II. Evaluation Parameters

E. M. Keen

1. Introduction

Evaluation of the SMART system centers on techniques for the measurement of retrieval performance. Some reasons for concentrating on this evaluation criterion have been given in [1]. This section discusses many aspects of retrieval performance measurement in general, describes several of the measures used by SMART, and gives a detailed account of the way in which results of individual requests are processed in order to present averaged results. Several measures professed by other researchers in the area are examined and evaluated, and some considerations relating to future testing are made.

2. Purposes, Viewpoints and Properties of Performance Measures

Since performance measures are used for different purposes according to test objectives, a division into three types is suggested. Firstly, there is the need for measures with which to make merit comparisons within a single test situation, that is, to make 'internal' comparisons only. In tests of this type the document collections, search requests, and relevance decisions are held constant while some system variable is altered, and this procedure has been used for almost all of the SMART experiments. Such situations are best characterized, in terms of performance measurement, by saying that comparisons are made in situations of constant generality, and a "generality number" may be computed in such cases [2]:

$$G = \frac{\text{Total Relevant in Collection} \times 1000}{\text{Total Documents in Collection}}$$

Although generality tends to vary between requests, an average value for a set of requests serves to characterize a particular series of experiments.

A second purpose of performance measurement is that of making 'external' comparisons between results obtained in different situations, in which generality is expected to differ. Such comparisons may be made even within an experimental test environment, if different request sets or collection sizes are introduced and compared.

A third purpose that may be distinguished is a specific need to interpret experimental results in terms of expected real-life merit, rather than merely comparing different techniques in a laboratory. Experimental tests of the kind conducted by SMART are simulation-tests, and any conclusions drawn from the results may need to be presented in a way that would be typical of the performance if the system were being used operationally.

The choice of performance measures is also affected by viewpoint, either the viewpoint of the user, or of a researcher seeking fundamental insight into retrieval capability. User satisfaction is restricted to properties "a", "b", and "c" in Figure 1, since a user is interested in examining as few non-relevant items as possible, and as many relevant items as he wishes to see, but he is not concerned about "d", or about the total collection size. From a system efficiency viewpoint, which is of concern in some types of research, the value of "d", and the collection size are needed. For example, test comparisons between situations of differing generality require measures that include "d" if a strict comparison of efficiency is the object. Still more sophisticated techniques may be needed, since correct system efficiency comparisons require adjustment for differing concentrations of documents by subject in different collections, so that the actual collection size can be replaced by the real number of documents within the subject

	Relevant	Non Relevant	
Examined	a	b	a + b
Not Examined	c	d	c + d
	a + c	b + d	a + b + c + d

$$\text{Recall} = \frac{a}{a + c}$$

$$\text{Cut-off} = \frac{a + b}{a + b + c + d}$$

$$\text{Precision} = \frac{a}{a + b}$$

$$\text{Generality Number} = \frac{(a + c)1000}{a + b + c + d}$$

$$\text{Fallout} = \frac{b}{b + d}$$

The Retrieval Table that Results from Searching a System, with Five Ratios Derived from the Table

Figure 1.

areas covered by a set of requests. No suitable method of achieving this type of comparison has yet been developed, but it is crucial to further research in this area because clearly some collections are more hostile to a good retrieval performance simply because these contain a large number of potentially retrievable non-relevant items.

Four desirable properties of retrieval performance measures are suggested by John Swets [3], namely that the measure should be:

— able to measure retrieval effectiveness alone, separately from other criteria such as cost;

— independent of any particular cut-off;

— a single number;

— on a number scale to give absolute and relative values.

Swets, however, does not recognize the possibility that different purposes and measurement viewpoints may be important, and the resulting measure proposed takes no account of the user viewpoint in a directly meaningful way. From matters discussed already, several other properties appear desirable:

— ability to reflect success of system in meeting needs of different types, such as high precision, or high recall;

— ability to interpret measures directly in terms of a user's experience: for example, 0.2 precision at 0.5 recall means that the user has examined half the relevant documents available, while at the same time four non-relevant document items were looked at for every one relevant;

— ability to compare systems of differing generality.

Other properties can be suggested, but the purposes and viewpoints here suggested should override such properties as the "single number" or "absolute and relative scales", which are desirable perhaps but not essential. The purposes, viewpoints and properties discussed are summarized in Figure 2.

Ⓐ Purposes

1. 'Internal' test comparisons, "G" constant.
2. 'External' test comparisons, "G" varies.
3. Interpretation of merit in simulated real-life terms.

Ⓑ Viewpoints

1. System efficiency
2. User satisfaction

Ⓒ Properties

1. Retrieval effectiveness alone
2. Independent of cut-off
3. Single number
4. Absolute and relative values
5. Differing user needs
6. Interpretation in users' terms
7. Comparisons involving "G" changes

Factors Affecting the Choice of Performance Measures

Figure 2.

3. Measures for Ranking Systems

The provision of a ranked output, in which documents are ordered according to the magnitude of their correlation coefficient with the search request, makes it possible to use evaluation measures of many types, since a direct evaluation of the rank positions occupied by the relevant documents may be made, or a series of cut-offs may be applied according to many different criteria. The requirement of a unique rank position for every document in the collection in SMART does require some procedures for dealing with tied ranks. Specifically, there always exist some documents that exhibit a zero correlation with the search request; these documents are given a rank position according to a random procedure in order that complete performance merit can be measured. Although relevant documents rarely take up such rank positions in the more usual processing runs, the use of titles alone does result in larger numbers of such cases. Also, tied ranks may occur with possitively correlated documents particularly when the overlap correlation is in use; this, however, occurs quite rarely with the better cosine correlation runs.

The performance measures used by SMART are now briefly described, and some additional suggested measures noted. The primary purpose of measurement in SMART has been that of internal comparisons (Purpose 1, Figure 2), and the viewpoint that of user satisfaction(Viewpoint 2, Figure 2).

A) Single Number Measures

Sets of measures known as Rank Recall and Log Precision, and Normalized Recall and Normalized Precision are in use, and have been described [1,4,5,6]. These measures are cut-off independent in that the rank positions of all the relevant documents to a request are compared with the ideal positions resulting from a perfect system. Results presented in other sections of this report employ the two normalized measures, so the formulas are repeated for convenience:

$$\text{Normalized Recall} = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N - n)}$$

$$\text{Normalized Precision} = 1 - \frac{\sum_{i=1}^{n} \log r_i - \sum_{i=1}^{n} \log i}{\log \frac{N!}{(N-n)!\ n!}}$$

where n = number of relevant documents

N = number of documents in collection

r_i = rank of i^{th} relevant document

i = ideal rank positions for the i^{th} relevant item.

The result obtained from one individual search request is given in Figure 3, and both the normalized measures are computed. Normalized recall gives equal 'weight' to documents with high rank positions as to documents with low rank positions, but normalized precision gives stronger weight to the initial section of the retrieval list, that is, to those with high rank positions.

An attempt to derive a single number measure of a quite different type is reported by John Swets [3]. It is different from the measures used by SMART since it does not directly use the ranked output list, but uses in the first place performance curves similar to those discussed in the next sub-section; examination of this measure is thus deferred. The "normalized 'sliding ratio' measure" proposed by Giuliano and Jones [8] appears to be designed for use at one selected cut-off point, and so again differs from the SMART measures.

B) Varying Cut-off Performance Curves

The most common measures of retrieval performance are the precision and recall ratios derived from the retrieval table, and given in Figure 1.

RELEVANT DOCUMENTS

Rank	Number	Correlation
1	588	.5764
2	589	.5477
4	590	.3833
6	592	.3523
13	772	.2092

$$\text{Normalized Recall} = 1 - \frac{\sum 1,2,4,6,13 \; - \; \sum 1,2,3,4,5}{5(200 - 5)}$$

$$= 0.9887$$

$$\text{Normalized Precision} = 1 - \frac{\sum \log 1,2,4,6,13 \; - \; \sum \log 1,2,3,4,5}{\log \frac{200!}{(200 - 5)!\; 5!}}$$

$$= 0.9238$$

Result of Cran-1 Individual Request Q268, Thesaurus-2,
Showing Evaluation Using the Normalized Measures.

Figure 3.

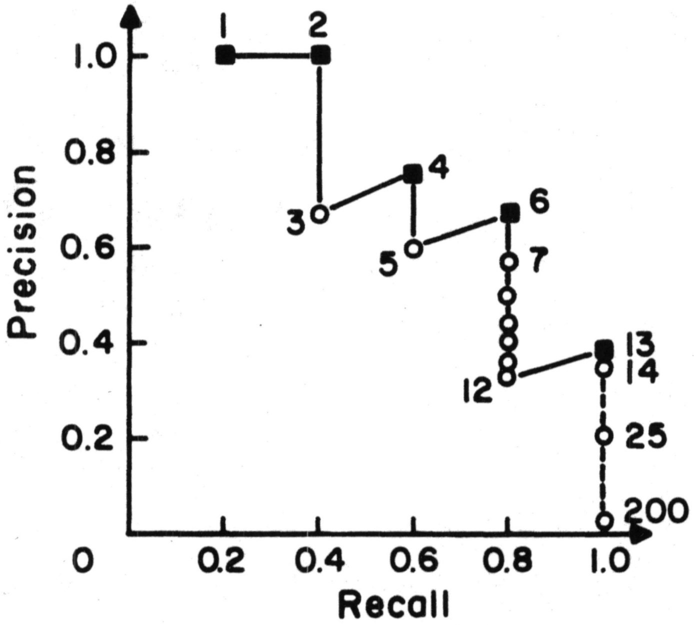
These measures are desirable even with a ranking system, since they alone seem capable of representing a user's viewpoint (Viewpoint 2, Properties 5 and 6, Figure 2). It is a simple matter to construct performance curves of this type from a ranked output, since a series of cut-off points may be chosen, precision and recall calculated, and the points joined for form a curve.

A precision versus recall curve for an individual request is presented in Figure 4, using the familiar graph, and showing the shape of the curve when a cut-off is established after each document. Results for a single request always exhibit the step pattern, but interpolation and extrapolation technique to be described in Part 4 produce a smoother curve. The practice, as with the normalized measures, is to present results averaged over a whole set of search requests, so Figure 5 shows as an example some averages for two retrieval runs in the form of a tabular computer print-out, and a graph of the precision versus recall curves.

A quite similar "performance characteristic" curve is proposed for use with ranking systems by Giuliano and Jones [8], it seems to offer no advantage over the precision versus recall curve. It is advocated for another reason to be discussed in Part 5. The normalized "sliding ratio" measure also proposed by Giuliano and Jones uses either the recall or precision ratios at each cut-off point. The equation is given and an example is calculated in Figure 6, showing that up to a cut-off equal to the number of relevant items, this measure is the precision ratio, and at higher cut-offs, the measure equals recall. While it is true that a perfect result would produce a perfect measure of performance, it would do so at every cut-off point, which would not seem to be a desirable result. In the perfect case, for example, a user who wanted high recall, not knowing how many relevant the system contained, might suggest a cut-off too 'early' in the list, and miss some relevant items, yet this measure would show a perfect result at

Recall Precision after retrieval of n documents

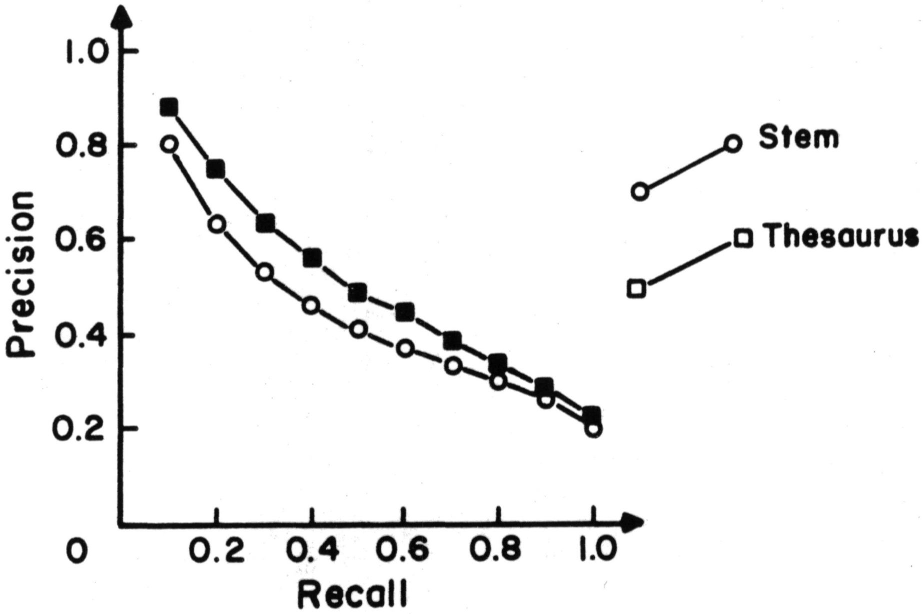
n	Number (x = Relevant)		Recall	Precision
1	588	x	0.2	1.0
2	589	x	0.4	1.0
3	576		0.4	0.67
4	590	x	0.6	0.75
5	986		0.6	0.60
6	592	x	0.8	0.67
7	984		0.8	0.57
8	988		0.8	0.50
9	578		0.8	0.44
10	985		0.8	0.40
11	103		0.8	0.36
12	591		0.8	0.33
13	772	x	1.0	0.38
14	990		1.0	0.36



Result of Cran-1 Individual Request Q268, Thesaurus-2 Showing Evaluation Using a Graph of Precision Versus Recall

Fig. 4.

ADI ABSTRACTS COLLECTIONS, AVERAGES OVER 35 REQUESTS			
STEM		THESAURUS	
REC.	PREC.	REC.	PREC.
0.1	0.7963	0.1	0.8788
0.2	0.6350	0.2	0.7567
0.3	0.5283	0.3	0.6464
0.4	0.4603	0.4	0.5577
0.5	0.4051	0.5	0.4912
0.6	0.3699	0.6	0.4470
0.7	0.3383	0.7	0.3893
0.8	0.2996	0.8	0.3287
0.9	0.2568	0.9	0.2726
1.0	0.2018	1.0	0.2093
RNK REC =	0.2415	RNK REC =	0.2534
LOG PRE =	0.3587	LOG PRE =	0.3837
NOR REC =	0.7652	NOR REC =	0.8045
NOR PRE =	0.5339	NOR PRE =	0.6075



Results of Two Retrieval Runs, showing the Averages Produced.

Fig. 5

Normalized "Sliding Ratio" Statistic:

$$\mu(n) = \frac{f(n)}{f^*(n)}$$

where n = rank position
f = number of relevant examined in ideal result
f^* = number of relevant actually examined.

Example using Cran-1 Request Q268, with Ranks of Five Relevant

1,2,4,6,13

Rank(n)	f	f^*	μ
1	1	1	1.00
2	2	2	1.00
3	3	2	0.67
4	4	3	0.75
5	5	3	0.60
6	5	4	0.80
7	5	4	0.80
⋮	⋮	⋮	⋮
13	5	5	1.00
14	5	5	1.00
⋮	⋮	⋮	⋮

Result of Cran-1 Individual Request Q268, Thesaurus-2, Showing Evaluation Using the Normalized "Sliding Ratio" Statistic Proposed by Giuliano and Jones [8].

