and 41 (non-relevant). It is seen that document 41 is more highly correlated (and therefore better ranked) than document 61 with titles; a reverse result is obtained with the abstracts. In this one example, the increase in document length improves performance, but many individual cases have been observed of the reverse trend.

It seems logical to postulate that, for a given set of search requests, relevance judgments, and document collection, there must exist an optimum document length that gives the best retrieval performance. However, in general, this is too simple a statement, and does not allow for the fact that performance requirements in terms of either high recall or high precision may demand different document lengths under different circumstances for optimum

Request data appears in Figure 2.

Document 41 (Non-Relevant) is entitled: Analysis, indexing and correlation of information.

Document 41 looked-up in Thesaurus Dictionary

	Concept	Weight	Concept	Weight	Concept	Weight
Title only (4 concepts)	1	1	33	1		
	4	1	296	1		
Abstract (17 concepts)	1	1	33	1	128	1
	4	1	36	1	140	1
	19	1	42	1	267	1
	20	1	48	1	286	1
	22	1	51	2	296	1
	32	2	66	1		

Cosine Correlation Coefficients

1. Title only $\quad \text{Cos} = \frac{2}{\sqrt{24}} = 0.4082$

2. Abstract $\quad \text{Cos} = \frac{2}{\sqrt{138}} = 0.1703$

Example of the change in cosine correlation between request QA8 (see Figure 2) and non-relevant document 41, when title only is compared with abstracts.

(ADI Collection, Thesaurus Dictionary)

Figure 3.

	Document 61 (Relevant)		Document 41 (Non-Relevant)	
	Correlation	Rank	Correlation	Rank
Title	.3651	10	.4082	8
Abstract	.3608	8	.1703	39

Cosine correlation coefficients and rank positions for two documents in relation to request QA8, giving an example of a superior result with abstracts compared to titles. (ADI Collection, Thesaurus Dictionary)

Figure 4.

performance. It is quite clear that where high recall is required, long documents are needed, since short documents, or low exhaustivity, constitute an absolute bar on the recall attainable; this "recall ceiling" is one of several important criteria for evaluating changes in document length. The opposite of this statement does not follow automatically, since it is not necessarily true that for high precision requirements short documents are needed. For a requirement of highest precision at low recall, some optimum document length normally exists in a given environment, and tests presented on SMART will give some idea of this optimum length for the different test collections used.

4. Test Results

Test results which consist of retrieval performance comparisons are given first for abstracts versus titles, then for abstracts versus full text, and finally for abstracts versus indexing. In each of these sub-sections, performance comparisons will be made using three main techniques:

— Overall performance measures, consisting of normalized recall values, normalized, precision values, and precision/recall graphs;

— Recall ceiling data, using recall alone;

— Individual request and relevant document data, using tables and graphs of the numbers of requests and documents that favor a given option.

After the main test results for each sub-section have been presented, additional test results of value are also described. All results are averages over the set of requests being tested, as indicated in the figures.

A) Abstracts versus Titles

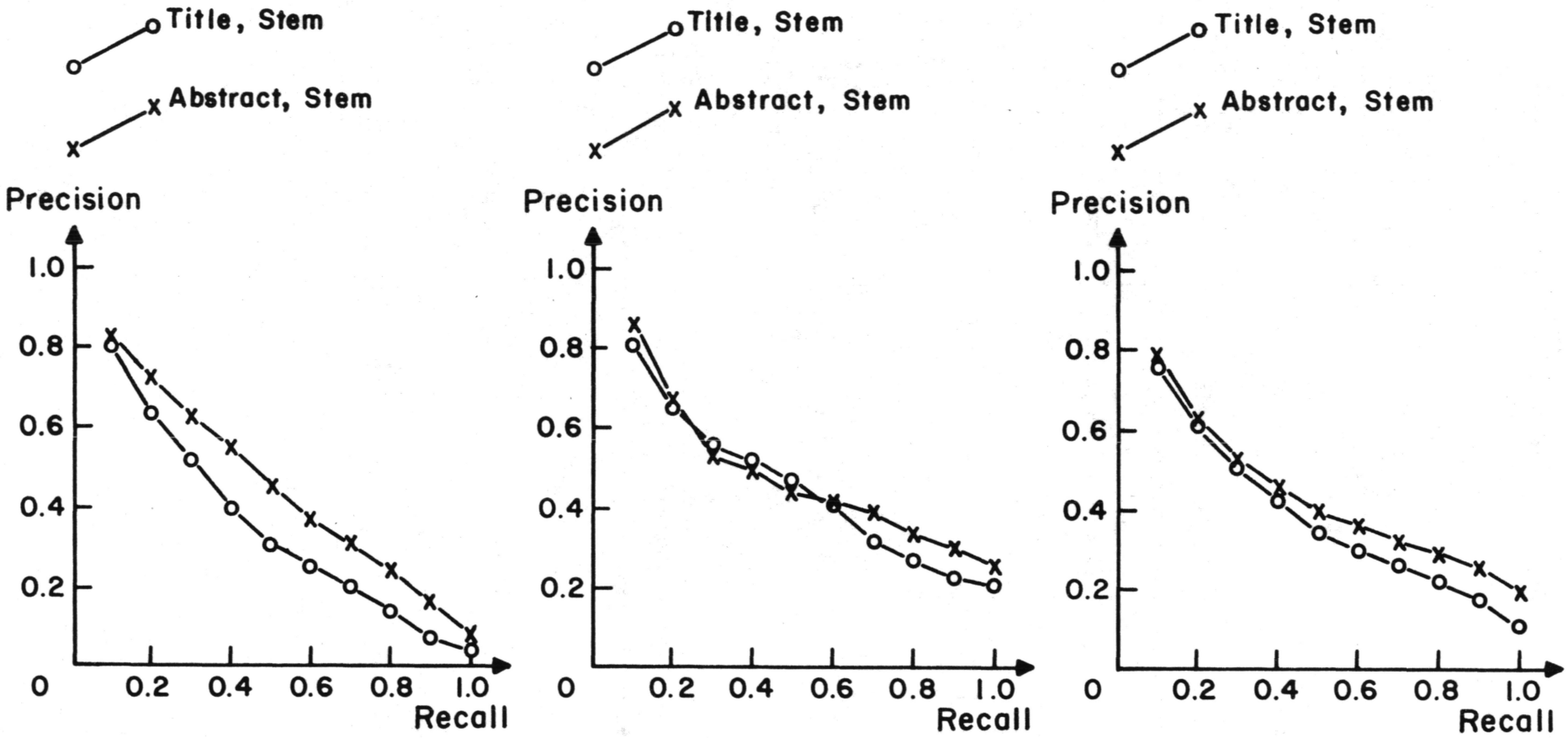
Overall performance measures are given in Figures 5, 6 and 7. Nine comparisons of abstracts and titles are presented using the normalized measures in Figure 5, the runs being made on different dictionaries using the cosine numeric matching algorithm for the three different collections. Every case shows the abstract to be superior to title, by as much as 0.0879 normalized recall in the case of ADI stem. Subsequent presentations will concentrate on the stem and thesaurus dictionaries only, for these three collections. Precision/recall graphs using stem are given in Figure 6. Title is slightly superior to abstract between 0.25 and 0.55 recall on the Cran-1 collection, otherwise the abstracts are always superior. Figure 7 repeats the comparison using thesaurus dictionaries, and in this case the abstract is superior to the title on Cran-1 over the whole curve, but on ADI the title is slightly superior to the abstract at low recall values. These graphs show that for the IRE-3 collection, abstract is always clearly superior to title, but on the ADI and Cran-1 collection the title is sometimes as good as the abstract in the low recall/high precision region.

Before presenting the recall ceiling data for these results, some explanations are necessary. For purposes of the experimental tests, requests are searched in the system and every single document in the collection is correlated with the request and is given a rank position in the output list. No cut-off is used to "retrieve", say, half the collection, since a cut-off might be made at any level by a user when he examines the output. In a real-life situation, it will be a rare thing for a user to examine documents that have a very low correlation with the request, and it seems certain that users would never examine documents with zero correlation; indeed, willingness to examine such would remove the need for the retrieval system altogether.

COLLECTION	DICTIONARY	EVALUATION MEASURE	ABSTRACT	TITLE
IRE-3 34 Requests	Stem	Normed. Recall	.8954	.8145
		Normed. Precision	.6746	.5547
	Thesaurus-2 (Harris 2)	Normed. Recall	.9191	.8436
		Normed. Precision	.7072	.5945
	Thesaurus-3 (Harris 3)	Normed. Recall	.9268	.8430
		Normed. Precision	.7382	.6068
CRAN-1 42 Requests	Stem	Normed. Recall	.8644	.8112
		Normed. Precision	.6704	.6185
	Thesaurus-3 (Harris)	Normed. Recall	.8833	.8374
		Normed. Precision	.6955	.6420
ADI 35 Requests	Suffix 's'	Normed. Recall	.7253	.6435
		Normed. Precision	.4997	.4209
	Stem	Normed. Recall	.7601	.6722
		Normed. Precision	.5326	.4537
	Thesaurus-1 (Harris)	Normed. Recall	.8016	.7324
		Normed. Precision	.6069	.5462
	Thesaurus-2 (Hastie)	Normed. Recall	.7548	.6877
		Normed. Precision	.5190	.4649

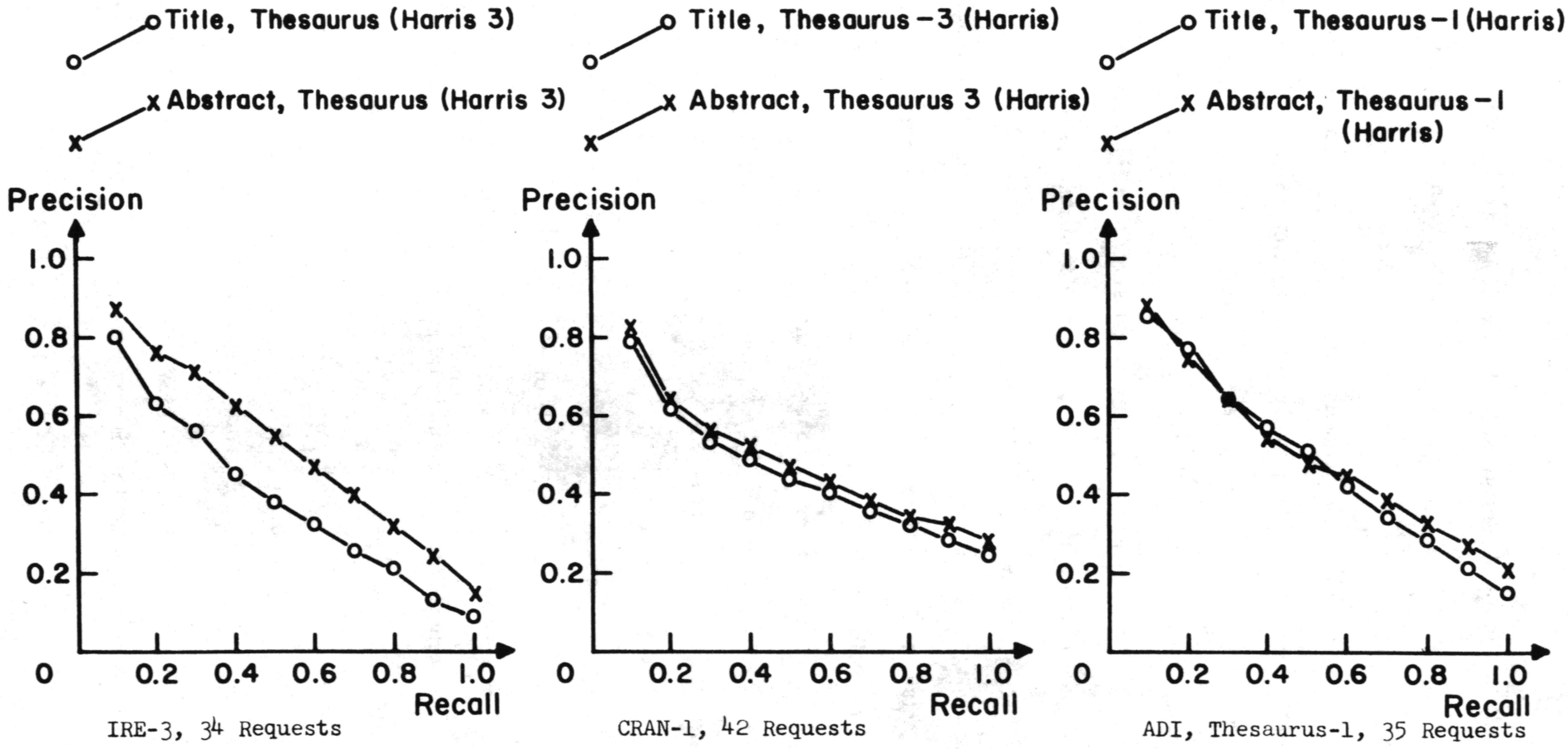
Performance Results comparing abstracts with titles for nine dictionaries on three collections, using normalized recall and precision.