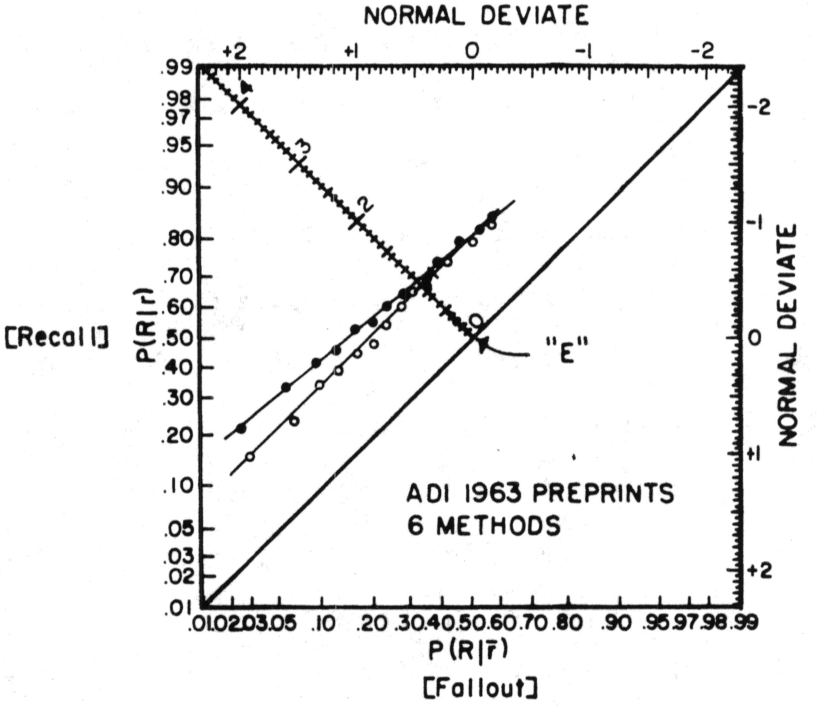
Fig. 6.

that cut-off point. Other similar examples can be constructed, and the conclusion is that the normalized "sliding ratio" measure does not include many valuable features.

The "operating characteristic" curves used by John Swets [3] use a graph of the recall ratio and the fallout ratio, described by him in terms of probabilities. The fallout ratio has been used in previous experiments [2], and is discussed in Part 6. Swets uses this measure because the operating characteristic curves may be examined in terms of statistical decision theory, and, hopefully, a single number measure may be derived to represent the whole curve, if the curves follow some suitable theoretical model. Some results from SMART and other experimental tests are used by Swets, but the resulting fit with the model curves is only partially successful, in that an "s" value as well as an "E" value are strictly required to characterize an operating characteristic curve, as shown in Figure 7. It should be noted that although this kind of measure is suitable for reflecting the system efficiency viewpoint, and meets nearly perfectly properties 1, 2, 3, 4, and 7, it does not and cannot display user satisfaction in terms of precision, and therefore does not meet properties 5 and 6 (Figure 2).

C) Comparison of Single Number and Curve Measures

The relationship between the single number normalized measures on the one hand, and the precision recall curve on the other has not yet been theoretically established. Both types of measures are obtained for every retrieval run, and in the vast majority of cases the two types of measure give the same merit when two runs are being compared for effectiveness. For example, the two average sets of results given in Figure 5 show that both normalized recall and normalized precision favor the Abstracts, Thesaurus option, and the same result is given by the precision recall curve, since



○—● Stem , "E" = 0.90 , Slope = 0.99

●—● Thesaurus , "E" = 1.00 , Slope = 0.83

ADI Abstracts, "Pseudo-Cranfield" Cut-off Micro

Evaluation of 35 Requests.

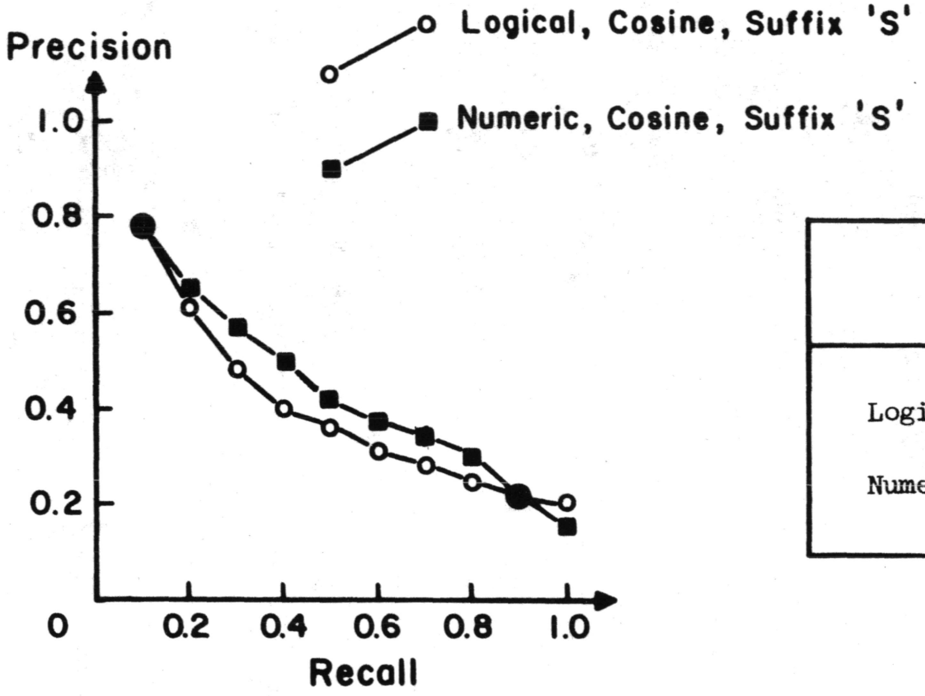
Two "Operating Characteristic" Curves of Fallout Versus Recall

and "E" values obtained, as suggested by Swets [3].

Fig. 7.

the curve for Abstracts Thesaurus is closer to the 1.0 precision at 1.0 recall corner over the whole of its range than the curve for Abstracts Stem.

A minority of results do not show such complete agreement, and a comparison presented in Figure 8 shows that only above 0.9 recall does the curve merit agree with merit assigned by the normalized measures. Two individual requests from the request set used are given in Figure 9, showing that although in both requests the normalized measures strongly favor the "Cosine Logical" option, some portions of the precision recall curve favor "Cosine Numeric". In request QA12 the ranks of the last two relevant documents favor cosine numeric, but the normalized measures are more directly influenced by the larger rank changes at the top rank positions that favor cosing logical. In request QA4 the same effects cause the high precision end of the curve to favor cosine numeric. Clearly single number measures cannot reflect crossing performance curves, unless the measures are specifically designed to reflect merit that exists at a particular point on the curve. But this possibility is not met by the normalized measures, and it is not always correct to say that normalized recall reflects merit at the high recall end of the curve, and normalized precision does so at the high precision end. For example, Figure 10 shows a result in which the average curve for "First Iteration" is at all points better than "Initial Search", yet normalized recall indicates that the latter appears to be better. This occurs because the "First Iteration" result improves ranks of quite a few documents that were already quite highly ranked in "Initial Search" (thus the normalized precision is best for "First Iteration"), but at the same time, some other relevant documents that were poorly ranked on "Initial Search"are worsened by quite large amounts in "First Iteration", thus causing normalized reeall to drop, without affecting the curve appreciably.



ADI, Text, 35 Requests

	Normed. Recall	Normed. Precision
Logical, Cosine	.7768	.5462
Numeric, Cosine	.7520	.5308

Comparison of Merit Assigned by Decision Recall Curve

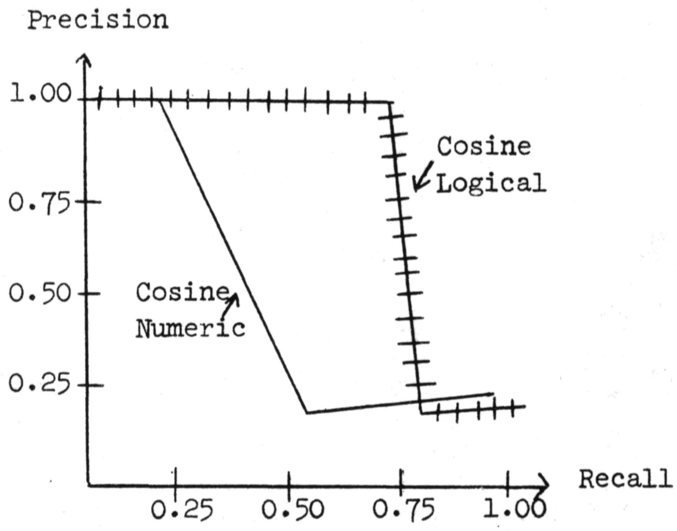
and Normalized Measures in run case of Poor Agreement

Fig. 8

Request QA12 5 Relevant Documents

Cosine Numeric — Ranks of Relevant 1,3,14,17,18

Cosine Logical — Ranks of Relevant 1,2,3,18,23

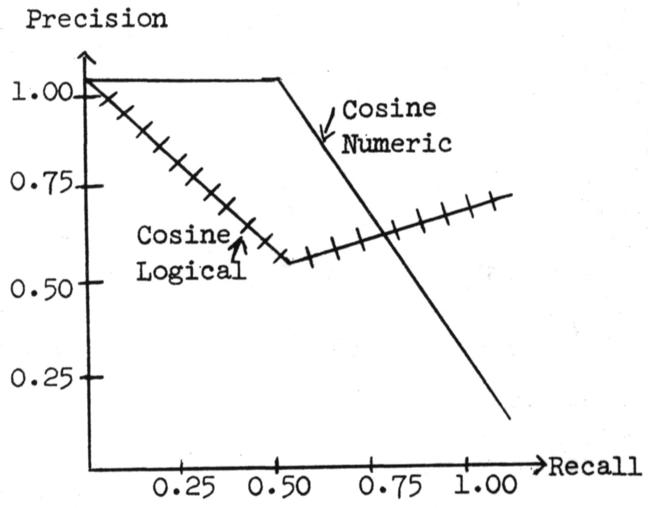


	Cosine Numeric	Cosine Logical
Normed Recall	.9013	.9169
Normed Precision	.7270	.8230

Request QA4 2 Relevant Documents

Cosine Numeric — Ranks of Relevant 1,15

Cosine Logical — Ranks of Relevant 2,3



	Cosine Numeric	Cosine Logical
Normed Recall	.9188	.9875
Normed Precision	.7515	.8645

NOTE: Precision Recall curves are extrapolated to 1.0 Precision 0.0 Recall, as discussed in Part 4C.

Results of Two Individual Requests from the ADI, Text, Suffix 's', Cosine, Numeric Versus Logical Runs, showing merit assigned by normalized measures and precision recall curves

Fig. 9

These examples given are practically the only such observed in over one hundred performance comparisons, and thus are definitely the exception rather than the rule. The reasons for the discrepancies lie in the way in which the different measures apply diferent weight to different distributions of the relevant documents; some research proposed by Michael Lesk is designed to investigate this problem.

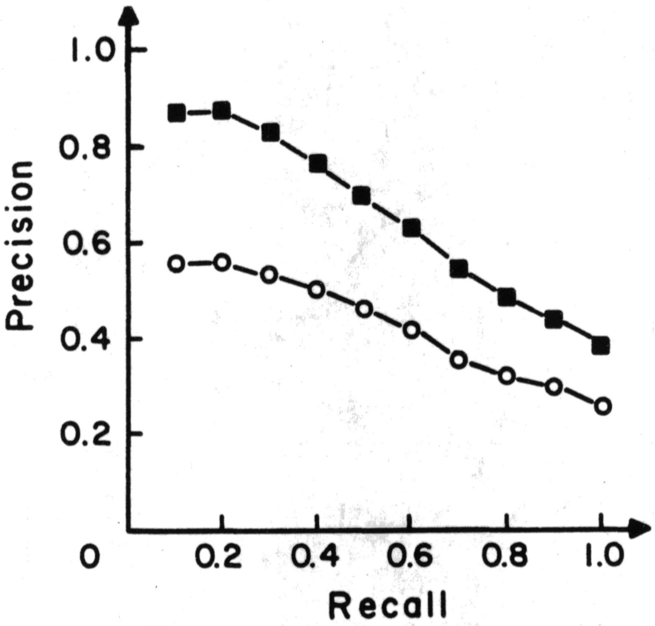
4. The Construction of Average Precision Versus Recall Curves

In the context of the SMART experiments, the construction of a precision versus recall curve for a set of search requests requires techniques for averaging over individual requests, chosing cut-off points to construct curves, and coping with problems that arise because individual requests have differing numbers of relevant documents. Different methods of meeting these three problems are suggested, and these methods are divided into those that are suitable only for test comparisons (Purposes 1 and 2, Figure 2), and those that satisfy the need to accurately simulate the result experienced by real users (Purpose 3, Figure 2). An additional problem that arises only for the fast cluster searches is also discussed.

A) Averaging Techniques

The two main alternative averaging techniques have been described as "micro evaluation" and "macro evaluation" [1,5,6]. The micro method requires the comulation over all requests of the number of documents both retrieved and relevant (for a given cut-off) so that one final precision-recall pair can be calculated, whereas the macro method requires the computation of precision-recall pairs for each request with the final precision-recall pair obtained by averaging, using the arithmetic mean. The macro method is generally preferred because it provides both adequate comparisons

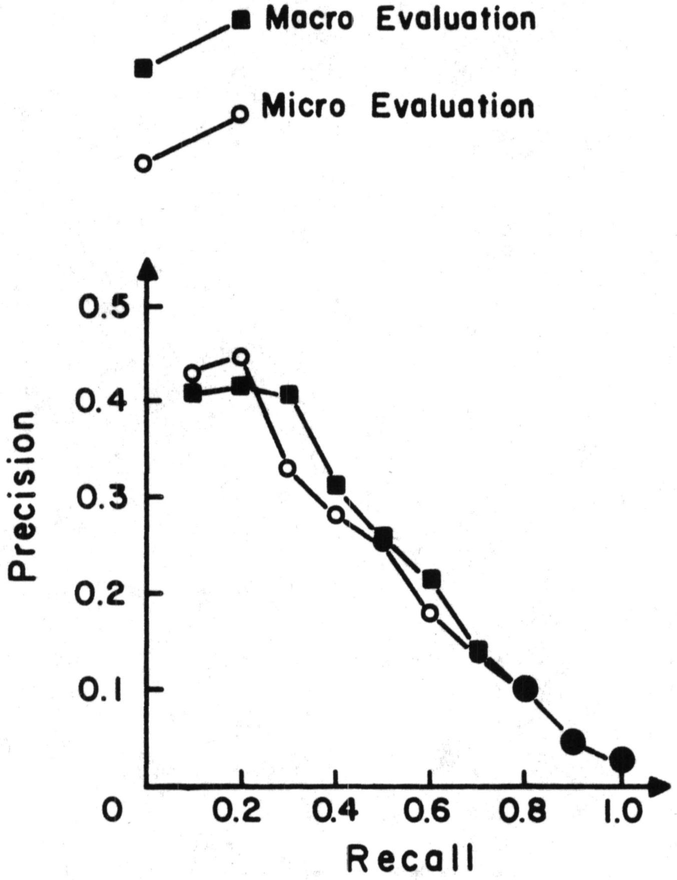
	Normed. Recall	Normed. Precision
○ Initial Search	.8772	.6815
□ First Iteration	.8680	.7622



Cran-1, Thesaurus-3, Averages over 42 Requests

Results of Two Searches in a Relevance Feedback Evaluation Run, Comparing Merit Assigned by the Normalized Measures with the Precision Recall Curve

Fig. 10.