Figure 5.



Performance Results Comparing Abstracts with Titles on Three Collections, Stem Dictionary, using Precision Versus Recall Graphs

Fig. 6



Performance Results Comparing Abstracts with Titles on Three Collections, Thesaurus Dictionary, using Precision Versus Recall Graphs.

Fig. 7.

In the experimental tests, documents with zero correlation are also given rank positions, (although, very low ones, in order that the normalized measures may be calculated, and also so that precision/recall curves may be drawn right up to 1.0 recall.) A statement was made earlier suggesting that short documents will present a barrier to perfect recall, because such short document identifications are likely in some cases completely to miss something important from the original full text of the document, thus resulting in a zero match between the search request and such a relevant document. Such an occurrence will cause recall loss, and the resulting recall ceiling will obviously be lower for short documents than long ones. The data of Figure 8 give results comparing abstracts and titles in six tests. The average recall ceiling is computed by accepting only those relevant documents with some positive correlation with the search requests; the recall ceiling with titles is seen to go down to .66 in one case. The recall ceiling for both abstracts and titles would in practice be lower than the values given, since many users would not be willing to examine all the documents with positive correlations (this would involve examining in the ADI Abstracts Collection, on average, 70% of the total collection). Comparing the results of Figure 8 with the data in Figure 1, the greater length of the abstracts and titles on aerodynamics over Documentation produces very slightly higher recall ceiling results, but the average length abstracts and titles on computer science give quite superior recall ceiling results. The conclusion is that for users needing high recall, titles only will not usually be adequate, and something nearer abstract length is required.

Since the results presented so far are all averages, and use the arithmetic means over the request sets, data are given in Figures 9, 10, 11, and 12 that are based on the individual requests and individual relevant

COLLECTION	INPUT AND DICTIONARY	TOTAL RELEVANT (ALL REQUESTS)	TOTAL DOCUMENTS WITH ZERO CORRELATION WITH REQUEST	AVERAGE* RECALL CEILING
IRE-3 34 Requests	Abstract Stem	592	23	.96
	Title Stem	592	129	.78
	Abstract Thesaurus-3	592	14	.98
	Title Thesaurus-3	592	92	.84
CRAN-1 42 Requests	Abstract Stem	198	17	.91
	Title Stem	198	59	.70
	Abstract Thesaurus-3	198	10	.95
	Title Thesaurus-3	198	44	.78
ADI 35 Requests	Abstract Stem	170	15	.91
	Title Stem	170	58	.66
	Abstract Thesaurus-1	170	14	.92
	Title Thesaurus-1	170	49	.71

*Computed using the aggregate recall ("micro" evaluation).

Average recall ceiling figures comparing abstract and titles searches on three collections each with two dictionaries.

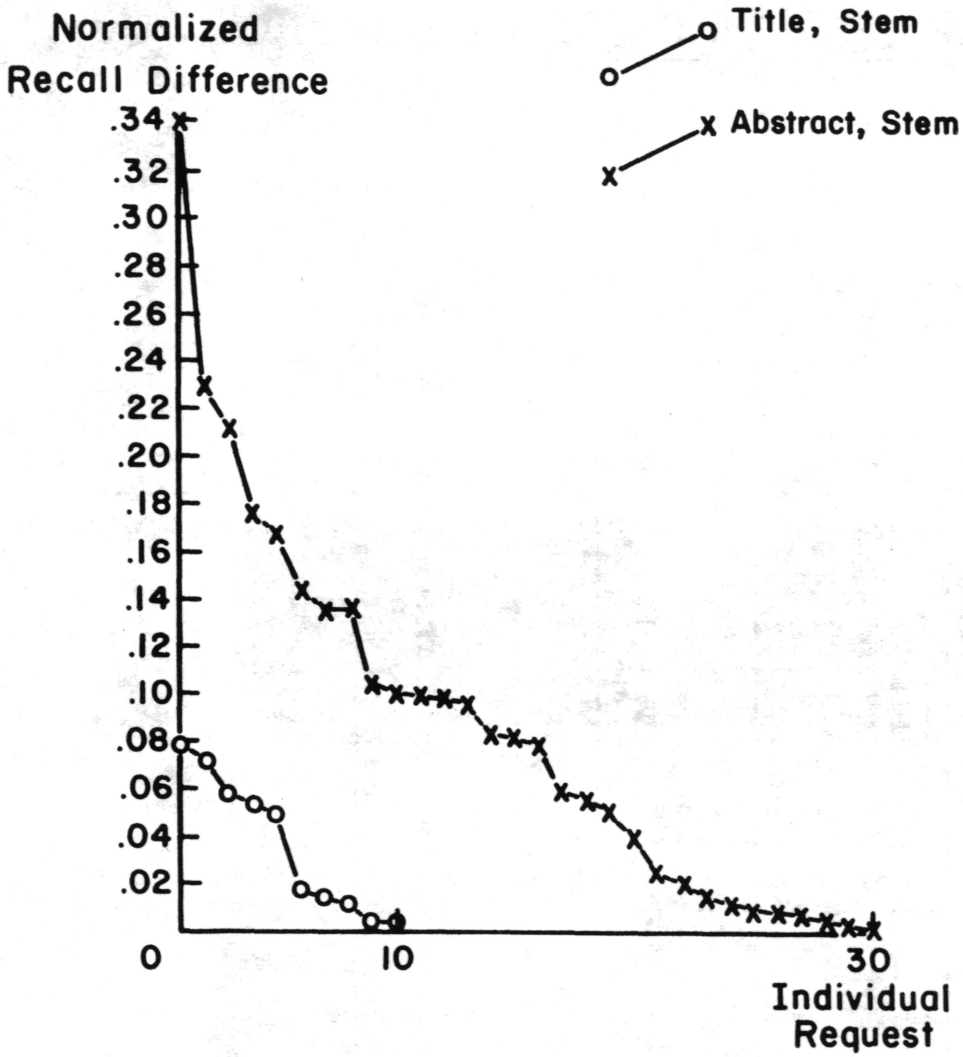
Fig. 8.

COLLECTION	DICTIONARY	EVALUATION MEASURE USED TO DETERMINE MERIT	NUMBER AND PERCENTAGE* OF INDIVIDUAL REQUESTS		
			ABSTRACT SUPERIOR	TITLE SUPERIOR	BOTH EQUAL
IRE-3 34 Requests	Stem	Normed. Recall	29 (85.3%)	5 (14.7%)	0
		Normed. Precision	28 (82.4%)	6 (17.6%)	0
	Thesaurus-3	Normed. Recall	30 (88.2%)	4 (11.8%)	0
		Normed. Precision	32 (94.1%)	2 (5.9%)	0
CRAN-1 42 Requests	Stem	Normed. Recall	30 (75.0%)	10 (25.0%)	2
		Normed. Precision	30 (75.0%)	10 (25.0%)	2
	Thesaurus-3	Normed. Recall	26 (66.7%)	13 (33.3%)	3
		Normed. Precision	22 (57.9%)	16 (42.1%)	4
ADI 35 Requests	Stem	Normed. Recall	24 (72.7%)	9 (27.3%)	2
		Normed. Precision	24 (72.7%)	9 (27.3%)	2
	Thesaurus-1	Normed. Recall	25 (75.8%)	8 (24.2%)	2
		Normed. Precision	26 (78.8%)	7 (21.2%)	2

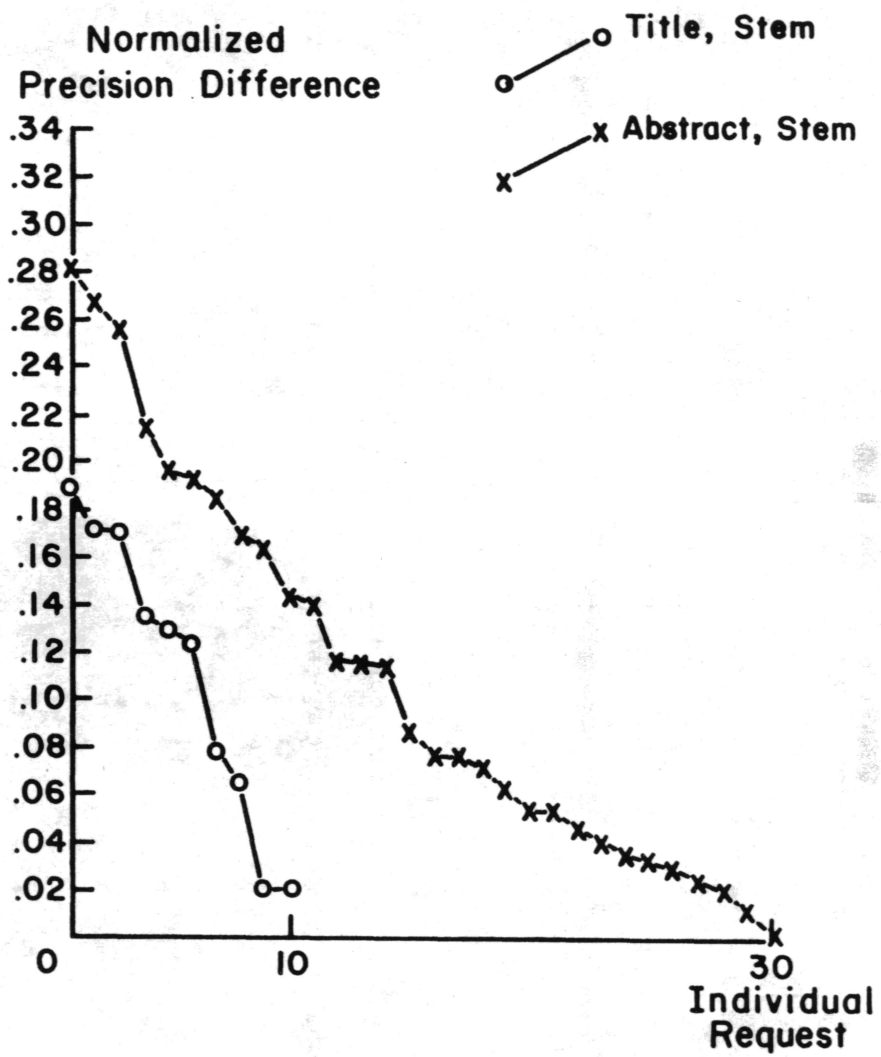
* Percentages exclude requests when both equal

Comparison of individual request merit, giving numbers and percentages of requests favoring abstracts and titles on three collections each using two dictionaries, according to merit assigned by the normalized evaluation measures.

Figure 9.