Cran-1, Abstracts, Thesaurus-3, "Pseudo-Cranfield" Averages
Over 42 Requests

Retrieval Comparison of the Macro and Micro Averaging Methods

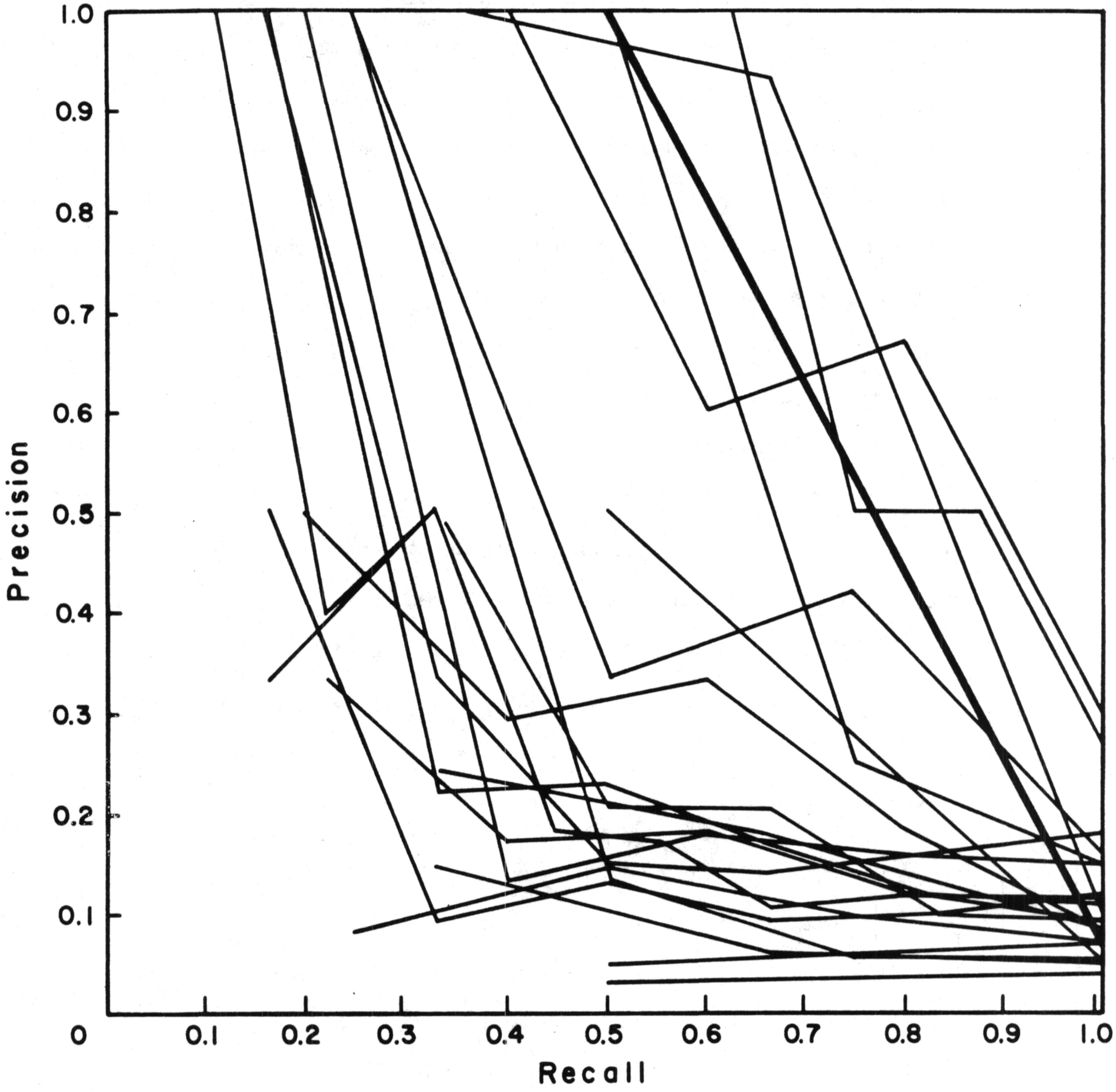
Fig. 11.

for test purposes and meets the need of indicating a user-oriented view of the result; the micro method on the other hand tends to give undue weight to requests that have many relevant documents. As Salton and Rocchio have shown [1,5], the macro method results in somewhat better precision recall curves, but the difference between the two methods with current collections and requests is near to or less than 5%, as seen in the comparison of Figure 11. An occasional use of the micro method has usually given the same performance merit when two options are compared, so that this issue does not affect comparative test results at all.

Further work on the averaging problem may reveal that the arithmetic mean is not the only suitable method to use. Averaging is a problem simply because of the extreme variance in individual results, as can be seen from the plot of individual precision recall curves for 22 requests given in Figure 12. The macro evaluation curve for these 22 requests is given in Figure 13, together with a curve based on the median, rather than the mean. The scatter of results raises the question of statistical significance; this matter is discussed elsewhere [9].

B) Cut-off Techniques

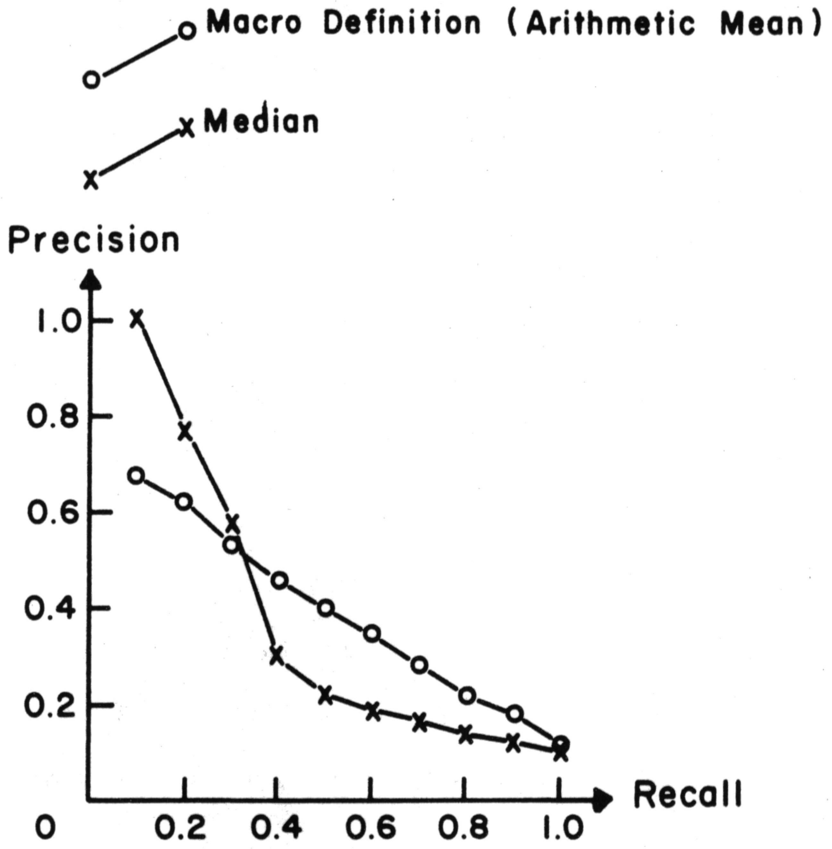
Cut-off techniques in conventional manual and mechanized retrieval systems usually depend on the search terms used, with specified term-matches establishing the cut-off points. The equivalent in SMART is the use of the correlation coefficient that is obtained between the request and each document, but the provision of ranked output permits other cut-off criteria to be used, specifically related to the exact number, or acceptability of the documents as they are examined. Cut-off techniques for experimental purposes must be based on methods applicable to all requests, regardless of variations in the number of relevant items. For this reason the ranked output list only is used.



ADI, Abstracts, Thesaurus

Individual Precision Recall Curves for 22 Requests, showing the Wide Scatter of Individual Results

Fig. 12



ADI, Abstracts, Thesaurus, Averages of 22 Requests

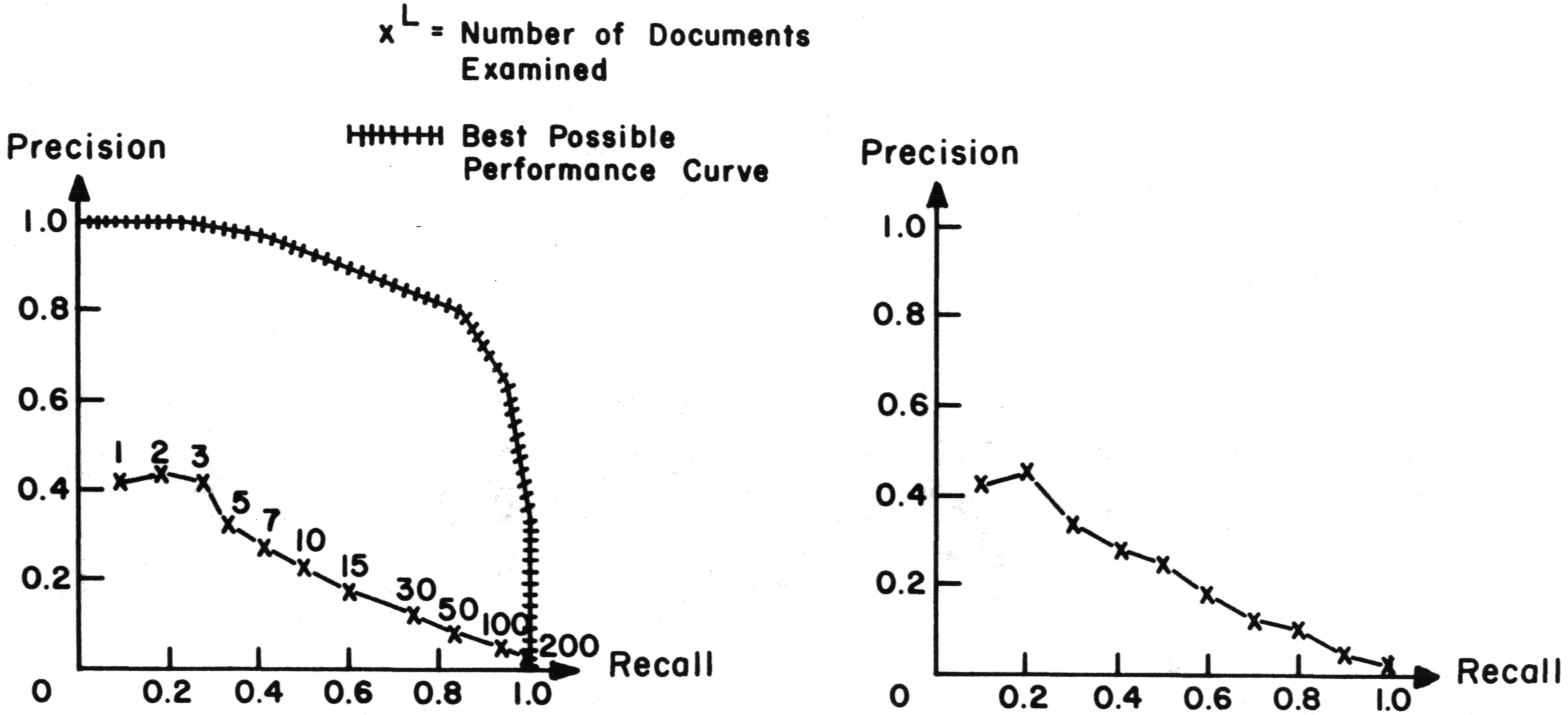
Comparison of Average Results Using Mean and Median of the Individual Requests.

Fig. 13

and no account is taken for cut-off purposes of the correlation, although one study using correlation magnitudes has been made [10].

Using the precision recall pairs that can be computed as each document in the output list is examined (Figure 4), three cut-off methods seem feasible. The first method is to obtain average curves from all requests just as drawn in Figure 4, by computing mean precision recall pairs for each document cut-off level. If done by hand, the cut-off points may be recorded on the curve as in Figure 14 a), or a computer-produced average may be used which produces precision at ten recall levels for plotting convenience, Figure 14 b). This technique is referred to as the "Pseudo-Cranfield"method, and although it is available for many runs it is not generally used for SMART evaluations. One advantage of this method is that is seems to be fully user-oriented, since the plot of Figure 14 a) shows how many documents a typical user must examine to get "x" recall. Another advantage is that computation does not depend on the interpolation and extrapolation techniques that are required for the other methods to be described. A disadvantage stems from the fact that the requests vary according to the number of relevant items so that if one of the requests has only a single relevant document, any cut-off made at 2 or more documents will not give 1.0 precision even if all requests have a quite perfect performance. One simple solution to this is to give the theoretical best possible curve for a given set of requests, as is done in Figure 14 a). It is a simple matter to use this cut-off method with macro evaluation, as the macro curve in Figure 11 was obtained this way.

The second and third cut-off techniques use, respectively, precision and recall ratios to determine the cut-off points at which averages will be computed. A set of precision or recall values are picked in advance, and requests are averaged essentially at the cut-off points at which the required



a) Document Cut-Off "Pseudo-Cranfield" computed by hand.

b) Document Cut-off - "Pseudo-Cranfield" as computed by evaluation routines at ten recall levels.

Cran-1, Abstracts, Thesaurus-3, Micro Evaluation of 42 Requests

An Illustration of the Document Cut-off or "Pseudo-Cranfield Method.

Fig. 14

precision or recall ratios are reached. The use of precision values, although theoretically possible, has not been tested, primarily because recall is more suitable for this, since precision does not monotonically decrease with rank (the upward sloping "steps" in Figure 4 indicate that more than one cut-off can achieve a given precision). Although recall does monotonically increase, there is still one problem that requires solution. The vertical segments of the 'step' curve for an individual request (Figure 4) show that at some recall points, more than one cut-off point may exist from which to choose, each giving a different precision ratio.

At least five possible solutions are available concerning the choice of a cut-off, namely, that having the highest precision, the lowest precision, the precision of the "middle" document, a precision ratio computed from the average precision over all cut-off points, or a precision ratio computed as the average of the top and bottom points only. Figure 15 indicates an example of each of these possible solutions.

There is a further question, relating to the precision values to be used at recall points where no vertical part of the step is encountered, such as at 0.5 recall in Figure 4. It is possible, for example, by using one of the five possible points at the vertical segments, to join up the chosen points on the vertical segment by a new interpolation line. Figure 16 a) shows that when the cut-off having the highest precision is chosen for use at the vertical segments, interpolation between these points of an individual request produces a smooth performance curve, that is quite suitable for averaging over sets of requests. This example of Figure 16 a) is the one most frequently used by SMART, and the description appeared first in [4]. This type of average curve normally uses ten recall levels, 0.1, 0.2, and so on, and is referred to as the "Quasi-Cranfield" method. Its advantage is that it can be quite

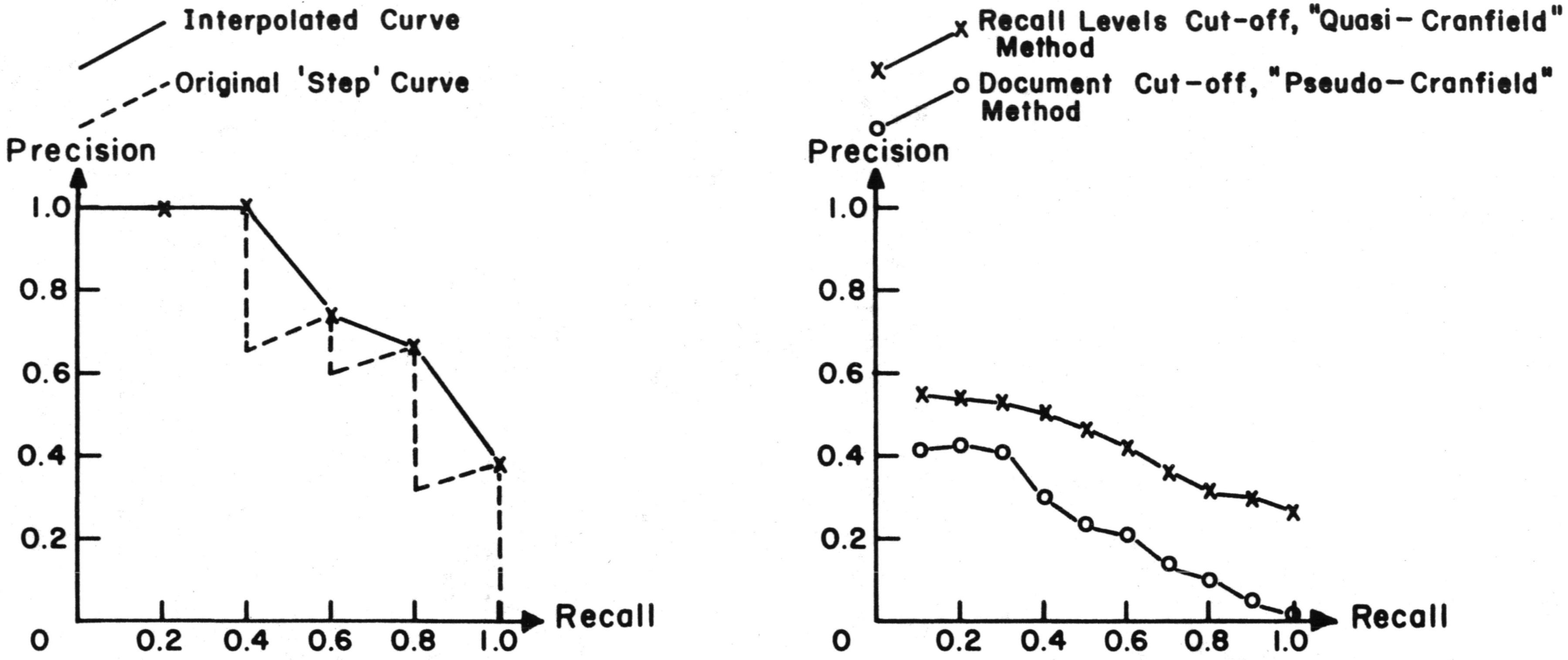
Ranks	Recall	Precision
5	.6	.6000
6	.8	.6667
7	.8	.5714
8	.8	.5000
9	.8	.4444
10	.8	.4000
11	.8	.3636
12	.8	.3333
13	1.0	.3846

Choices of Precision to Use at .8 Recall:

1. Document with highbest precision, rank 6, precision .6667
2. Document with lowest precision, rank 12, precision .3333
3. Document in 'middle' position, rank 9, precision .4444
4. Documents at all ranks, average precision .4685
5. Documents at best and worst ranks, average precision .5000.

Calculation Examples of Five Choices of Precision to be Used at Constant Recall, Using One Vertical 'Step' of the Precision Recall Curve from Figure 4.

Figure 15.



Ranks of Five Relevant are 1,2,4,6,13

a) Curve for Individual Request Q 268, Cran-1, Thesaurus-2, using Recall levels Cut-off - "Quasi Cranfield" Method.

Cran-1, Abstracts, Thesaurus-3. Macro Evaluation of 42 Requests.

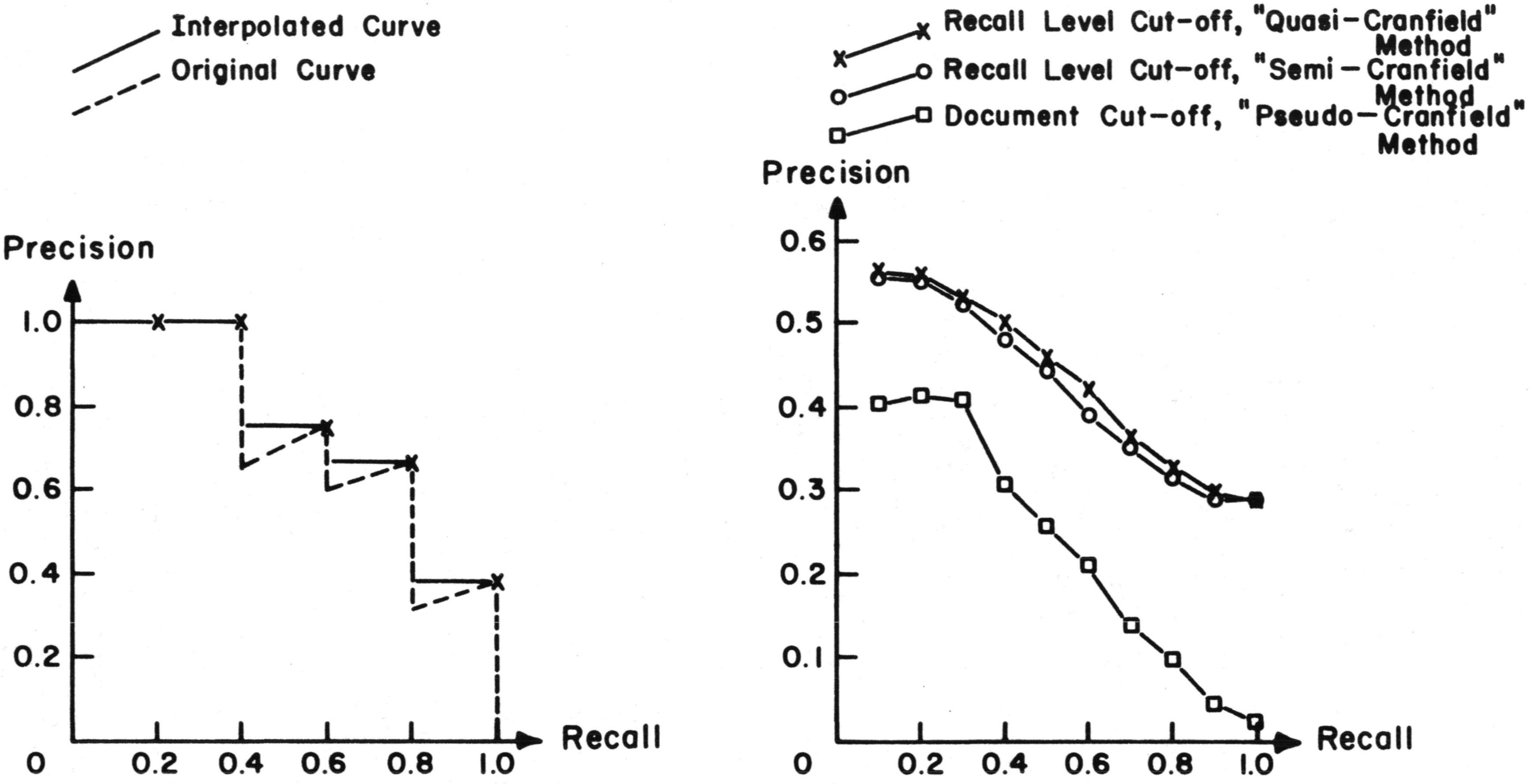
b) Average Curves to compare Cut-offs.

Illustrations of the Recall Levels Cut-off - "Quasi-Cranfield" Method.

Figure 16.

simply interpreted by noting that a cut-off is established immediately when a relevant document is encountered in each output list. It very effectively reflects merit at the high recall end of the plot, since the lowest precision ratio for any individual request is computed when a recall of 1.0 is reached, unlike the "Pseudo" plot which continues making cut-offs until the last document in the collection is reached. This technique is quite adequate for making comparisons with in SMART, but a possible disadvantage in some circumstances is that the curve is not typical of a real user environment since it produces too optimistic a result. Figure 16 b) compares a "Pseudo" and a "Quasi" curve for the same set of averaged results.

A modification to the technique is being tested, in which the cut-off having the highest precision at each vertical segment is still used, but the interpolation lines are altered to produce what is believed to represent the best possible curve that a user could achieve, assuming that almost optimum choices of cut-off are made. Figure 17 a) gives an example of this, and the reason for this type of extrapolation line which retains constant precision resides in the fact that user requirements would ask for the best possible precision above "x" recall. Whatever the value of "x" is, the best possible precision is always the next peak in the step curve, so a line of constant precision leading to that peak is thought to give the required result. A slight modification, yet to be made, is that sometimes, the next peak encountered above "x" recall is eclipsed by a higher peak at still greater recall (occurring, for example, when one relevant document is followed by another in the rank list). The line should thus be connected to the highest peak. This technique is known as the "Semi-Cranfield" method, and an average curve is presented in Figure 17 b), together with curves of the other two types. The comparison is slightly affected by a different tied rank procedure used for the "Semi-



Ranks of Five Relevant are 1,2,4,6,13

a) Curve for Individual Request Q 268 Cran-1, Thesaurus-2, using Recall levels cut-off - "Semi-Cranfield" Method

Cran-1, Abstracts, Thesaurus-3, Macro Evaluation of 42 Requests

b) Average Curve to Compare Cut-offs

Illustrations of the Recall Levels Cut-off - "Semi-Cranfield" Method

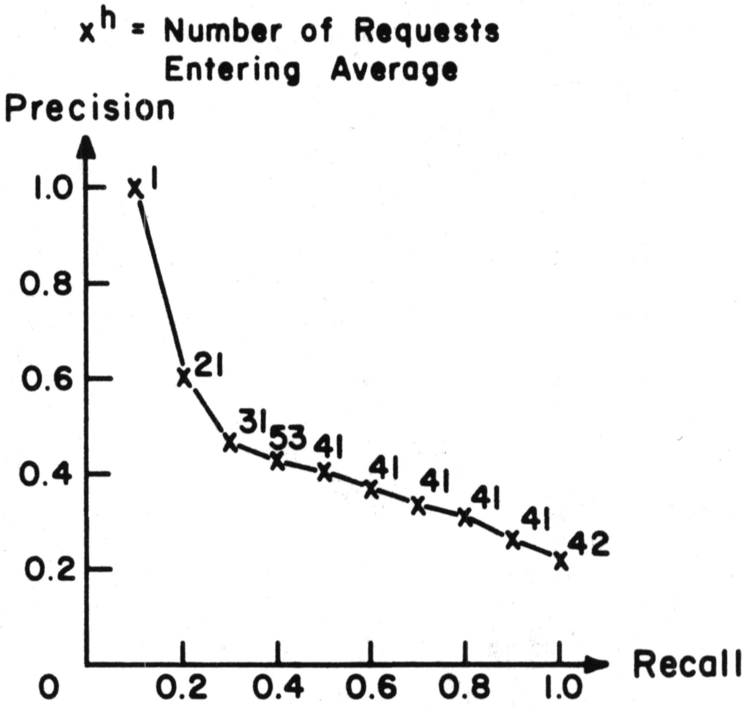
Fig. 17

Cranfield curve, but any differences due to this effect are very small indeed. In fact the "Quasi-Cranfield" and "Semi-Cranfield" methods result in a quite similar performance curve, but the latter does give the theoretical maximum performance that a user could achieve. Other choices of cut-off to be used at the vertical segments would give curves positioned lower on the graph than for these two methods, and would probably give performance curves that would be more typical of user experience. However, for experimental test comparisons, the procedures used are completely adequate.

C) Extrapolation Techniques for Request Generality Variations

Discussion of the recall level cut-off techniques suggests consideration of one further problem, caused by the variation in numbers of relevant documents for different requests. The problem is that requests having few relevant documents cannot exhibit low recall values, and therefore have shorter precision recall curves than those that have many relevant documents. The extreme example is furnished by a request with only one relevant document, where the performance on a graph is reflected by only a single point on the graph, somewhere at 1.0 recall. The question arises as to whether the performance of such a request should still be incorporated in the average results at recall levels lower than 1.0, and five possible methods are suggested.

The first method is to use individual precision-recall curves only at points where they can in fact be drawn by methods discussed in Part 4B; at low recall values, only those requests having many relevant documents will then enter into the averages. Figure 18 gives an example based on 42 requests, where the numbers of requests that would enter into the averages are given at each of ten recall levels. Although this method is quite simple to use and gives quite acceptable results for 'internal' test comparisons, any attempts to compare dissimilar request sets are complicated by different request



Cran-1, Abstracts, Stem, "Quasi-Cranfield" Cut-off, Macro Evaluation of 42 Requests.

Illustration of First Method of Averaging Where Individual Requests have Varying Amounts of Relevant Documents

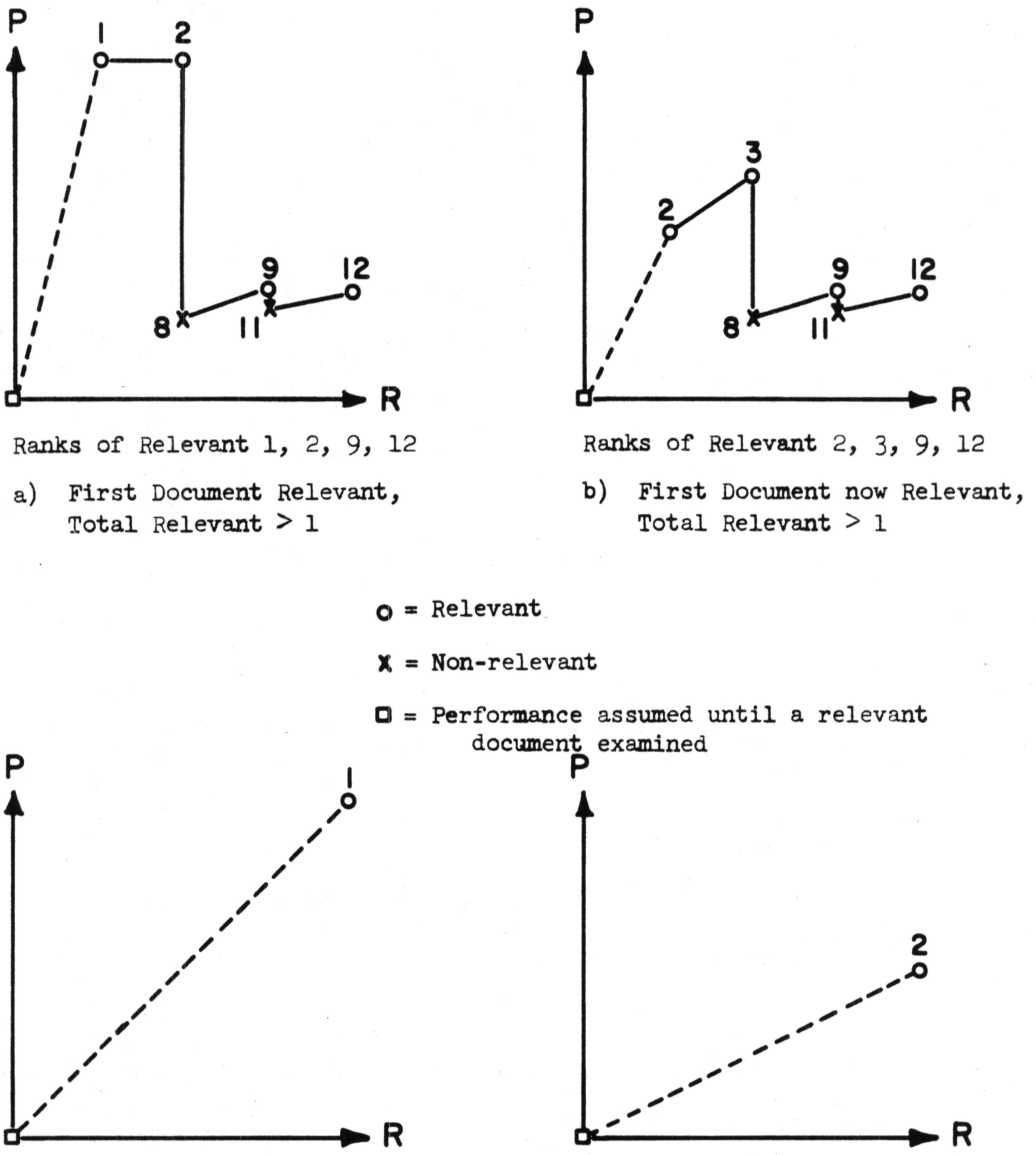
Fig. 18

generality distributions.