30 Requests superior on Abstracts by amounts on plot
10 Requests superior on Titles, by amounts on plot
2 Requests equal on both



30 Requests superior on Abstracts, by amounts on plot
10 Requests superior on Titles, by amounts on plot
2 Requests equal on both

CRAN-1, 42 Requests, Stem Dictionary
Graphs of the Magnitudes of the Differences of Individual Requests Comparing Abstracts and Titles, using the Normalized Evaluation Measures

Fig. 10

COLLECTION	DICTIONARY	NUMBER OF RELEVANT DOCUMENTS WITH RANK POSITIONS SHOWING: ABSTRACT SUPERIOR	TITLE SUPERIOR	BOTH EQUAL	TOTAL DOCUMENTS
CRAN-1	Stem	99	84	15	198

Rank position merit of the 198 individual documents relevant to the 42 requests comparing abstract and title.

Figure 11.

	NUMBERS OF RELEVANT DOCUMENTS WITH RANK CHANGES IN RANGES: 1-5	6-10	11-20	21-30	31-40	41-50	51-75	76-100	101-125	126-150	
Abstract Superior	29	17	12	5	2	2	4	20	6	2	Total 99
Title Superior	28	16	15	11	7	3	4	0	0	0	Total 84

Changes in rank positions between abstracts and titles of 183 of the individual documents relevant to the 42 requests.(Cran-1 Collection, Stem Dictionary)

Figure 12.

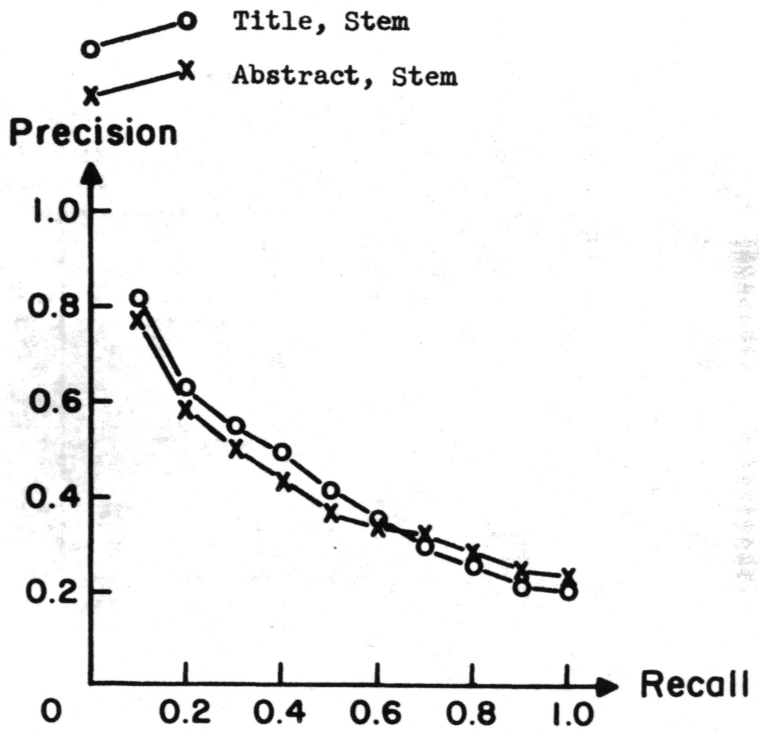
documents. Figure 9 gives the numbers of requests that favor abstracts and the number that favor titles, using both normalized recall and normalized precision, for six results. The data given reflect the fact that the precision/recall curves in Figures 6 and 7 closely represent the actual situation, namely, that abstracts are superior to titles, since between 57.9% and 94.1% of the requests favor abstracts on the six runs, using normalized recall (ties being ignored). The superiority of abstracts is again most evident with the computer science collection, and least so in the aerodynamics collection. Since the aerodynamics result in Figure 6 produces a crossing curve, two plots are given in Figure 10 of the normalized recall and normalized precision values for each of the 42 requests, showing the magnitude of the differences, comparing the 30 requests that favor abstracts and the 10 that favor titles. For example, one request had a normalized recall difference of 0.34 between abstracts and titles, while another request was better by 0.08 on titles than abstracts. The requests are arranged in an order of decreasing difference, and it is seen, using both normalized recall and normalized precision, that although ten requests did perform better on titles, there are ten requests that performed better on abstracts with a larger increase in performance. This result does not explain the superiority of titles over a small range in the middle of the precision recall curve seen in Figure 6 b, so that further data are given in Figures 11 and 12 to explain this fact. In these tables, the individual relevant documents are examined, and the ranks of the 198 documents concerned are compared on abstracts and titles. Figure 11 shows that 99 are superior on abstracts, and 84 on titles, a close result that accurately describes the situation. Figure 12 further breaks down these 99 and 84 documents, showing by a series of 10 ranges, the difference in rank positions for the 99 superior on abstracts, and the 84 superior

on titles. It can be seen that 28 of the documents superior on abstracts have improved rank positions (compared with titles) by 76 to 150 places, thus explaining why many of the requests do work better on abstracts. Since a large number of documents exhibit quite significant improvements in rank on titles compared with abstracts, however, the results that show superiority of titles in the middle of the precision recall curve seem quite reasonable.

The results presented so far have all been based on the cosine correlation and numeric vector matching procedure, which is generally superior to simpler procedures. Results are given in Figures 13 to 16 based on unweighted vectors (logical) using the overlap correlation, comparing titles and abstracts with the stem, Cranfield collection. For this process, the title match is superior by a small amount at the high precision end of the curve, below 0.65 recall, and this result is also reflected in the normalized measures (Figure 13). This same precision superiority is seen in the number of requests favoring abstracts and titles in Figure 14, where using normalized recall, the abstracts are superior, but using normalized precision the titles do better. The difference curve also given in Figure 14 shows that using normalized precision all but 2 of the 24 requests performing better with titles do so by a greater difference than the 18 which are better on the abstracts. Figures 15 and 16 give data for the 198 individual relevant documents involved, showing that 13 relevant documents changed rank by over 100 places in favor of abstracts, but that more documents changed a smaller number of places in favor of titles than abstracts.

These figures are presented in order to show that there is no inconsistency between the results on abstracts and titles obtained with SMART and those obtained on the Cranfield Project [2]. The results of searches on the same titles and abstracts, using the same collection, requests and relevance

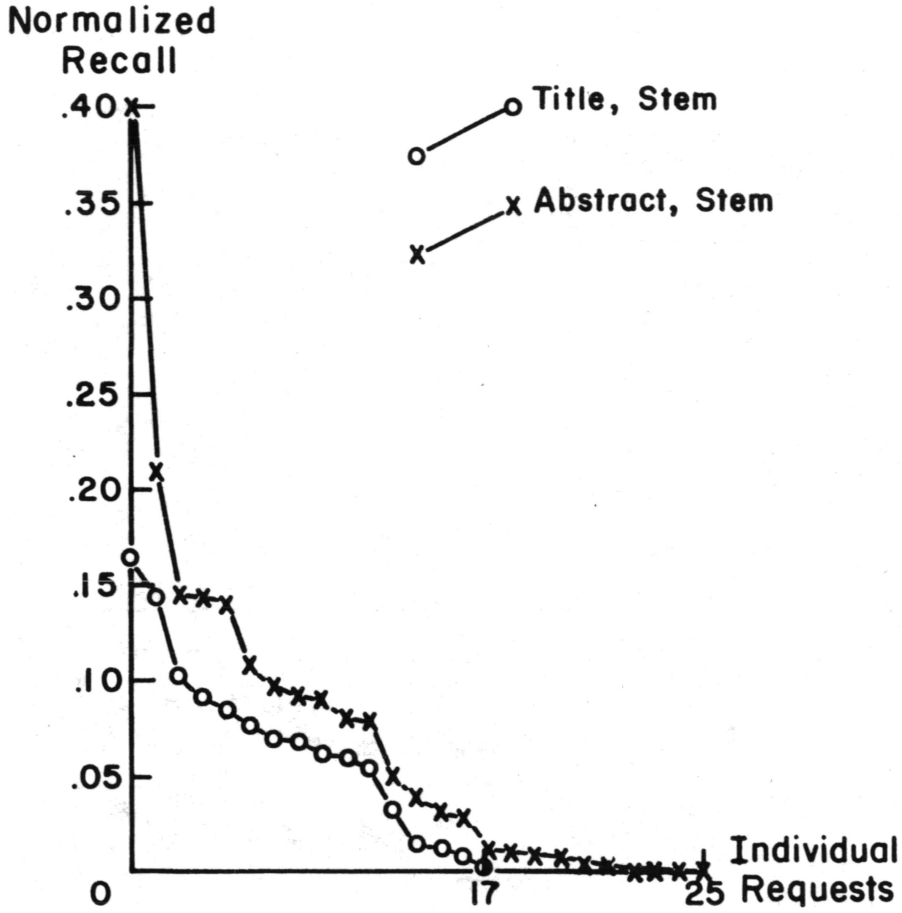
EVALUATION MEASURE	ABSTRACT	TITLE
Normed. Recall	.8237	.8082
Normed. Precision	.5830	.5979



CRAN-1, 42 Requests.

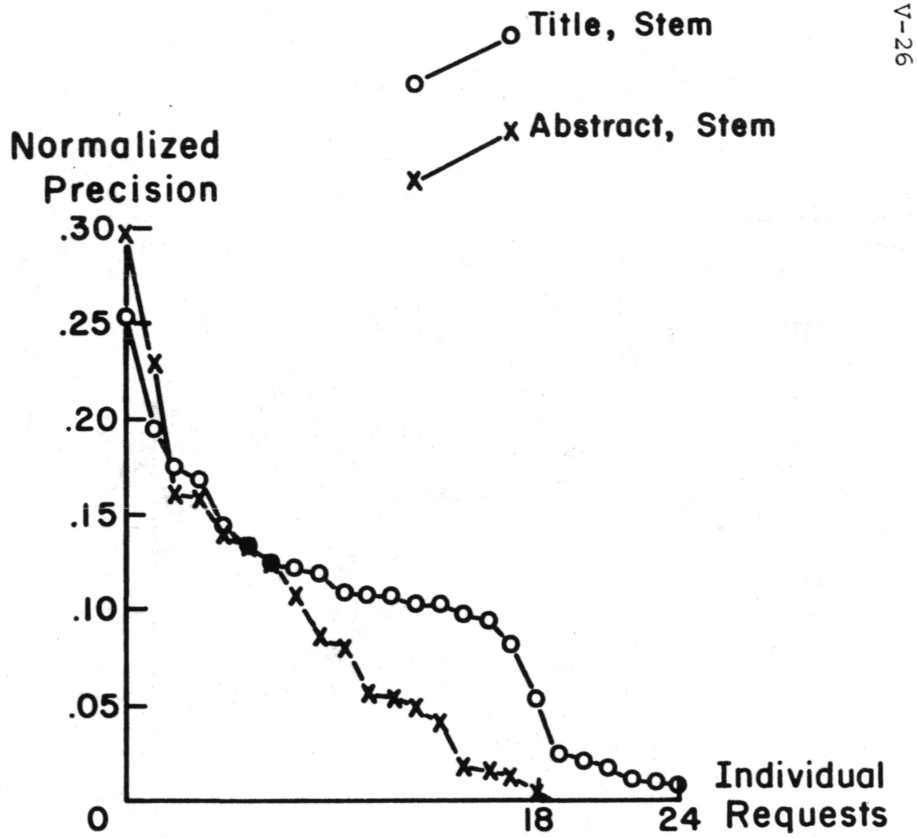
Performance Results of Abstracts Versus Titles using Overlap Correlation and Logical Vectors, Stem Dictionary, CRAN-1 Collection.

Fig. 13



25 Requests superior on abstracts by amounts on plot

17 Requests superior on title, by amounts on plot



18 Requests superior on abstracts, by amounts on plot

24 Requests superior on titles, by amounts on plot

CRAN-1, 42 Requests, Stem Dictionary, Overlap Correlation, Logical Vectors.