The other suggested methods all use some technique of extrapolation, so that all requests have full length precision recall curves that extend from 0.0 to 1.0 recall. The second method involves extrapolation of the beginning of all curves to 0.0 precision at 0.0 recall. Four examples using different numbers of relevant and different rank positions are given in Figure 19. This method is justified mathematically, since if no documents are retrieved (cases a) and c)) recall is 0, and precision is strictly zero, and if the first document retrieved is non-relevant, recall is zero, and precision zero ($\frac{0}{\geq 1} = 0$). The disadvantage of this method is that the intermediate values introduced by the extrapolation lines do not make much sense.

The third method uses extrapolation of all curves to 1.0 precision at 0.0 recall, and is normally used by SMART together with the "Quasi-Cranfield" recall level cut-off. Figure 20 reproduces the four previous examples processed in the indicated manner. In documentary terms, when no documents are examined (cases a) and c)) precision may in a sense be regarded as perfect, hence the 1.0 precision point is used. Cases b) and d) pose a problem for the precision ratio, since retrieval of non-relevant documents only, normally indicates zero precision, but the 1.0 precision ratio is used here for these cases also for reasons of simplicity. As with the second method, the main disadvantage is that intermediate values introduced by the extrapolation lines have no user-oriented meaning.

The fourth method is proposed in an attempt more correctly to reflect precision in cases b) and d), where only non-relevant documents are retrieved. Thus if no documents are retrieved at all a 1.0 precision and 0.0 recall is used; but if non-relevant documents only are retrieved first, then 0.0 precision at 0.0 recall is used. Figure 21 gives the examples, but this hybrid combination of methods 2 and 3 still provides poor meaning to a user.



Ranks of Relevant 1, 2, 9, 12

a) First Document Relevant, Total Relevant > 1

Ranks of Relevant 2, 3, 9, 12

b) First Document now Relevant, Total Relevant > 1

Rank of Relevant 1

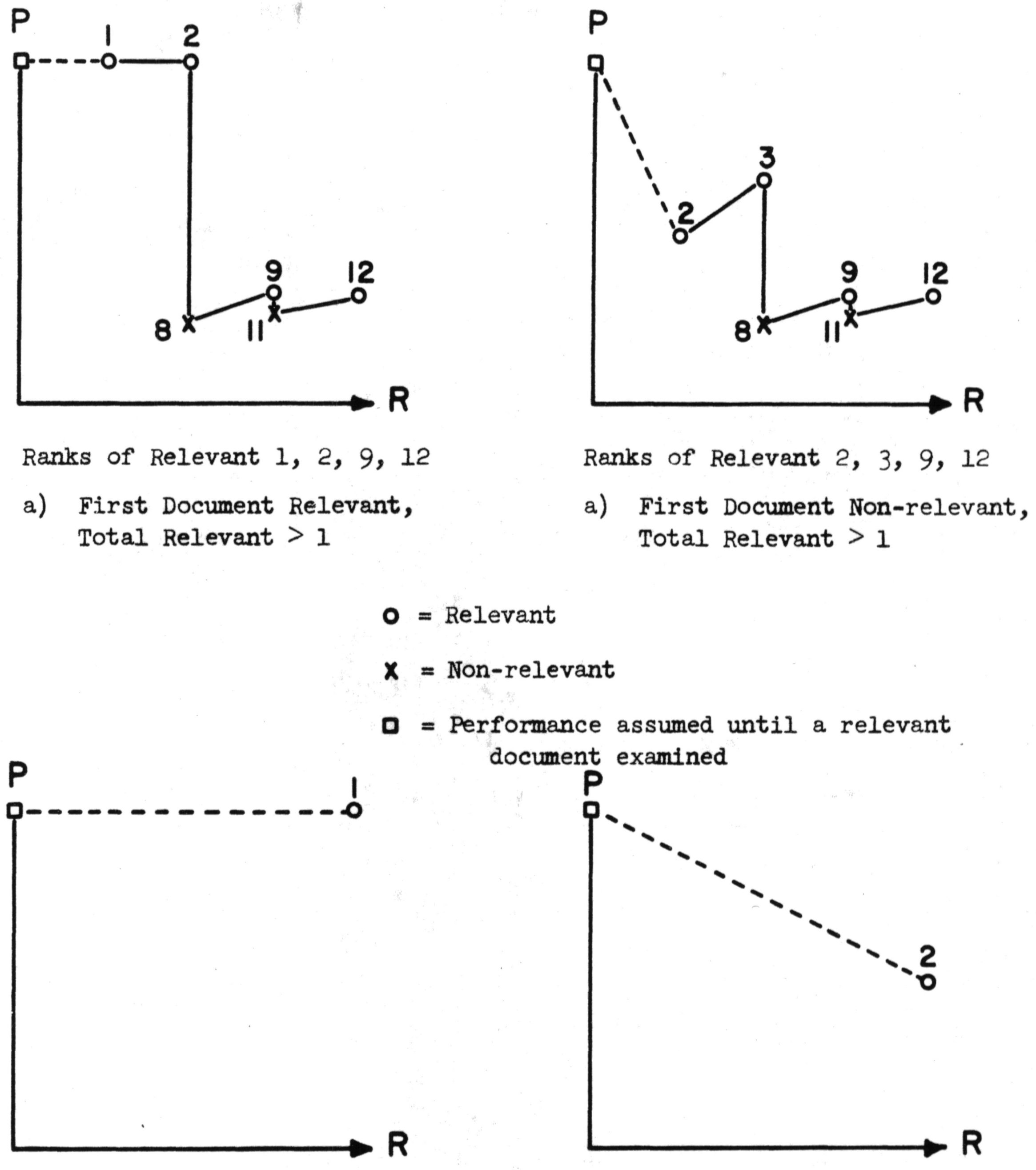
c) Total Relevant 1, First Document Relevant

Rank of Relevant 2

d) Total Relevant 1, First Document Non-relevant

Illustration of Second Method of "Left End Extrapolation"

Fig. 19



Ranks of Relevant 1, 2, 9, 12

a) First Document Relevant, Total Relevant > 1

Ranks of Relevant 2, 3, 9, 12

a) First Document Non-relevant, Total Relevant > 1

Rank of Relevant 1

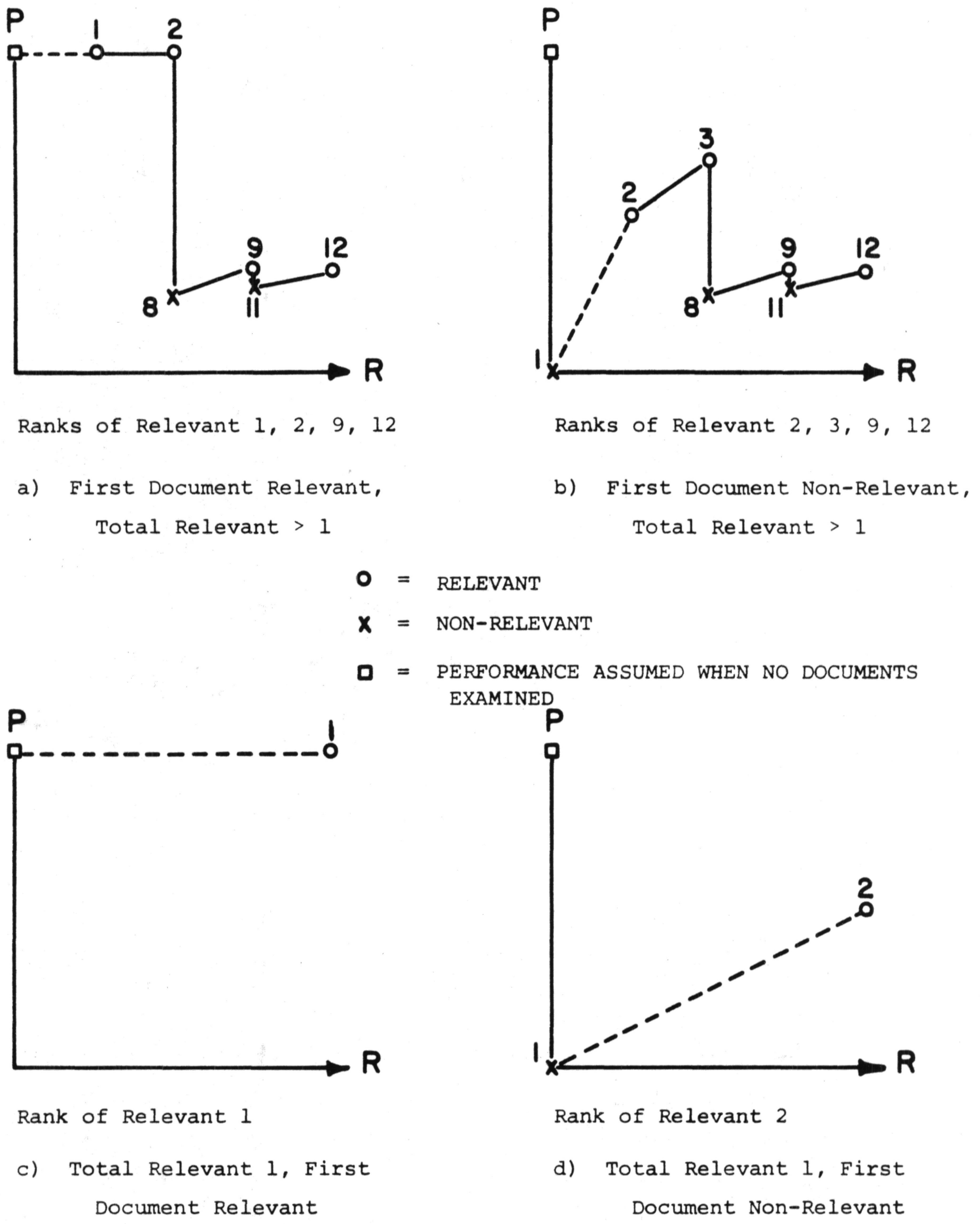
c) Total Relevant 1, First Document Relevant

Rank of Relevant 2

d) Total Relevant 1, First Document Non-relevant

Illustrations of Third Method of "Left End Extrapolation"

Fig. 20



Ranks of Relevant 1, 2, 9, 12

a) First Document Relevant, Total Relevant > 1

Ranks of Relevant 2, 3, 9, 12

b) First Document Non-Relevant, Total Relevant > 1

○ = RELEVANT

X = NON-RELEVANT

□ = PERFORMANCE ASSUMED WHEN NO DOCUMENTS EXAMINED

Rank of Relevant 1

c) Total Relevant 1, First Document Relevant

Rank of Relevant 2

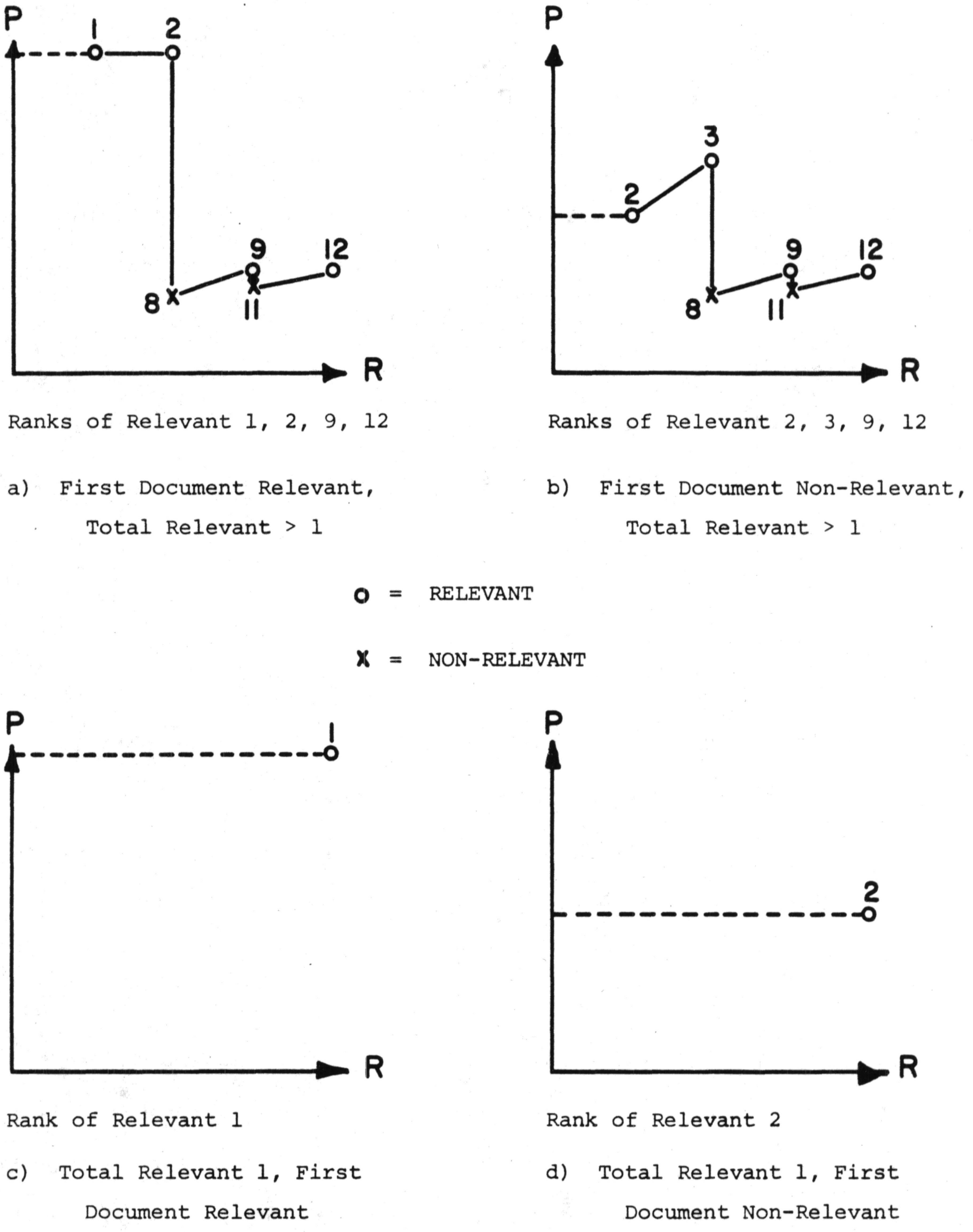
d) Total Relevant 1, First Document Non-Relevant

Illustrations of Fourth Method of "Left End Extrapolation"

Fig. 21

The fifth method uses an extrapolation at constant precision, that is, the precision ratio of the first relevant document retrieved is held constant as the curve is extrapolated to 0.0 recall. Figure 22 includes the four examples for this method. This method has the best documentary interpretation from a user viewpoint, since intermediate points on the extrapolated part of the curve do give an accurate precision ratio that can be achieved at low recall value in cases a) and b), and in cases c) and d) this extrapolation seems to be fairer for averaging purposes than any of methods 2 to 4. This does mean that the precision value at low recall is dependent on the precision achieved when the first relevant document is encountered, and a later relevant document may give slightly higher precision (as in Figure 22 case b)); usually, the extrapolation is sensible.