Plots of the Magnitudes of the Differences of Individual Requests Comparing Abstracts and Titles, using the Normalized Evaluation Measures.

Fig. 14

COLLECTION	DICTIONARY	NUMBER OF RELEVANT DOCUMENTS WITH RANK POSITIONS SHOWING: ABSTRACT SUPERIOR	TITLE SUPERIOR	BOTH EQUAL	TOTAL DOCUMENTS
CRAN-1	Stem (Overlap Logical)	89	103	6	198

Rank position merit of the 198 individual relevant documents, comparing abstract and title with overlap correlation and logical vectors.

Figure 15.

	NUMBERS OF RELEVANT DOCUMENTS WITH RANK CHANGES IN RANGES: 1-5	6-10	11-20	21-30	31-40	41-50	51-75	76-100	101-125	126-150	
Abstract Superior	25	11	6	6	5	4	7	12	11	2	Total 89
Title Superior	36	14	15	8	12	5	8	5	0	0	Total 103

Changes in rank positions between abstracts and titles of 192 of the individual relevant documents, Cran-1 Collection, Stem Dictionary, overlap correlation and logical vectors.

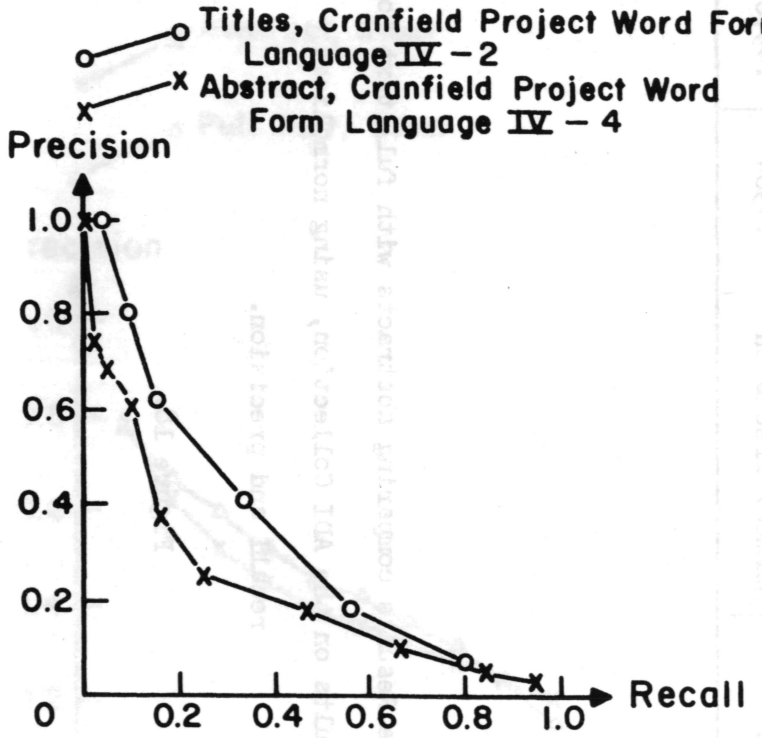
Figure 16.

judgment, and using coordination level matching (which is virtually identical to overlap-logical matching on SMART) obtained at Cranfield are given in Figure 17. The differences between test techniques in the case of SMART and Cranfield reside in the dictionaries used (Cranfield word forms language is similar, but not identical, to SMART stem dictionary), and also in the methods of calculating the average recall/precision curves. It is seen that this last matter is still a partly unsolved problem, since the two Cranfield plots presented are not totally consistent. Figure 17 b) comes closest to the methods used by SMART, and comparison with Figure 13 shows a similar result except at the low recall end. It would seem that the addition of a weighting scheme, as used in SMART, does not help the title performance much, but does improve the abstracts, so that in circumstances where such weighting may be practiced even the Cranfield results do show a reasonable superiority of abstracts over titles.

B) Abstracts versus Full Text

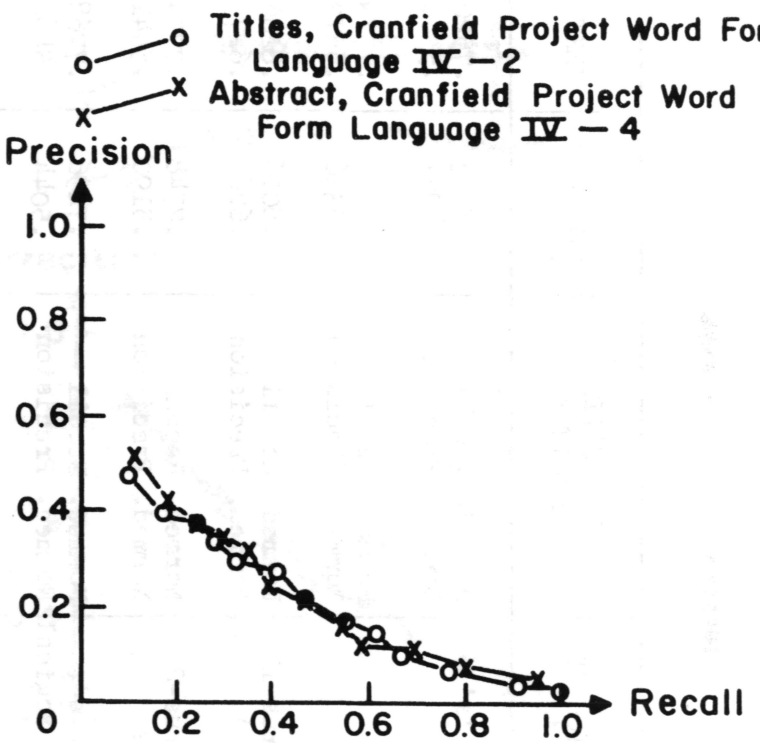
Overall performance measures are given in Figures 18 and 19. Seven comparisons of abstracts and full text are given in Figure 18 using the normalized measures, and two comparisons using precision/recall curves in Figure 19. In all cases the full text is superior to abstract alone, but the difference is always small. The precision/recall curves do cross over at the high recall end with stem and at the high precision end with thesaurus Figure 19), but this is due in the former case to the fact that 10% of the relevant documents have zero correlation with the request when abstracts are used, and the ranks assigned to these documents are higher than the ranks given when full text is in use.

The recall ceiling data are given in Figure 20, where it is seen that high ceilings are present, with the expected superiority of full text.



CRAN-1, 42 Requests

a) Plot constructed by keyword matching (coordination levels).



CRAN-1, 42 Requests

b) Plot constructed by a simulated ranking method and a document output cut-off.

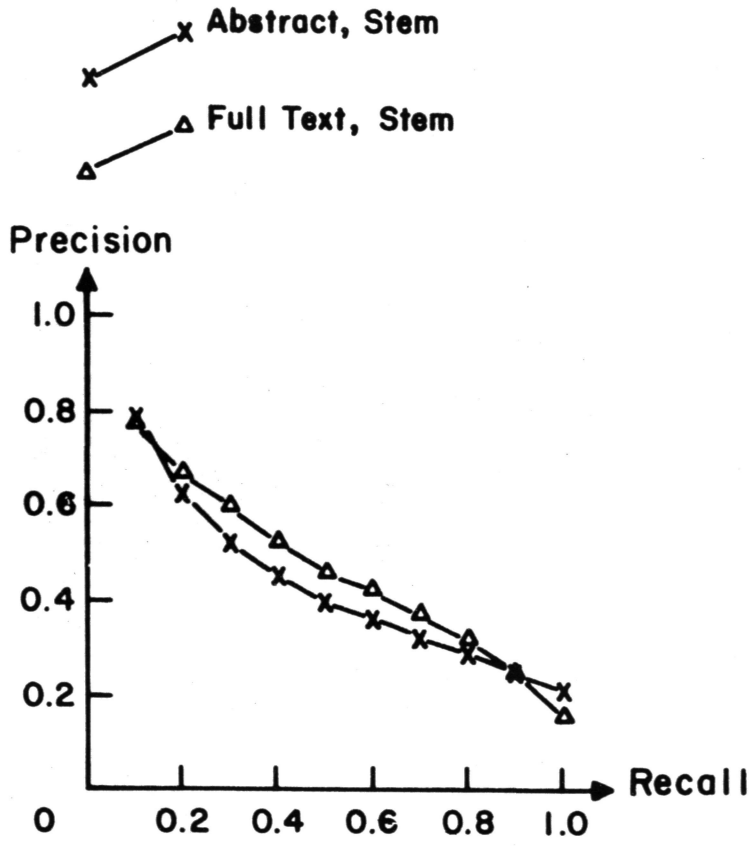
Cranfield Project Results Comparing Abstracts and Titles, using Precision Versus Recall Plots Constructed by Two Cut-off Methods.

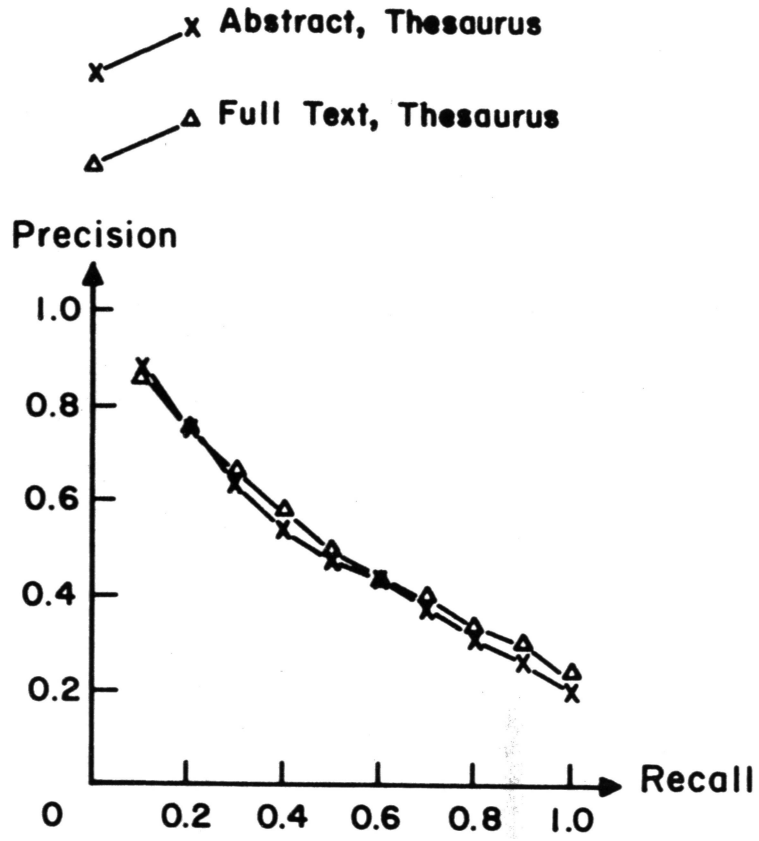
Fig. 17.

COLLECTION	DICTIONARY	EVALUATION MEASURE	ABSTRACT	TEXT
ADI 35 Requests	Suffix 's'	Normed. Recall	.7253	.7520
		Normed. Precision	.4997	.5308
	Stem	Normed. Recall	.7601	.7779
		Normed. Precision	.5326	.5573
	Thesaurus-1 (Harris)	Normed. Recall	.8016	.8206
		Normed. Precision	.6069	.6273
	Thesaurus-2 (Hastie)	Normed. Recall	.7548	.7774
		Normed. Precision	.5190	.5441
	Suffix 's', Cosine Logical	Normed. Recall	.7296	.7768
		Normed. Precision	.5044	.5462
	Stem, Cosine Logical	Normed. Recall	.7546	.7695
		Normed. Precision	.5221	.5248
	Stem, Overlap Logical	Normed. Recall	.7423	.7434
		Normed. Precision	.4904	.4978

Performance Results comparing abstracts with full text for seven results on the ADI Collection, using normalized recall and precision.