The foregoing discussion of different techniques for extrapolation is partly an academic one, since in the test comparisons made within SMART comparative merit will not be affected by choice of extrapolation method when the request set is unaltered. Method 3, which has been used in runs made at Harvard, does not correctly indicate merit at the left end of the curve if comparisons involving changes in request sets, or average generality are to be made. For example, three hypothetical requests with differing numbers of relevant items are seen in Figure 23 a) to be badly served by this method at say 0.2 recall, where merit of the three requests is really the reverse of the fact. For this reason, it is preferable that in further work extrapolation method 5 be used. A comparison of methods 3 and 5 is made in Figure 23 b), showing that the difference in curves averaged by a recall level ("Quasi-Cranfield") cut-off is quite small except at the high precision end. If it is thought important to know, at each recall level on the curve, how many of the requests were averaged using an extrapolated part of the individual curves, and how many have enough relevant items to actually enter the average



Ranks of Relevant 1, 2, 9, 12

a) First Document Relevant, Total Relevant > 1

Ranks of Relevant 2, 3, 9, 12

b) First Document Non-Relevant, Total Relevant > 1

o = RELEVANT

x = NON-RELEVANT

Rank of Relevant 1

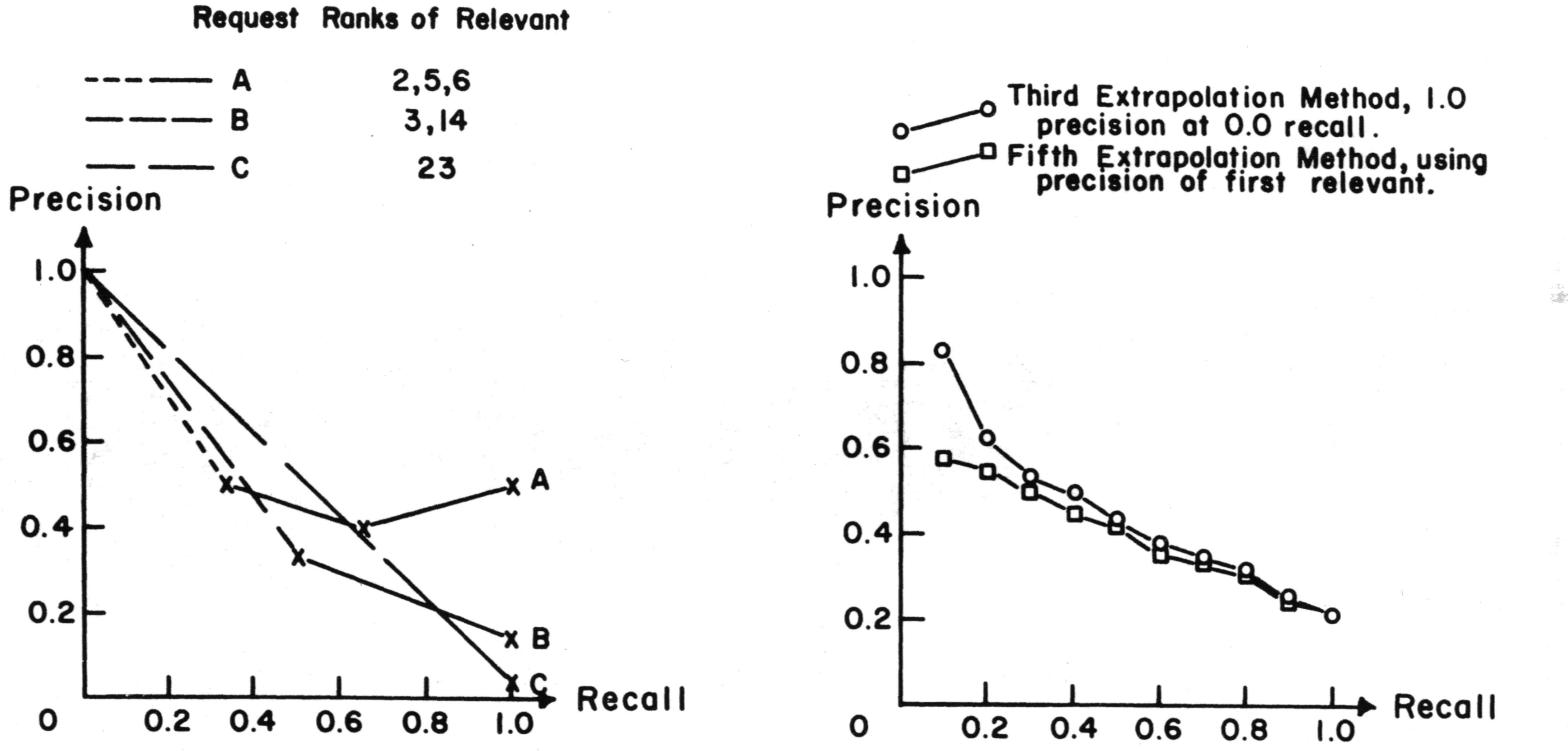
c) Total Relevant 1, First Document Relevant

Rank of Relevant 2

d) Total Relevant 1, First Document Non-Relevant

Illustrations of Fifth Method of "Left End Extrapolation"

Fig. 22



a) Three Individual Requests using 1.0 precision at 0.0 recall

Extrapolation Method

b) Comparison of Two Extrapolation Methods using Averaged Results

Cran-1, Abstracts, Stem, "Quasi-Cranfield" Cut-off, Macro Evaluation of 42 Requests

Comparisons of Extrapolation Methods Three and Five

Fig. 23

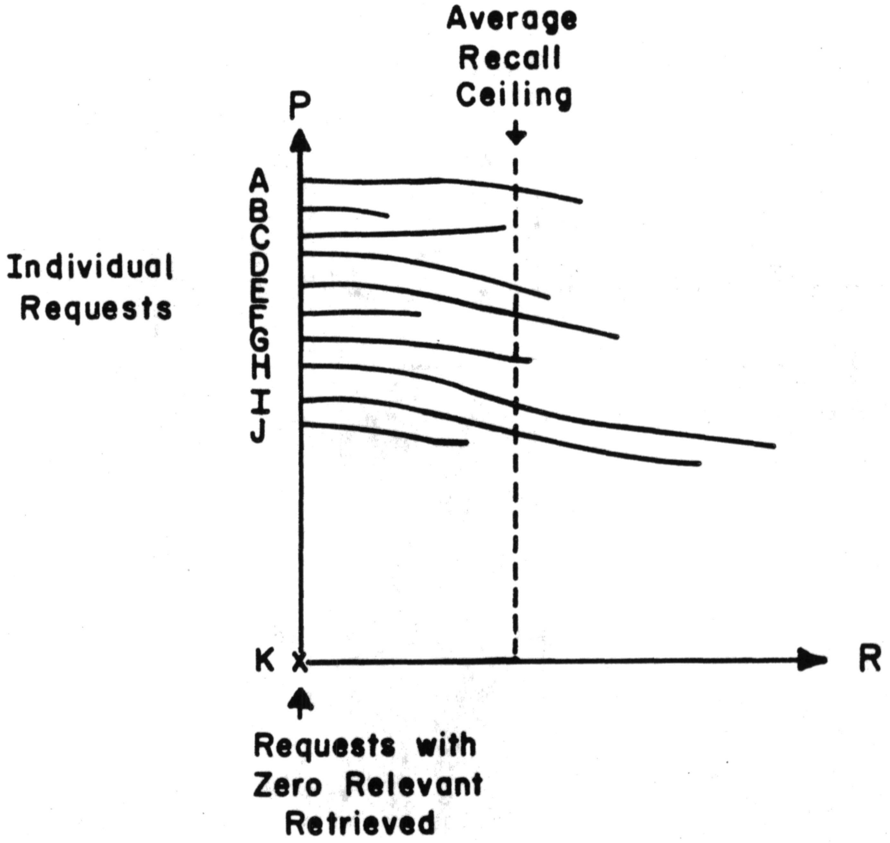
without extrapolation, then this data can be recorded at the ten recall levels on the curve, as was done in Figure 18.

D) Extrapolation Techniques for Evaluation of Cluster Searching

Experiments on cluster searching, many of which are described in report I.S.R.-12, raise an additional problem when precision recall curves of cluster results are to be averaged. The difficulty arises because, when only certain clusters of documents are searched, rather than the total collection, some of the relevant documents are frequently not examined, so that no rank positions exist for some of the relevant documents. This phenomenon is both an expected and an important one, since this "recall ceiling" is one of the vital factors that is used to evaluate cluster searching. An ideal precision curve that would result from a cluster search averaged over many requests would commence in the usual manner at the high precision end but would go only as far as the recall ceiling, thus allowing a comparison with the ordinary full search curve only up to that recall ceiling.

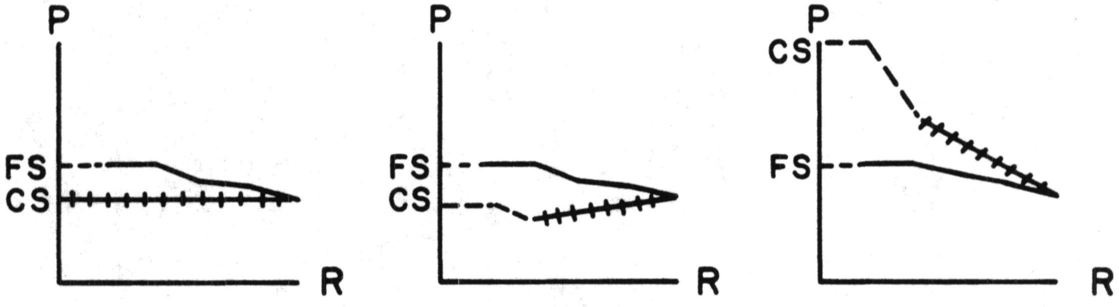
The problem is reflected in Figure 24 for some hypothetical individual requests, it is seen there that some requests naturally do not reach the average recall ceiling, some exceed it, and others are not included on the plot at all, since no relevant documents at all are found in the cluster search. One solution would be to include in the average curve only those requests which supply some results, so that as the average curve approaches the recall ceiling, it would be based on fewer than the total requests. Other methods can also be suggested which employ extrapolation techniques so that every request enters into the whole of the average curve.

The first additional suggested extrapolation technique, has been used exclusively in test results obtained so far with the SMART system. As Figure 25 shows for three individual requests, the recall ceiling reached



Representation of the Cluster Search Evaluation Problem

Fig. 24



a) Cluster Ranks - none b) Cluster Ranks - 3, 7 c) Cluster Ranks - 1, 3

Full Search ----——, Ranks 2, 4, 7, 10, 14.

Cluster Searches, result --- , extrapolation —+++++—

The Extrapolation Technique Used to Construct
Average Precision Recall Curves for Cluster Searches

Fig. 25

by the search results (0.4 in cases b) and c)) is extrapolated linearly to the 1.0 recall points, using the precision gained in the full search. Since the full search curve is drawn by the Quasi-Cranfield cut-off method, this means that cluster results are extrapolated to the precision achieved by the last relevant document in the full search. Figure 25 a) shows what happens to a cluster result in which no relevant documents at all are found: using the left-end extrapolation method recommended in part 4C, the whole cluster curve is an extrapolation from the chosen point at 1.0 recall in the full search curve.

Extrapolation could also be done by assigning to those relevant documents not found in the cluster search a random rank position, bounded by the rank of the last document recovered by the cluster search and the total collection size. It would be feasible also to extrapolate by use of the precision achieved if the relevant documents not found were ranked in the worst possible positions, that is, assuming that recall 1.0 is obtained only as the last document in the collection is examined. A further suggestion is to make use of the full search curve before it reaches 1.0 recall, and use some method of joining the end of the cluster curve to some point along the full search curve.

No comparison of these methods has yet been made, since the technique in use is conceptually as satisfactory as any of the other suggestions.