Figure 18.





ADI, 35 Requests

Performance Results Comparing Abstracts and Full Text on Two Dictionaries.

Fig. 19

COLLECTION	INPUT AND DICTIONARY	TOTAL RELEVANT ALL REQUESTS	TOTAL DOCUMENTS WITH ZERO CORRELATION WITH REQUEST	AVERAGE* RECALL CEILING
ADI 35 Requests	Abstract Stem	170	15	.91
	Text Stem	170	0	1.00
	Abstract Thesaurus-1	170	14	.92
	Text Thesaurus-1	170	1†	.99†

* Computed using the aggregate recall ("micro" evaluation).

† The one document having zero correlation with the thesaurus does correlate with request concepts "technique" and "system" in the stem dictionary, but these terms are in the common words list for the thesaurus dictionary.

Average recall ceiling figures comparing abstract and full text searches on two dictionaries.

Figure 20.

Figures 21 and 22 show, respectively, the number of requests favoring abstract and text using two dictionaries, and magnitude difference plots for the stem dictionary, since stem favors abstracts more than text in Figure 21, using normalized precision.

The differences between text and abstract are always small, and usually in favor of text. The precision/recall curves for titles only are added to those abstract and text in Figure 23; the data on individual requests in Figure 24 comparing the three document lengths again shows the expected order of merit. Data for the 170 relevant documents concerned are given in Figure 25. Taking results of the six possible orders of merit for the three document lengths, it is interesting to note that merit orders "A" and "F" are observed for more documents than any of the other orders of merit. Documents in A are clearly matched poorly with the request using titles, and the two increases in length improve the match and rank positions of the 47 documents concerned. Documents in F probably match the requests quite well on titles, and increases in document length only serve to increase the matches with non-relevant documents, thus worsening the ranks of these 36 relevant documents. The abstracts came off worse by this evaluation, but text is best for many relevant documents.

Retrieval runs using full text were also made without the abstracts, although the title was always included. In the results presented here text includes abstract, and this change does provide a slight improvement in performance as the normalized measures in Figure 26 show.

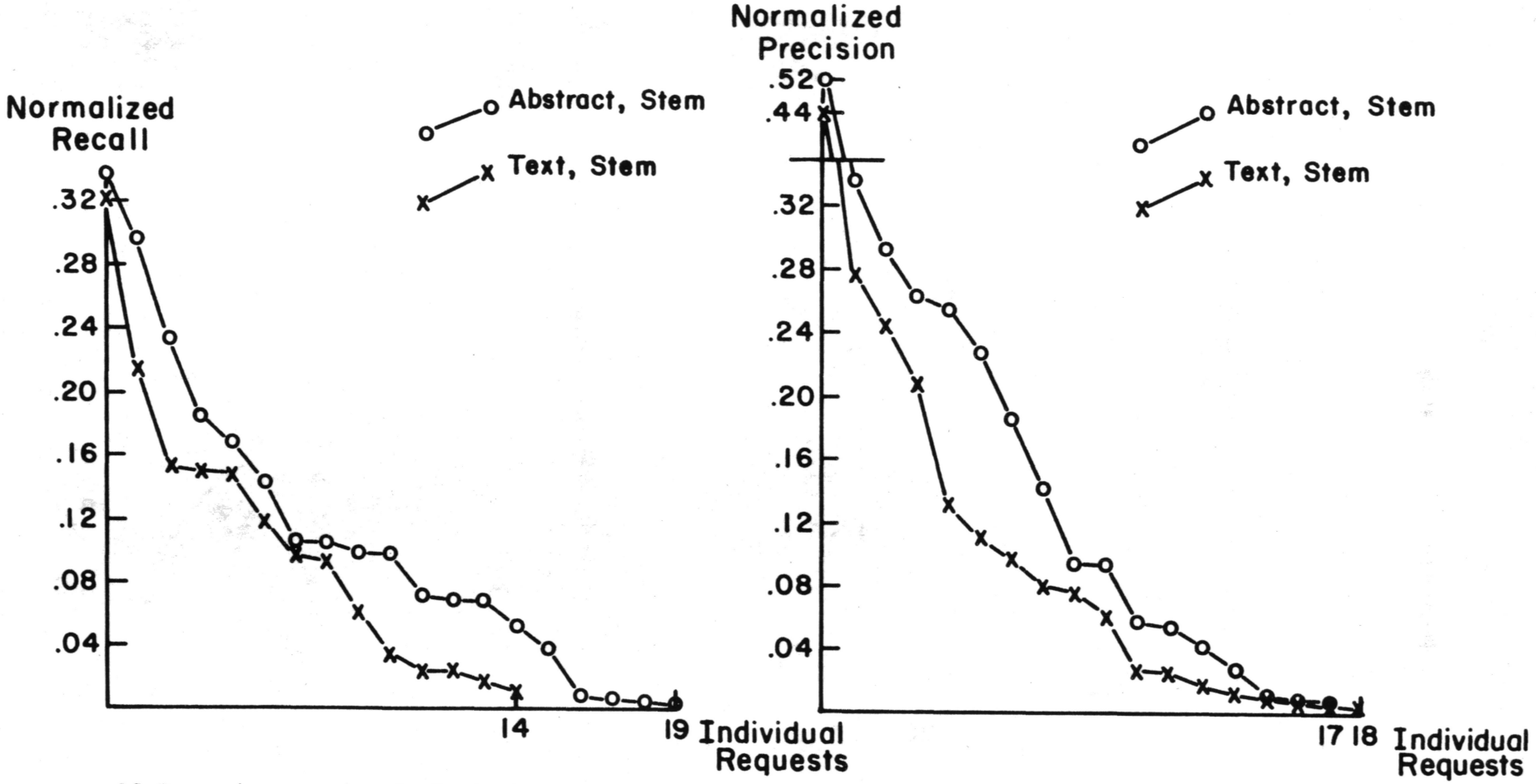
Despite this outcome, the ADI abstracts are thought to be rather poor; some are rather short, and do not seem adequately to cover the text for document retrieval purposes. It is suggested that if better abstracts were available they might have a superior performance (apart from recall ceiling) to full text.

COLLECTION	DICTIONARY	EVALUATION MEASURE USED TO DETERMINE MERIT	NUMBER AND PERCENTAGE* OF INDIVIDUAL REQUESTS		
			TEXT SUPERIOR	ABSTRACT SUPERIOR	BOTH EQUAL
ADI 35 Requests	Stem	Normed. Recall	19 (57.6%)	14 (42.4%)	2
		Normed. Precision	17 (48.6%)	18 (51.4%)	0
	Thesaurus-1	Normed. Recall	20 (58.8%)	14 (41.2%)	1
		Normed. Precision	21 (61.8%)	13 (38.2%)	1

* Percentages exclude requests where both equal.

Comparison of individual request merit, giving the numbers and percentages of requests favoring text and abstracts on the ADI collection, according to merit assigned by the normalized evaluation measures.

Figure 21.



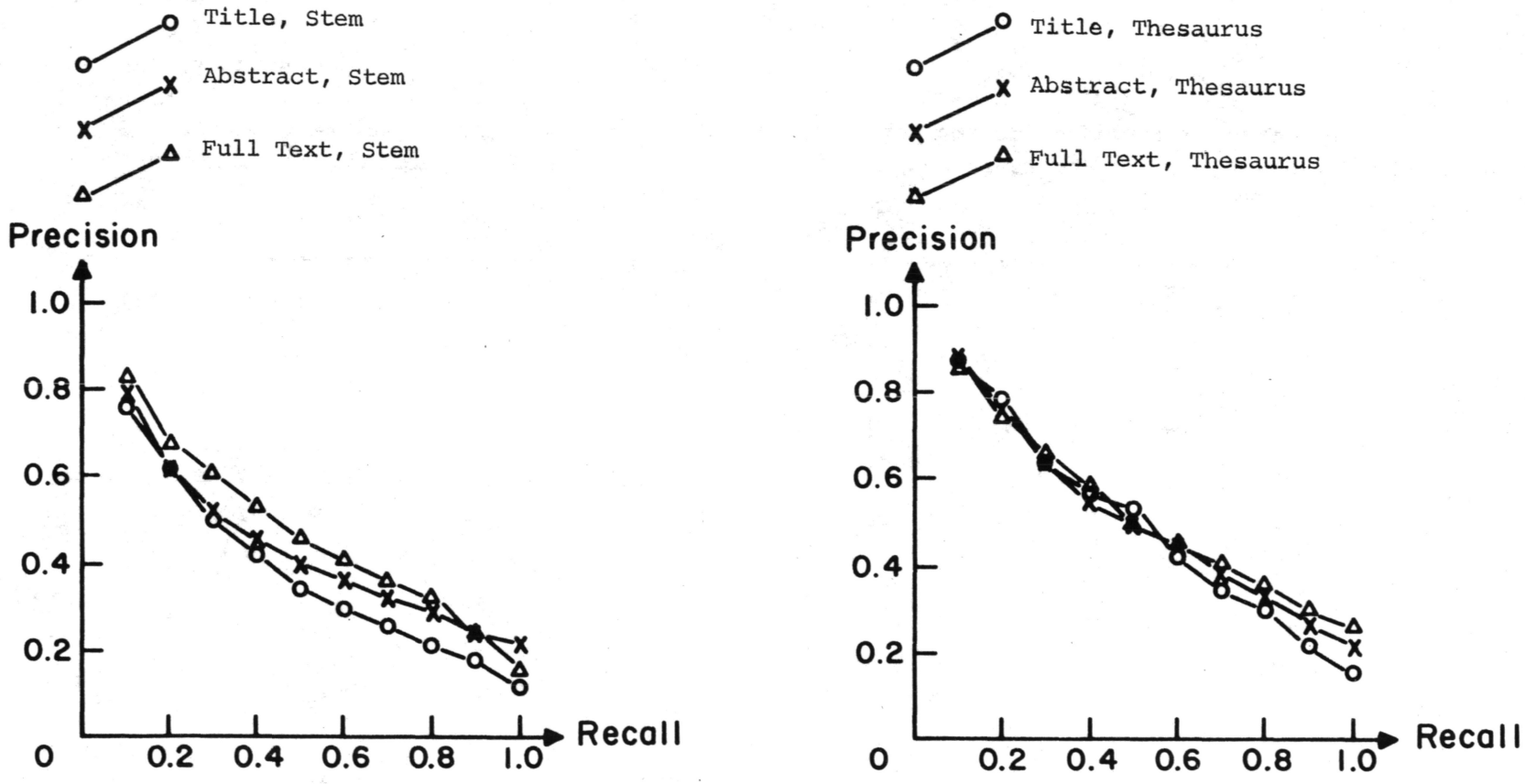
19 Requests superior in Text, by amounts on plot
14 Requests superior on Abstracts, by amounts on plot
1 Request equal on both

17 Requests superior on Text, by amounts on plot
18 Requests superior on Abstracts, by amounts on plot
1 Request equal on both

ADI, 35 Requests, Stem Dictionary.

Plot of the Magnitudes of the Differences of Individual Requests Comparing Full Text and Abstracts, using the Normalized Evaluation Measures.

Fig. 22.



ADI, 35 Requests

Performance Results Comparing Titles, Abstracts and Full Text on Two Dictionaries.

Fig. 23.

COLLECTION	DICTIONARY	EVALUATION MEASURE USED TO DETERMINE MERIT	NUMBER OF INDIVIDUAL REQUESTS				
			TEXT SUPERIOR	ABSTRACT SUPERIOR	TITLE SUPERIOR	TEXT, ABS. AND TITLE EQUAL	ABS. AND TITLE EQUAL
ADI 35 Requests	Stem	Normed. Recall	16 (47.1%)	13 (38.2%)	5 (14.7%)	0	1
		Normed. Precision	15 (44.1%)	14 (41.2%)	5 (14.7%)	0	1
	Thesaurus-1	Normed. Recall	17 (53.1%)	10 (31.3%)	5 (15.6%)	1	2
		Normed. Precision	17 (51.5%)	9 (27.3%)	7 (21.2%)	1	1

Comparisons of individual request merit giving the numbers of requests favoring full text, abstracts and titles, using two dictionaries, according to merit assigned by the normalized evaluation measures.

Figure 24.

ORDER OF MERIT OF DOCUMENT LENGTHS USING RANK POSITIONS WORST MIDDLE BEST		NUMBERS OF RELEVANT DOCUMENTS	
A	Title – Abstract – Text	47	74
B	Abstract – Title – Text	27	
C	Text – Title – Abstract	23	43
D	Title – Text – Abstract	20	
E	Abstract – Text – Title	17	53
F	Text – Abstract – Title	36	

Rank position merit of 170 individual documents relevant to 35 requests comparing title, abstract and full text, Stem dictionary, ADI Collection.

Figure 25.

COLLECTION	DICTIONARY	EVALUATION MEASURE	TEXT INCLUDES ABSTRACT	TEXT WITHOUT ABSTRACT
ADI 35 Requests	Stem	Normed. Recall	.7779	.7674
		Normed. Precision	.5573	.5538
	Thesaurus-1	Normed. Recall	.8206	.8136
		Normed. Precision	.6273	.6223

Performance Results comparing full text including abstract and full text without abstract, on two dictionaries.

Figure 26.

C) Abstracts versus Indexing

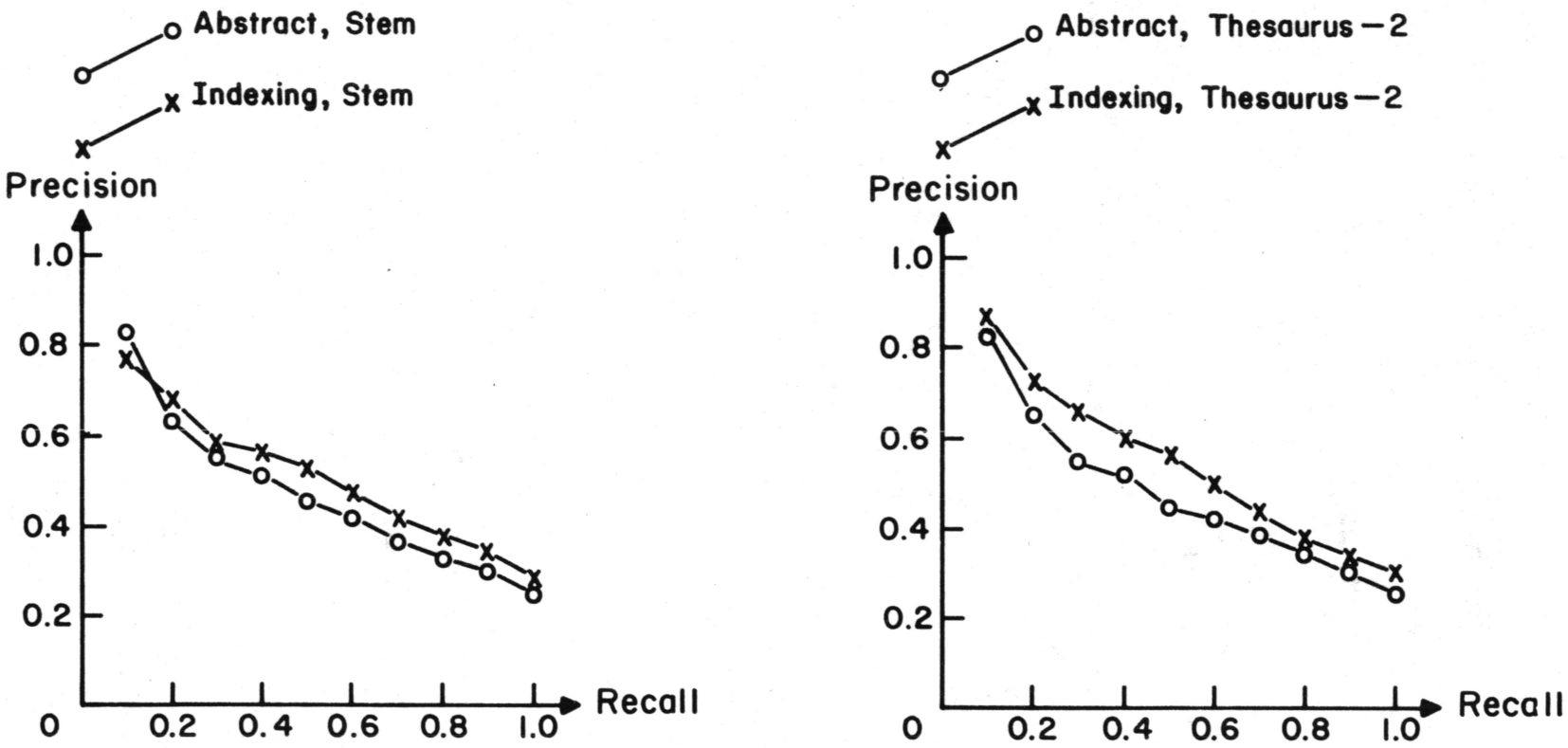
Overall performance measures are given in Figures 27 and 28. The indexing is in all cases superior to the abstracts, except in the stem dictionary at the very low recall end of the precision/recall curve (Figure 28 a). Indexing has a slightly superior recall ceiling also, as seen in Figure 29. The individual request data and difference plots in Figures 30 and 31 reinforce these results: between 51.3% and 64.1% of the requests are superior on indexing.

The superiority of the indexing is small but quite marked, and was observed to be similar in the tests conducted at Cranfield, see Figure 32. A positive explanation as to why the indexing is superior awaits analysis not yet performed, because the effects of two separate factors which differ between the indexing and abstracts cannot be distinguished. The first point relates to the fact that the indexers were free to choose terms out of the whole documents, so that it is expected that the indexing incorporates at least some subject notions that the abstractors did not include. The second factor is the one of primary interest here, namely document length (or indexing exhaustivity), which for the indexing was roughly half that of the abstracts. Both these factors may be observed in the results from the Cranfield Project, presented in Figure 33. The tables give search results at two coordination levels (corresponding to a demand of two matching keywords) for five different document lengths, the shortest being titles only, then three levels of exhaustivity of indexing, and finally the longest being the abstracts. The indexing results previously examined used the "Indexing 3" level. Figure 33 b) shows the indexing which probably included some ideas not in the abstract, since 5 additional relevant documents were found in the indexing as against the abstracts. The effect of document length is seen in

COLLECTION	DICTIONARY	EVALUATION MEASURE	ABSTRACT	INDEXING
CRAN-1 42 Requests	Stem	Normed. Recall	.8644	.8897
		Normed. Precision	.6704	.6831
	Thesaurus-1 (Old Q.S.)	Normed. Recall	.8602	.8629
		Normed. Precision	.6319	.6335
	Thesaurus-2 (New Q.S.)	Normed. Recall	.8864	.8992
		Normed. Precision	.6864	.7094

Performance results comparing abstracts with indexing, on three dictionaries, using normalized recall and precision.

Figure 27.



CRAN-1, 42 Requests

Performance Results Comparing Abstracts with Indexing, on Two Dictionaries.

Fig. 28.

COLLECTION	INPUT AND DICTIONARY	AVERAGE DOCUMENT LENGTH, re CONCEPTS PER DOCT.	TOTAL RELEVANT (ALL REQUESTS)	TOTAL DOCUMENTS WITH ZERO CORRELATION WITH REQUEST	AVERAGE* RECALL CEILING
CRAN-1 42 Requests	Abstract, Stem	61	198	17	.91
	Indexing, Stem	32	198	11	.94
	Abstract, Thesaurus-2	40	198	4	.98
	Indexing, Thesaurus-2	30	198	0	1.00

*Computed using aggregate recall ("micro" evaluation).

Average recall ceiling figures comparing abstracts and indexing on two dictionaries, Cran-1 Collection.

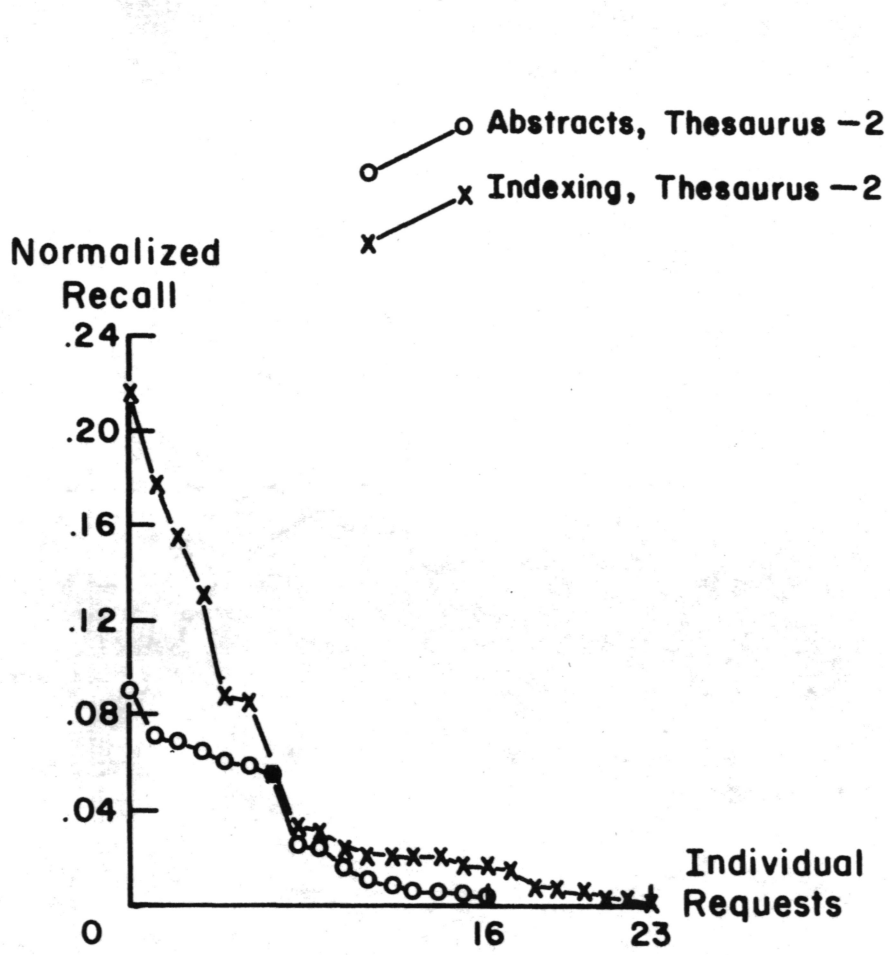
Figure 29.

COLLECTION	DICTIONARY	EVALUATION MEASURE USED TO DETERMINE MERIT	NUMBER AND PERCENTAGE* OF INDIVIDUAL REQUESTS		
			INDEXING SUPERIOR	ABSTRACT SUPERIOR	BOTH EQUAL
CRAN-1 42 Requests	Stem	Normed. Recall	25 (64.1%)	14 (35.9%)	3
		Normed. Precision	20 (51.3%)	19 (28.7%)	3
	Thesaurus-2	Normed. Recall	23 (59.0%)	16 (41.0%)	3
		Normed. Precision	23 (60.5%)	15 (39.5%)	4

* Percentages exclude requests where both equal.

Comparison of individual request merit giving the numbers and percentages of requests favoring indexing and abstracts on two dictionaries according to merit assigned by the normalized evaluation measures.