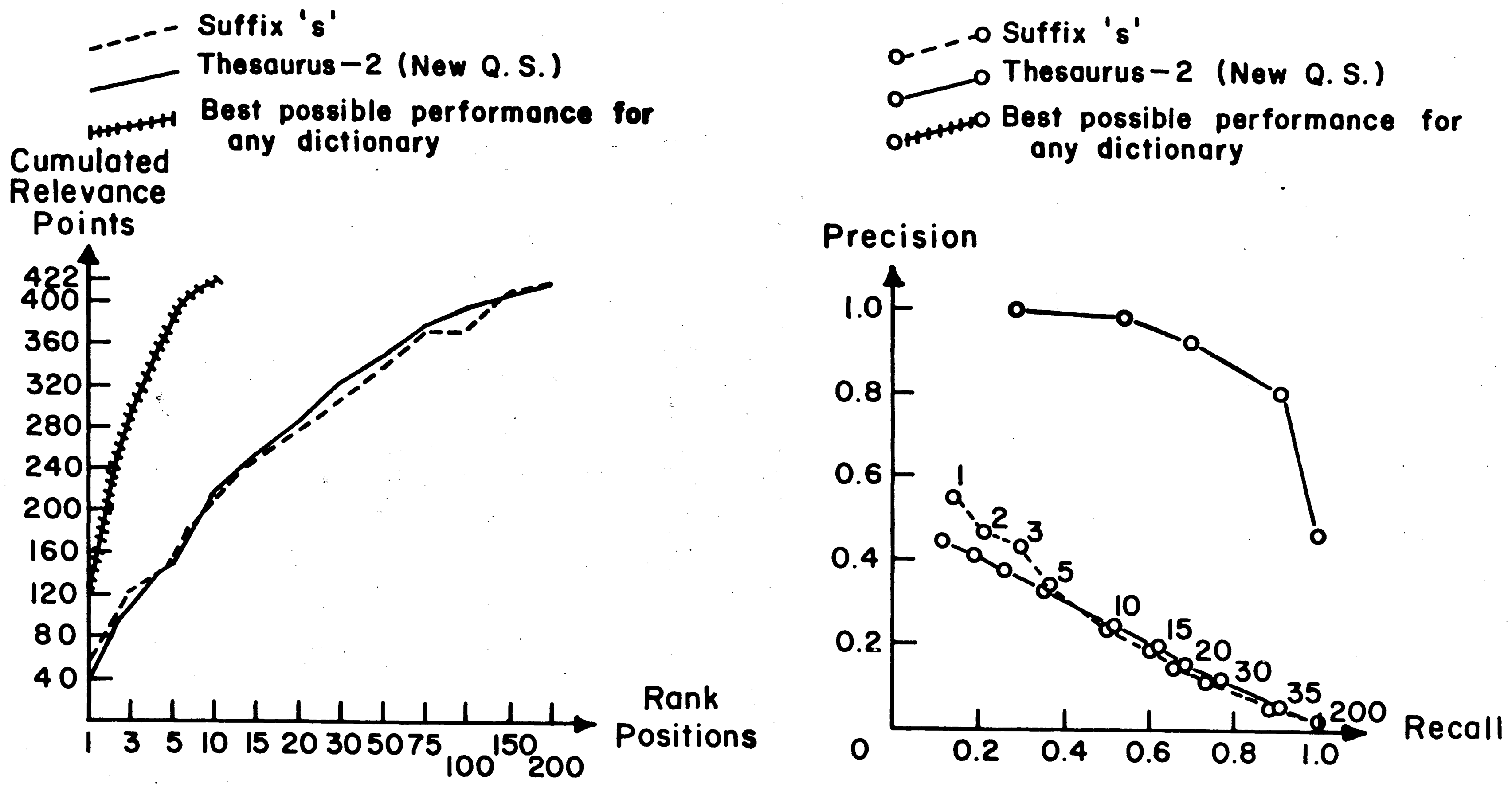
5. Measures for Varying Relevance Evaluation

Although the rendering of relevance decisions is a task quite separate from the considerations which go into the construction of performance measures reflecting system effectiveness, it may be desirable to use performarce measures based on grades of relevance rather than on binary decision of "relevant" or "non relevant" alone. The performance characteristic curve suggested by Giuliano and Jones [8] is designed to use spectra of relevance, since in their view the usual precision and recall can only be used in situations where relevance decisions are black or white. An Example of a performance characteristic curve using relevance grades is given in Figure 26(a). The Cran-1 collection is used because grades of relevance on a scale of four are available for these relevance decisions; thus a "point score" is assigned to those requests, giving a score of four to the most relevant documents, three to the next, and two and one to the final two grades. Figure 26(a) then uses these cumulated relevance points on the y axis as indicating a type of recall, and uses rank positions (cut-off ratio) on the x axis. Two dictionaries are compared, and the best possible performance curve is displayed.

However, as has been demonstrated in [2], it is not correct to assume that precision and recall are incapable of handling relevance grades. Figure 26(b) uses the same data and displays two precision recall graphs, where recall is based on the relevance points score rather than on the more usual document score. In fact, the merit of the two options compared is quite identical — and must be so mathematically — so that the curves cross at the same point; furthermore, the rank position value can be indicated on the precision recall graph as shown. The performance characteristic curve does not give any directly visible information about the amount of non-relevant material being retrieved; the conclusion is then that precision is of value here. Additional precision recall graphs based on relevance grades are given in Section I of this report.

It is also a quite simple matter to modify the single number measures to incorporate grades of relevance. For example, using the normalized recall measure, a "Weighted Normalized Recall" may be defined:

$$\text{Weighted Normalized Recall} = 1 - \frac{\sum_{i=1}^{n} (r_i w_i) - \sum_{i=1}^{n} (iw)}{n(N-n)}$$



Cran-1, Abstracts, Micro Evaluation over 42 Requests.

a) Performance characteristic curve using relevance points

b) Precision recall curve, with recall based on relevance points, "Pseudo-Cranfield" cut-off.

Performance Curves that are Based on Grades of Relevance

Fig. 26

where n = number of relevant documents

N = number of documents in collection

r_i = rank of i^{th} relevant document

w_i = weight score derived from relevance grade of i^{th} relevant document

This equation therefore uses the sum of the products of the ranks and the weight scores of the relevant documents, rather than the sum of the ranks alone as in conventional normalized recall. Some examples given in Fig. 27 will clarify the use of this measure. Fig. 27(a) illustrates a perfect case, where the four relevant documents are given relevance grade weights of 4 (most highly relevant), 3, 2, and 1 (least relevant). Performance in rank position is perfect, as is the order in which the relevant documents are ranked, so a weighted normalized recall of 1.0 results. Fig. 27(b) and (c) show cases of less than optimum relevance grades and ranks, respectively, although both have equal merit in weighted normalized recall. This illustrates the fact that a different range of weights assigned to the relevance grades could be used to adjust the relative effect of ranking and relevance grade ordering. An actual result is given in Fig. 27(d).

6. Measures for Varying Generality Comparisons

The generality number defined in part 2 reflects the concentration of relevant documents in a given collection. From a user viewpoint, the greater the number of relevant documents in a system, the higher probability there is of finding relevant documents at a given cut-off point. Comparing the ADI and Cran-1 collections, for example, although the average request has

	Relevance	
Rank	Grade	Products
1	4	4
2	3	6
3	2	6
4	1	4

Sum of Products 20

Weighted Normed. Recall 1.000

(a) Perfect Ranks and Perfect Relevance Grade Order

	Relevance	
Rank	Grade	Products
1	1	1
2	2	4
3	3	9
4	4	16

Sum of Products 30

Weighted Normed. Recall .9872

(b) Perfect Ranks and Worst Relevance Grade Order

	Relevance	
Rank	Grade	Products
1	4	4
3	3	9
4	2	8
9	1	9

Sum of Products 30

Weighted Normed. Recall .9872

(c) Less than Perfect Ranks and Perfect Relevance Grade Order

	Relevance	
Rank	Grade	Products
3	3	9
13	2	26
19	4	76
41	2	82

Sum of Products 193

Weighted Normed. Recall .7844

(d) Actual Performance of Cran-1 Request Q167, Suffix 's' Dictionary

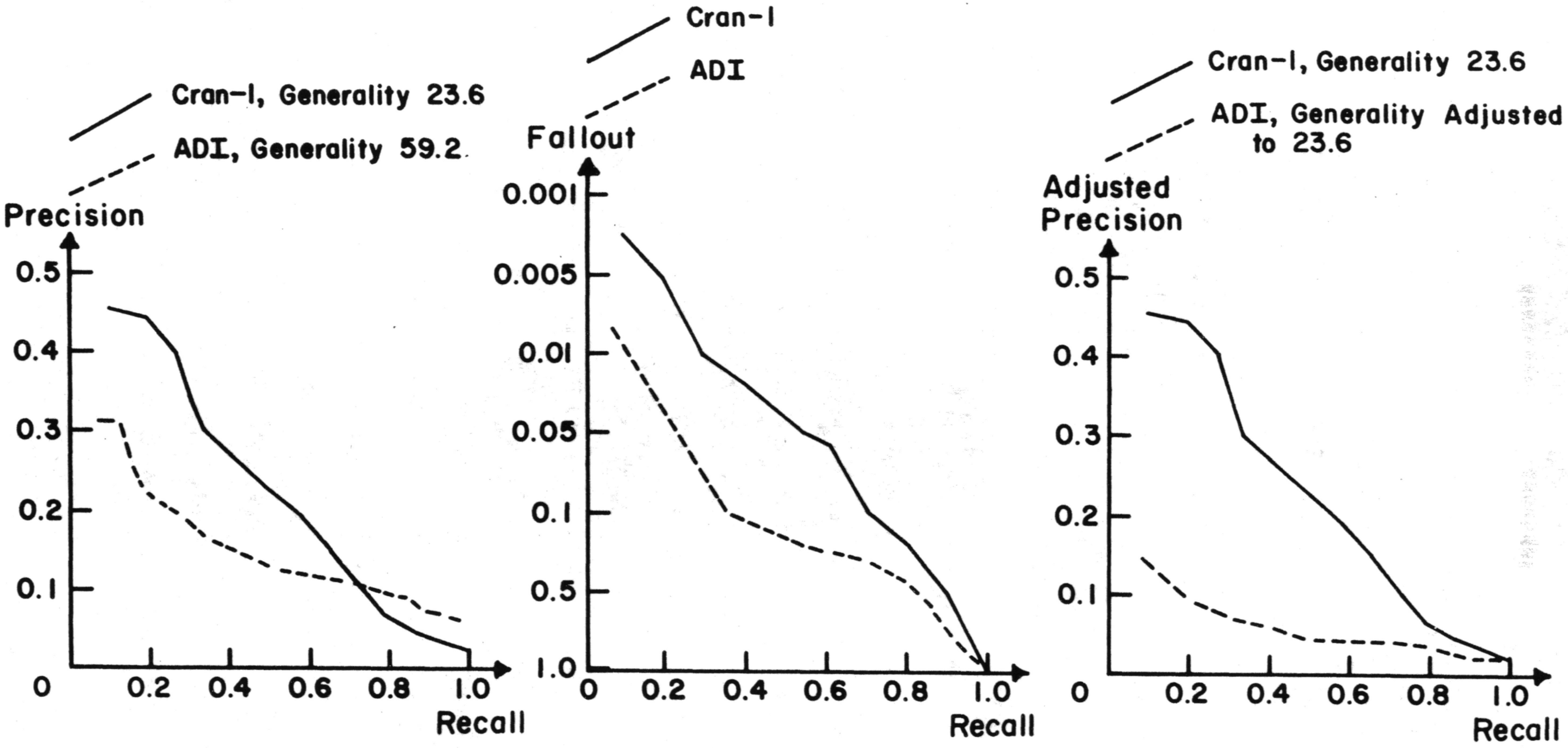
Examples of Use of Weighted Normalized Recall.

Fig. 27.

a similar number of relevant documents in both collections, (4.9 ADI, 4.7 Cran-1), the differing collection sizes (82 ADI, 200 Cran-1) show that the concentration of relevant items favors the ADI collection. This may be observed by imagining a user, who examines every document in both collections in order to be certain of gaining 1.0 recall, and who will finally end up with a precision ratio of .0592 in ADI, and .0236 in Cran-1 (at this cut-off point, the precision ratio becomes the generality number itself). Thus higher precision ratios are expected with higher generality numbers, at all cut-off points in the curve, unless some other factor such as subject language or request and relevance decisions causes some strong effect over the low and middle recall regions of the curve.

Fig. 28(a) includes a comparison of this type using the ADI (Documentation) and Cran-1 (Aerodynamics) collections, where the expected merit is found above 0.7 recall; below that point, the ADI collection falls below Cran-1. The reasons for this result are not important to the present discussion relating to measurements. From the user viewpoint, the comparison in Fig. 28(a) accurately reflects merit, but from a system viewpoint, the change in generality number makes the ADI collection more hospitable to good retrieval than Cran-1; a measure is thus needed to take this into account.

As was suggested in part 2, the value of "d" (Fig. 1, total non-relevant items not examined) is needed for system comparisons, and Fig. 1 also defines the fallout ratio as used in the Cranfield Project [2]. Fig. 28(b) gives a fallout recall graph of the ADI and Cran-1 results, which shows that Cran-1 is now correctly superior over the whole performance range, except at 1.0 recall where both curves meet. It is also possible to represent this system-



a) Precision recall, with variation in Generality.

b) Fallout recall, which allows for Generality.

c) Precision recall with adjustment giving constant generality.

Retrieval Options: Cran-1, Abstracts, Stem, Micro Evaluation Pseudo cut-off averages over 35 requests.
ADI, Abstracts, Stem, Micro Evaluation Pseudo cut-off averages over 35 requests.

Techniques for Performance Comparison in Situations of Different Generality

Fig. 28

viewpoint result in a precision-recall graph, since an equation to adjust precision for generality is given in [2], namely,

$$\text{Adjusted Precision Ratio} = \frac{R_1 \times G}{(R_1 \times G) + F_1(1000 - G)}$$

where R_1 = Recall ratio at a given cut-off point
F_1 = Fallout ratio at a given cut-off point
G = Generality number $\left(\frac{1000 \times \text{total relevant}}{\text{collection size}}\right)$ to which it is desired to alter the results.

Thus, in Fig. 28(c), the ADI recall and fallout ratios are recorded as R_1 and F_1 for a series of cut-off points, and G is set to 2316, in order to adjust the generality of ADI to fit the generality of Cran-1. The adjusted precision versus recall curve is given in Fig. 28(c). It should be noted that the precision for ADI does not now represent a user-oriented evaluation, but has been artificially adjusted to give a system oriented evaluation. A series of tables appears in [2] in which the fallout values for ranges of recall and precision values have been computed, for a range of generality numbers, primarily to permit quick calculation of adjusted precision ratios.

Some comparisons involving changes in generality are given in section I and Appendix A, and further comparisons using the Cran-1 and larger Cran-2 collections will require performance measures of this type. It should be emphasized, however, that the ordinary precision-recall curve still gives a valid and useful user-oriented result, and it is in experimental test comparisons only that the two viewpoints for evaluation (Fig. 2) give different and complementary results. The normalized evaluation measures appear to reflect a system-oriented result since the equations both contain "N", the total number of documents in the collection. For example, the normalized measures corresponding to the presentation of Fig. 28 show results favoring Cran-1. Rank recall and log precision appear to follow the pattern expected of a user-oriented evaluation. However, additional theoretical work is required to establish the nature of these single number measures.

7. Techniques for Dissimilar System Comparisons and Operational Testing

Comparisons between systems of a semi-automatic nature, such as SMART, with more conventional mechanized or manual systems, such as the Medlars system, introduce many theoretical and practical problems. Although direct comparisons of such dissimilar systems are almost impossible to make, one small part of the problem concerning performance measurement can be discussed. This relates to the ability to compare the retrieval performance of a system that produces a ranked output, such as SMART, with a system that conventionally uses a search term matching cutoff, retrieving unordered sets of documents of generally uncontrollable numbers.

For experimental systems that use search term matching cut-offs, such as the Cranfield Project which uses techniques of "coordination levels", it is possible to obtain full precision versus recall curves if very exhaustive search programs are used to establish many cut-off points; the resulting curves can then be compared to the curves produced by SMART. If a direct comparison of this sort is not possible, then an alternative is to apply to the non-ranking system a simple random ranking technique that places relevant documents in random positions in each of the large sets of retrieved documents, as has been done at Cranfield.