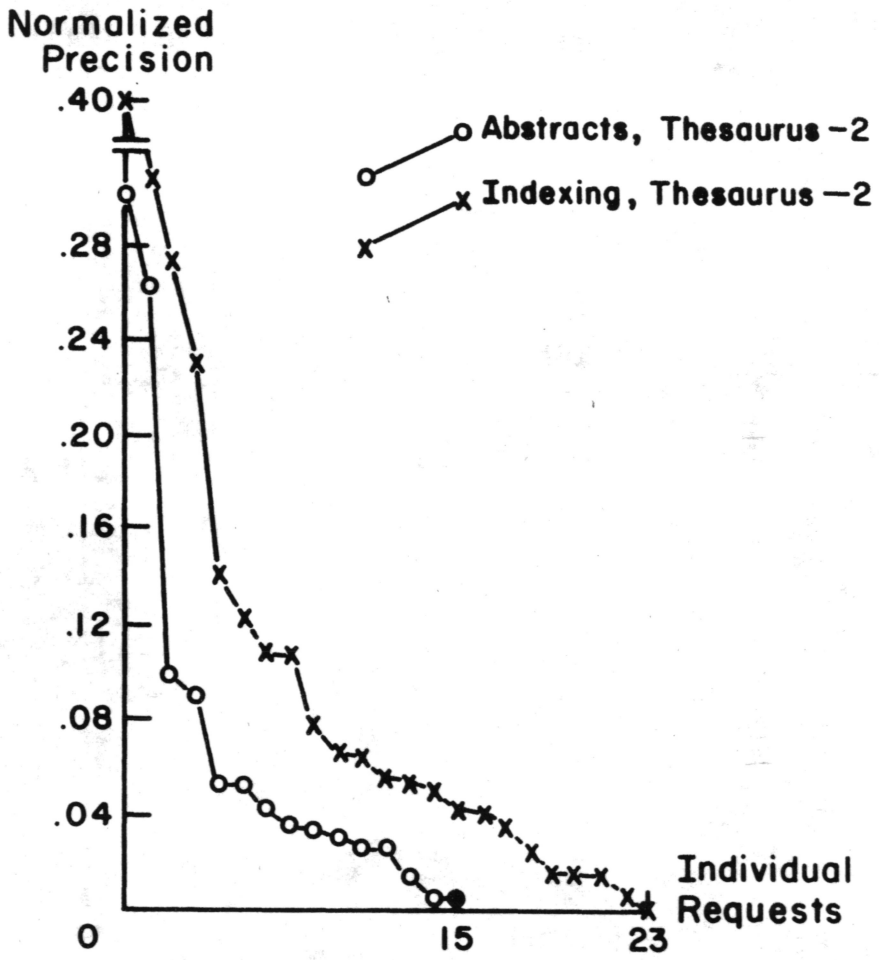
Figure 30.



23 Requests superior on Indexing, by amounts in plot

16 Requests superior on Abstracts, by amounts in plot

3 Requests equal on both



23 Requests superior on Indexing, by amounts in plot

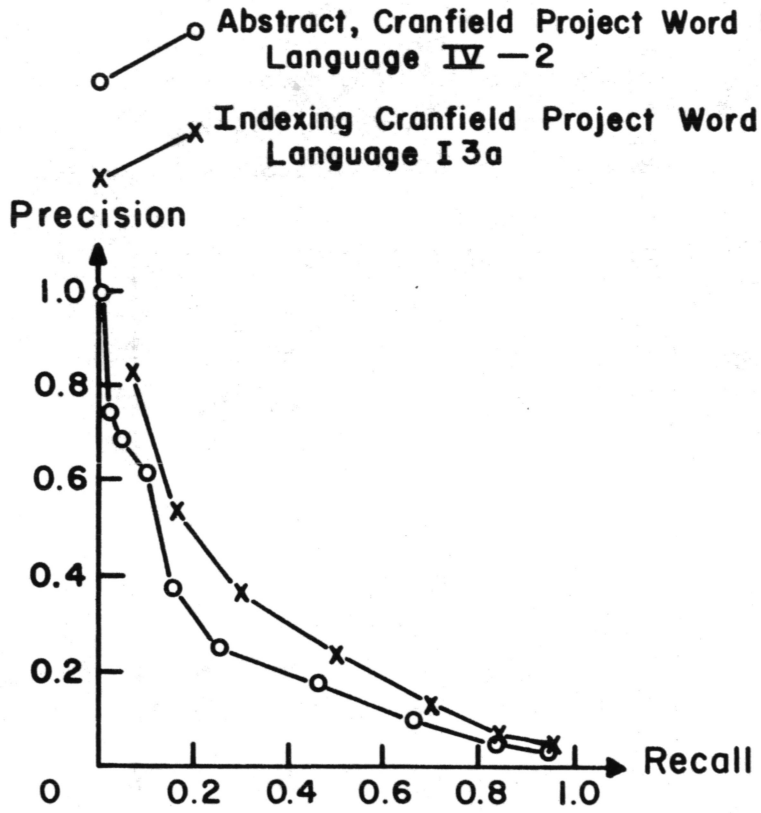
15 Requests superior on Abstracts, by amounts in plot

4 Requests equal on both

CRAN-1, 42 Requests, Thesaurus-2 (New Quasi-Synonym) Dictionary.

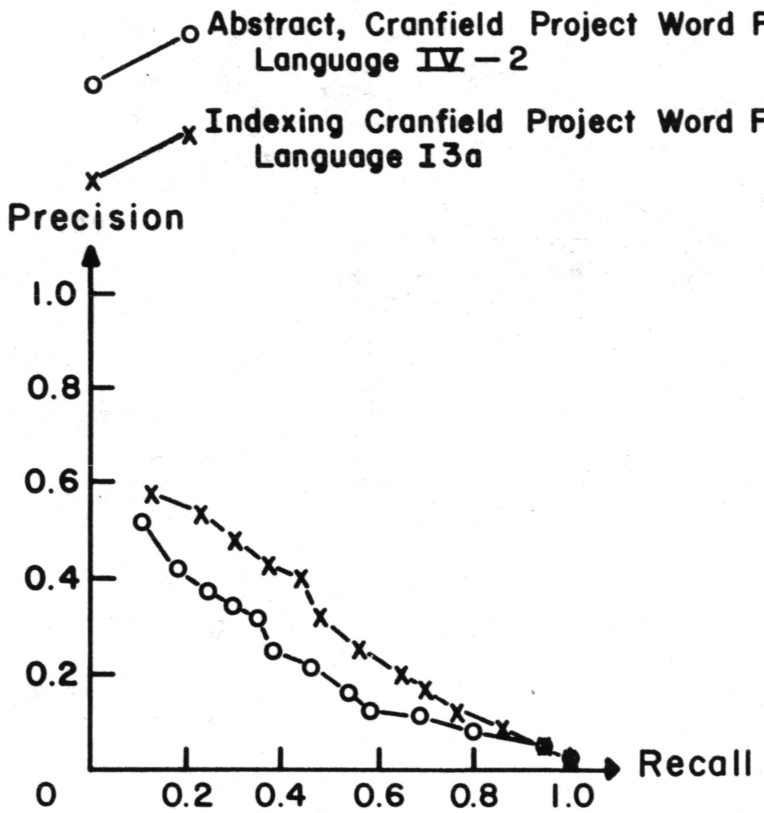
Plots of the Magnitudes of the Differences of Individual Requests Comparing Indexing and Abstracts, using the Normalized Evaluation Measures.

Fig. 31.



CRAN-1, 42 Requests

a) Plot constructed by keyword matching (coordination levels)



CRAN-1, 42 Requests

b) Plot constructed by a simulated ranking method and a document output cut-off

Cranfield Project Results Comparing Abstracts and Indexing, using Precision versus Recall Plots Constructed by Two Cut-off Methods.

Fig. 32

Co-ordination of Two Terms

INPUT TEXT	AVERAGE LENGTH	TOTAL DOCUMENTS RETRIEVED	RELEVANT DOCUMENTS RETRIEVED	NON-RELEVANT DOCUMENTS RETRIEVED	RECALL* RATIO	PRECISION* RATIO
Title	7	623	111	512	56.1%	17.8%
Indexing 1	14	1,239	138	1,101	69.7%	11.1%
Indexing 2	22	2,029	162	1,867	81.8%	8.0%
Indexing 3	33	2,381	166	2,221	83.8%	7.0%
Abstract	60	2,820	166	2,654	83.8%	5.9%

Co-ordination of Four Terms

INPUT TEXT	AVERAGE LENGTH	TOTAL DOCUMENTS RETRIEVED	RELEVANT DOCUMENTS RETRIEVED	NON-RELEVANT DOCUMENTS RETRIEVED	RECALL* RATIO	PRECISION* RATIO
Title	7	47	29	18	14.6%	61.7%
Indexing 1	14	140	60	80	30.3%	42.9%
Indexing 2	22	295	87	208	43.9%	29.5%
Indexing 3	33	412	97	315	49.0%	23.5%
Abstract	60	524	92	432	46.5%	17.6%

* These ratios are computed using the aggregates ("micro" evaluation).

Cranfield Project results comparing titles, abstracts, and indexing at three levels of exhaustivity, using search term co-ordination of two and four terms, Word Form Language I3a.

Figure 33.

Figure 33 a) and b), where it is seen that for a similar recall ratio the abstracts retrieve many more non-relevant documents than the indexing. These results suggest that the abstracts are too long (too exhaustive), for the particular requests and environment of this test, compared to the indexing. However, the abstract searches are not so inferior to the indexing that the use of abstracts could not be considered for an operational system; indeed to a system manager the loss of performance due to use of abstracts might be lessened by use of some selective effort at the input stage, thus resulting in a very acceptable substitute for the effort of complete manual indexing.

5. Individual Requests and Discussion of Results

Some data on individual requests is presented in order to support and illustrate the average results already given.

Comparing abstracts and titles, Figure 34 gives results for four cases, each case corresponding to a different request/relevant document pair, using results of the Cran-1 collection. Case A gives an example where the abstract provides two more matching terms and a better retrieval performance than the title, but in case B the greater match achieved by the abstract results in a worse performance for the abstract compared with the title. The latter case may be explained by remembering that the use of abstracts provides on average more matching terms between the requests and many of the documents in the collection; the reason for the case B result is that many non-relevant documents achieve better improvements in matching on abstracts compared with titles than the relevant document number 713 in question. In cases C and D, the abstract searches do not provide additional matching concepts, although the weights are increased on abstracts. In case C the abstract provides superior retrieval to the title, and in Case D the opposite result is seen to hold.

CASE	REQUEST DATA	RETRIEVAL CHARACTERISTICS	ABSTRACT	TITLE
A	Request Q266, 6 Request Concepts, Relevant Document 965	Number of matching concepts	3	1
		Sum of doct. wts. of matching concepts	7	1
		Cosine numeric correlation	.2619	.1666
		Rank position	9	27
B	Request Q122, 9 Request Concepts, Relevant Document 713	Number of matching concepts	3	2
		Sum of doct. wts. of matching concepts	6	2
		Cosine numeric correlation	.1383	.1571
		Rank position	47	25
C	Request Q250, 4 Request Concepts, Relevant Document 36G	Number of matching concepts	1	1
		Sum of doct. wts. of matching concepts	3	1
		Cosine numeric correlation	.2022	.1212
		Rank position	3	15
D	Request Q116, 13 Request Concepts, Relevant Document 574	Number of matching concepts	3	3
		Sum of doct. wts. of matching concepts	7	3
		Cosine numeric correlation	.1277	.2631
		Rank position	27	5

Four cases of request/relevant document analysis comparing abstracts and titles, stem dictionary, Cran-1 Collection.

Figure 34.

The use of titles only for input to a retrieval system may be expected to provide a widely differing performance efficiency depending on two circumstances:

1. The degree to which titles contain specific and exhaustive descriptions of the document content, as opposed to "novelty" titling designed only to draw attention to the document;
2. The type of documentary need demanded by the set of requests in use, ranging from a need which is satisfied by a total document only (thus enabling a good title to provide a satisfactory link), to a need which is satisfied by a small portion often unrelated to the major subject of the document (where titles will be quite unsatisfactory).

The first factor may be expected to differ with the subject field and the amount of control exercised in the technical writing (technical reports may differ from journal articles, for example). Figure 1 shows that the Cran-1 Aerodynamics titles are the longest, with IRE-3 Computer Science titles second longest, and ADI Documentation the shortest, on average. For example, a Cranfield title picked at random reads "Static Longitudinal Stability Characteristics of a blunted glider reentry configuration having 79.5° sweepback and 45° dihedral at a mach number of 6.2 and angles of attack up to 20°". Many of the Cranfield documents are technical research reports, whereas documents in the ADI collection are all conference 'short' papers, and documents in the IRE collection are predominantly journal articles. The Cranfield titles are undoubtedly the best for retrieval, thus explaining the smallest difference that exists between title and abstract performance on that collection.

The IRE titles are all quite short, but only a very few contain novelty titles, such as "A new concept in computing". The IRE requests are quite long, and do match at least one word in most of the titles of the relevant

documents, but the abstracts give a much better performance. The ADI titles are short also, but analysis has shown that in cases where the title does not give a good performance, a search of the full text frequently results in a poor performance for these cases also. The cause of this is probably a combination of the relevance decisions used, and synonym recognition problems, since the subject terminology in documentation is thought to be less precise than in the other collections.

Where requestor needs are covered by whole documents treating the topic of the request, titles alone may frequently be adequate, and KWIC title indexes have proved to be useful tools for such needs. It was noted that in a subset of the requests used in the Cranfield Project tests, 31% of the relevant document titles in a set of 35 requests had a strong match with the search request [2, pages 36-39]. But requestor needs are not always for whole documents, since relevant portions of a document frequently answer a need as completely as a whole document. In these cases, titles are quite inadequate, and a more exhaustive selection from the text is essential for good retrieval.

Four examples are given comparing abstracts to full text using the ADI Collection, in Figures 35 to 39. The request statements are given, together with the words matching the documents, with matching aided by a thesaurus dictionary. Figure 35 shows a case where the increased matching on full text improves performance, whereas Figure 36 shows how an increased matching on full text can worsen performance. Figure 37 shows a case where matching and performance were unchanged by use of full text, since the weight of the important term "journals" was increased from 4 to 30 on text from abstracts. Figure 38 shows a case where the increased weights provided by full text fail to prevent a non-relevant document from receiving a rank

Request QB15 Retrieval Systems which provide for the automated transmission of information to the user from a distance.

Relevant Document 18, Using the Abstract

Matching Concepts and weight in document — Transmission (1)

Total Concepts in Document	16
Total Concepts in Request	6
Total Concepts that match	1
Cosine Numeric Correlation	0.0690
Rank position	58

Relevant Document 18, Using the Full Text

Matching Concepts and weight in document
— Automated (8)
— Transmission (15)
— Information (13)
— User (8 1/2)
— Distance (2)

Total Concepts in Document	139
Total Concepts in Request	6
Total Concepts that match	5
Cosine Numeric Correlation	0.3596
Rank position	4

Analysis of request QB15 and relevant document 18, comparing abstract and full text result, thesaurus, ADI Collection.

Figure 35.

Request QA12 Give methods for high speed publication, printing, and distribution of scientific journals.

Relevant Document 07, Using the Abstract

Matching Concepts and Weights in document
- speed (1)
- scientific (1)

Total Concepts in Document	19
Total Concepts in Request	6
Total Concepts that match	2
Cosine Numeric Correlation	0.1490
Rank position	15

Relevant Document 07, Using Full Text

Matching Concepts and Weights in Document
- speed (2)
- printing (1)
- distribution (4)
- scientific (1)
- journals (8)

Total Concepts in Document	126
Total Concepts in Request	6
Total Concepts that match	5
Cosine Numeric Correlation	0.1103
Rank position	41

Analysis of request QA12 and relevant document 07, comparing abstract and full text result, thesaurus, ADI Collection.

Figure 36.

Request QB12 Information dissemination by journals* and periodicals* [*these terms are mapped into one concept class by the thesaurus].

Relevant Document 04, Using Abstracts

Matching concepts, and weights in document		— Information (1)
Total Concepts in Document	13	— Dissemination (2)
Total Concepts in Request	3	— Journals (4)
Total Concepts that match	3	
Cosine Numeric Correlation	0.7145	
Rank position	1	

Relevant Document 04, Using Full Text

Matching concepts, and weights in document		— Information (2)
Total Concepts in Document	86	— Dissemination (2)
Total Concepts in Request	3	— Journals (30)
Total Concepts that match	3	
Cosine Numeric Correlation	0.6828	
Rank position	1	

Analysis of request QB12 and relevant document 04, comparing abstract and full text result, thesaurus, ADI Collection.

Figure 37.

Request QB10 Computerized information systems in fields related to chemistry.

Relevant Document 09, Using Abstracts

Matching Concepts, and weights in document	
Total Concepts in Document	14
Total Concepts in Request	5
Total Concepts that match	5
Cosine Numeric Correlation	0.5622
Rank position	1

— Computerized (3)
— Information (1)
— Fields (1)
— Related (1)
— Chemistry (1)

Relevant Document 09, Using Full Text

Matching Concepts, and weights in document	
Total Concepts in Document	111
Total Concepts in Request	5
Total Concepts that match	5
Cosine Numeric Correlation	0.3736
Rank position	2

— Computerized (19)
— Information (8)
— Fields (4 1/2)
— Related (1)
— Chemistry (10 1/2)

Non-relevant Document 76, Using Full Text

Matching Concepts, and weights in document	
Total Concepts in Document	110
Total Concepts in Request	5
Total Concepts that match	4
Cosine Numeric Correlation	0.3910
Rank Position	1

— Computerized (3)
— Information (36)
— Fields (10)
— Related (1)

Analysis of Request QB10, relevant document 09 and non-relevant document 76, comparing abstract and full text results, thesaurus, ADI Collection.

Figure 38.

COLLECTION	DICTIONARY	ABSTRACT(A) VERSUS TITLE(T) I R-2 R-5 R-8	II NR NP	III (NR)(NP)	ABSTRACT(A) VERSUS TEXT(X) I R-2 R-5 R-8	II NR NP	III (NR)(NP)	ABSTRACT(A) VERSUS INDEXING I R-2 R-5 R-8	II NR NP	III (I) (NR)(NP)
SMART IRE-3 34 Requests	Stem	[A] [A] [A]	[A] [A]	(A) (A)						
	Thesaurus -3	[A] [A] [A]	[A] [A]	(A) (A)						
SMART ADI 35 Requests	Stem	A [A] [A]	[A] [A]	(A) (A)	X [X] X	X X	X A			
	Thesaurus -1	T T A	[A] [A]	(A) (A)	A X X	X X	X (X)			
SMART CRAN-1 42 Requests	Stem	A T [A]	[A] [A]	(A) (A)				I [I] I	I I	(I) I
	Thesaurus 2†,3*	A* A* A*	A* [A*]	(A*) A*				[I†] [I†] [I†]	I† I†	I† (I†)
	Stem(Overlap Logical)	T T A	A T	A T						
CRANFIELD CRAN-1 42 Requests	Word Form (Co-ord. Cutoff)	[T] [T] T						[I] [I] I		
	Word Form (Doc. Cutoff)	A T A						[I] [I] I		

I Precision at Recall 0.2(-2), 0.5(R-5), and 0.8(R-8). [ ] = over .05 better. [No Box] = up to .05 better.

II Normalized Recall(NR) and Normalized Precision(NP) [ ] = over .05 better. [No Box] = up to .05 better.

III Individual Requests based on merit assigned by Normed. Recall (NR) and Normed. Precision (NP). ◯ = over 60% of requests better. [No Circle] = between 50% and 60% better

Summary of Performance Results of Document Length Comparisons

Figure 39

position better than the relevant one; this is due to a highly weighted common term ("information") which gives a high correlation to the non-relevant document. For this request, some extra weight placed on the important term "chemistry" would preserve the perfect rank position of relevant document 09 on full text.

The average performance results show that although text is superior to abstracts, the improvement is small. Since the ADI abstracts are shorter than those used in Cran-1 or IRE, and probably do not include so much useful information, longer abstracts might perform better than full text. Any validation of full text searching would need to be carried out with text lengths more comparable to the average journal article or report than the short papers used, but even the use of these somewhat unsatisfactory documents suggests that text searching is feasible and worth further study.

The small superiority of indexing over abstracts can be explained by two possible reasons:

1. The indexers chose some terms from the full texts of the documents that the abstractors failed to include, and some of these terms represented subject notions that were asked for in the requests.
2. By choosing nearly half the number of terms contained in the abstracts, the indexers avoided notions that are not asked for in the requests, notions which only serve to increase the matches between requests and non-relevant documents.

The second case was previously illustrated by Figure 33. Several examples of the first reason have been found; for example, in three documents the terms "bust", "Stalling" and "Quasi-conical (flow)" appear in the indexing but not in the abstracts, and these ideas are all demanded in requests. There are cases also where the specific subject such as "wing" and "channels"

is included in the indexing, but the abstract mentions only the more generic ideas of "surface" and "walls" respectively, and the dictionaries in use do not make the necessary connections. Another example of this type is a request involving "transonic", where the indexer included that word from the text of the document, but the abstractor just used the more specific notion "Mach 0.6 – 1.6". This is basically a difficult synonym recognition problem.

An analysis has also been made to attempt to find important subject ideas that the indexers omitted but the abstractors included, but few examples were found. One case is the concept of "Computing (time)", mentioned in the abstract, but not in the indexing. It must be concluded that the main reason for the superiority of the indexing is that the indexers did a better job of making a précis of the full text than did the abstractors, at least in relation to the search requests tested. The indexers both selected subject notions that the abstractors missed, and also made shorter précis, which prevented retrieval of non-relevant documents and thus increased precision.

6. Conclusions

A simplified summary of the precision recall curves, the normalized measures and numbers of individual requests favoring a given option is presented in Figure 39. Conclusions may be enumerated as follows:

a) The use of very short documents, namely, titles only, is unsatisfactory in all collections for users requiring high recall. Recall ceilings are 0.71 (Documentation), 0.78 (Aerodynamics) and 0.84 (Computer Science).

b) The use of titles only for users requiring high precision performance is inferior to abstracts in all tests on the IRE-3 Collection

(Computer Science), with the stem dictionary on the ADI Collection (Documentation), and with both stem and thesaurus dictionaries employing weighting and the cosine correlation on the Cran-1 Collection (Aerodynamics). Titles perform better than abstracts on ADI using the thesaurus, which is probably due to poor abstracts rather than good titles. Titles also perform well on Cran-1 when simple matching (overlap correlation) and no weights (Logical Vectors) are used; this is due to the very good length and quality of titling in aerodynamics.

c) The use of abstracts in the ADI collection was only slightly inferior to full text at high precision using the stem dictionary, and at high recall using the stem and thesaurus dictionaries. It is suggested that the increase in recall/precision performance and increase in recall ceiling from 0.92 to 1.00 is unlikely to be worth the increased input and storage costs, and extended search time, and the use of slightly longer abstracts would show the text to have no advantages at all. Further work on full text processing of a more typical set of technical documents in another subject area is required.

d) The use of abstracts in the Cran-1 Collection gave a somewhat inferior performance to the shorter précis made by the manual indexers on the Cranfield Project. Further work is required to determine whether the apparently good quality abstracts suffer either from excessive length or failure to include some vital subject notions that the indexers included. The abstract performance is, however, sufficiently good to question the need for indexing for high performance, particularly since the indexing was more exhaustive than is practiced in many operational situations.

References

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

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