For operational system comparisons, however, such exhaustive searching is rarely possible, and tests of such systems usually produce just one precision recall pair, or at the most, three or four quite closely positioned pairs. In such cases, a comparison may be made by making the SMART results fit in with those of the non-ranking system by choosing cut-offs in SMART searches that are in some way identical to the cut-offs made in the non-ranking system. In a quite simple test comparison, for example, the 35 ADI requests were hand searched in a KWIC type concordance of the ADI Abstracts collection, and the result compared with the SMART Abstracts Thesaurus retrieval run (see Section X). The hand searches were based on four or five keywords for each request, and the final performance of what was intended to be a medium-precision at medium-recall search was 0.22 precision at 0.72 recall. Comparison with SMART requires an examination of each individual hand searched request to see how many documents were retrieved, followed by the generation of a cut-off in the SMART ranked output at an identical point to obtain one comparable precision recall pair. The SMART result produced 0.16 precision at 0.64 recall: naturally the hand search benefited from the free choice that was allowed of any synonyms known to the searcher, and higher recall in the hand search would have required choices of further keywords. SMART's fully ranked output would allow high precision at low recall (o.31 precision, 0.31 recall, cut-off 4 documents), or high recall (0.84 recall, 0.11 precision, cut-off 33 documents) simply by examining more or less of the output. Techniques of this type will be used in future comparisons of SMART and Medlars searches using a common set of documents and requests.

A final consideration for evaluation of operational tests pertains to the appropriate measures to be used. Experimental tests of the SMART system have so far measured the recall ratio on the basis of the total relevant items in the collection. Although this accurately simulates users with a high recall requirement, those users with a high precision requirement are probably not too well served by the high precision end of the same curve. The reason is that at least some users wanting high precision are not at all concerned about getting high recall, and since they wish only to see, say one, two, or three relevant items, they are clearly satisfied on the recall side long before 100% recall of the total relevant items in the collection is achieved. It is suggested that in semi-operational tests that will be made in SMART in the future, a "Relative Recall" be computed:

$$\text{Relative Recall} = \frac{\text{Total Relevant Examined}}{\text{Total Relevant User Would Like to Examine}}$$

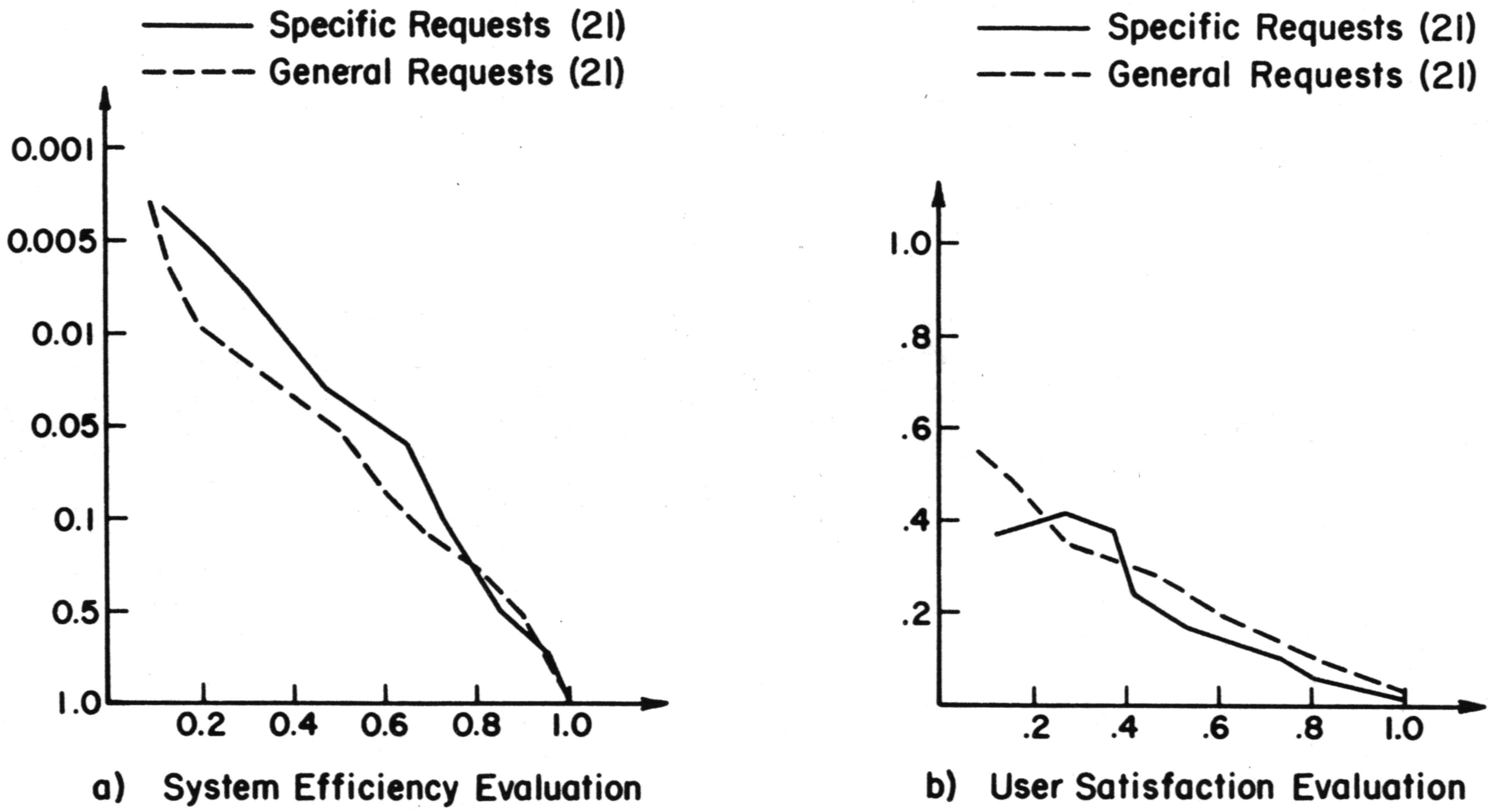
This ratio is relative to user satisfaction rather than to toal system resources. Several adjustments might be made for actual tests, since some users would perhaps examine more relevant than they intended (1.5 recall would not be very useful for evaluation purposes), and other users might wish to see more relevant than were available in the system at all (an acquisitions, rather than retrieval failure).

8. The Comparison of Specific and General Requests and the Viewpoints of the "higher precision" and "high recall" user.

The comparison of a set of 'specific' requests with a set of 'general' requests provides an environment of acute change in request generality. Isolation of specific from general requests is carried out by dividing a given request set into equal or nearly equal groups according to the numbers of documents in the collection that are relevant. The comparison of the specific and general request sets then involves a very large change in average generality, although the collection size is unaltered. To illustrate further the problems caused by this type of comparison the set of 21 specific requests will be compared with the 21 general requests in the Cran-1 aerodynamics collection, using the stem dictionary results.

Since the generality change suggests that fallout should be used in place of precision, a fallout versus recall plot is given in Fig. 29(a). Apart from a slight crossing of the curves between .8 and .9 recall, the specific requests are seen to have a superior performance, from the point of view of system efficiency. The precision versus recall plot, however, will reflect a direct performance comparison ignoring the generality change, so a plot of this type is given in Fig. 29(b) where it is now seen that except between .25 and .4 recall, the general requests have a superior performance. It should be noted that a Pseudo-Cranfield type of cut-off is used here for comparison of specific and general requests, since a plot of the Quasi-Cranfield type as used in [4] give a large bias in favor of the specific requests. This occurs because the specific requests all require greater lengths of left end extrapolation and the technique used for extrapolating to 1.0 precision at 0.0 recall (method 3, part 4c, Fig. 20) gives the specific requests falsely high precision values at low recall.

A partial explanation for the facts reflected in Fig. 29 is shown by the data in Fig. 30. At each of the cutoff points shown, the general requests produce a greater number of relevant and a smaller number of



a) System Efficiency Evaluation

b) User Satisfaction Evaluation

Cran-1 Abstracts Stem, Micro Evaluation, Pseudo-Cranfield Cutoff, Averages over 21 Requests

Fig. 29

PERFORMANCE	SPECIFIC REQUESTS (63 relevant)					GENERAL REQUESTS (135 relevant)				
	CUT-OFF, n DOCUMENTS					CUT-OFF, n DOCUMENTS				
	n= 1	2	3	4	5	n= 1	2	3	4	5
TOTAL RETRIEVED	21	42	63	84	105	21	42	63	84	105
RELEVANT RETRIEVED	8	17	24	25	26	11	20	27	33	38
NON-RELEVANT RETRIEVED	13	25	39	59	79	10	22	36	51	67
RECALL	.13	.27	.38	.40	.41	.08	.15	.20	.24	.28
FALLOUT	.00314	.00604	.00943	.01426	.00910	.00246	.00541	.00886	.01255	.01648
PRECISION	.38	.41	.38	.30	.25	.52	.48	.43	.39	.36

Cran-1 Collection, Abstracts, Stem dictionary, Macro Average
over 21 Specific and 21 General Requests.

Details of Comparison of Specific and General Requests
at the First Five Document Cut-offs.

Fig. 30

non-relevant compared with the specific requests. But also at each cutoff, the recall ratios favor the specific requests, and both fallout and precision ratios favor the general requests. The recall merit is explained by noting that the smaller number of total relevant in the specific requests means that each one found "counts" for more in recall, since one relevant found in the specific requests increases recall by .016, and one relevant found in the general requests increases recall by .008. The fallout and precision merit is clearly affected by the higher concentration of relevant documents that is found in the general requests. It is not clear that fallout is free from the effects of generality in this sense, and therefore it is not certain that the fallout versus recall plot truly reflects system effectiveness when a generality change of this type is encountered. Also, since recall is here affected by the difference in request generality, it is not certain that recall accurately reflects user satisfaction, although it obviously does measure what the user examines.

This last difficulty arises because it is not really clear just how the positions should be weighed when specific and general requests are compared. Six cases for comparison are given in Fig. 31: if some rational hand ranking of the merit of these six requests is not possible, then no satisfactory performance measure to compare specific and general requests can be derived. One obvious solution is to recognize formally what has often been stated, namely, that users' needs differ considerably, and the two ends of the spectrum may be represented by the high recall need and the high precision need. For example, if the high precision need is defined to mean that the best precision should be obtained in the process of finding just two relevant documents only, then the cases A, D, and F, in Fig. 31 are superior to B, E, and C. Also, if the

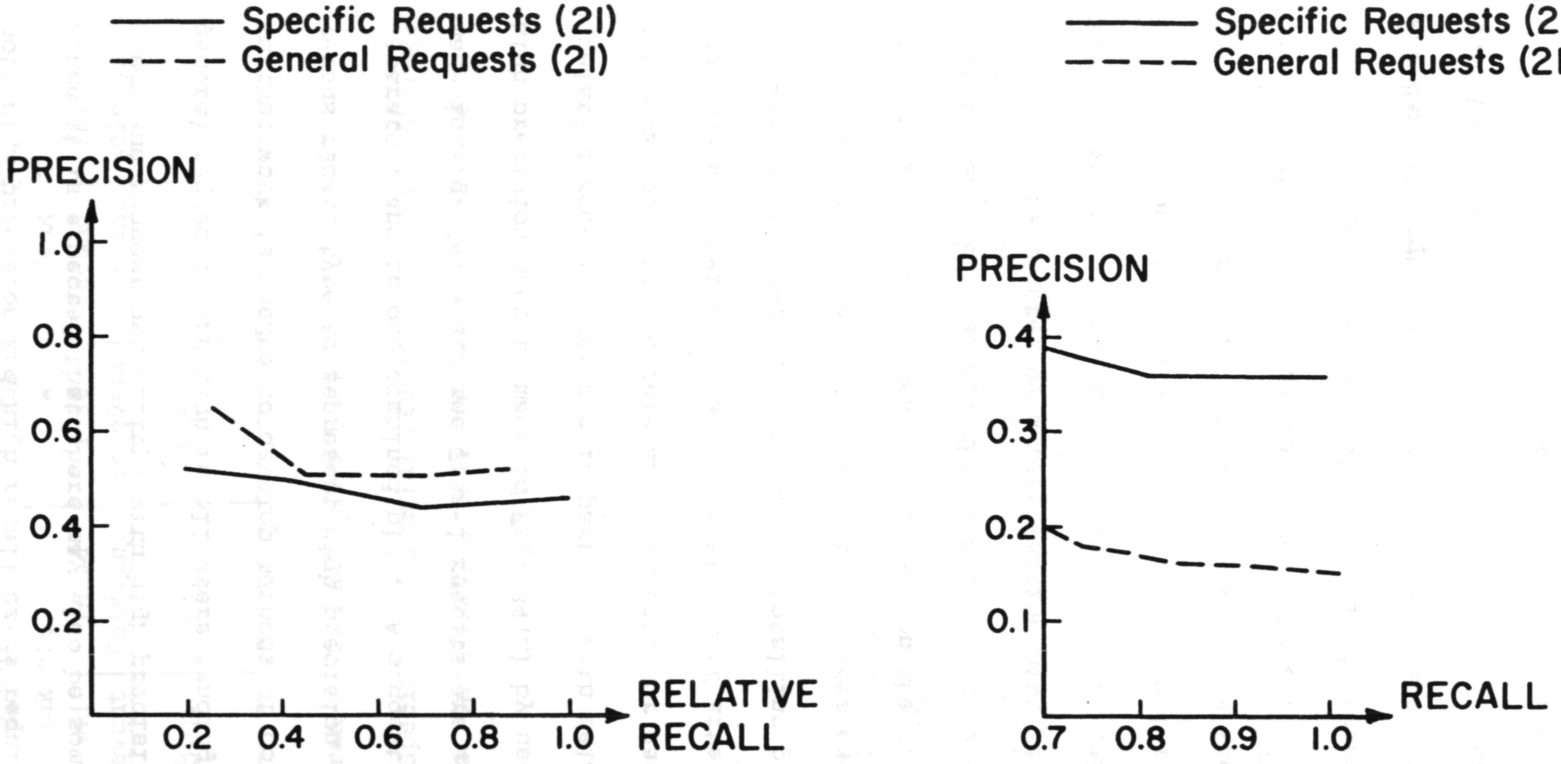
Generality	Request	Ranks of Relevant
SPECIFIC	A	1,2,10
	B	3,4,17
	C	7,21,45
GENERAL	D	1,2,5,7,8,9,14,15
	E	3,7,10,22,33,37,49,51
	F	1,2,8,10,11,29,36,47

Rank Positions of the Relevant Documents for Six Hypothetical Requests.

Fig. 31.

high recall need is defined to mean that a full 1.0 recall is required, then the best performance will be achieved when perfect recall is quickly reached and has high precision, so that in Fig. 31 cases A, D, and B are superior to C, F, and E. Making the further distinction that A, B, and C are specific and D, E, and F are general requests, this hypothetical example shows that the high precision user is served best on the average by the general requests, and the high recall user by the specific requests.

The cases in Fig. 31 are chosen to be typical of the results obtained in the Cran-1 request sets being used, and full discussion of these results appears in section I part 6B. One method of presenting average results that reflects the success achieved in meeting the two different types of user need is given in Fig. 32. The high precision and high recall needs are based on the definitions given in the previous paragraph. An average rank position is thus calculated for the first and second ranked relevant documents (for a high precision merit), and for the last ranked relevant document (for a high recall merit). It can now be concluded that the high precision user is served best by the general requests, and the high recall user by the specific requests. However, the computation of the arithmetic mean rank is sometimes a poor representation of the data since the variance can be large and one or two very bad requests can unduly influence the average. Some type of histogram would solve this problem, but at the cost of a somewhat more complex presentation. One compromise solution is suggested by Fig. 33, where data on the rank of the first relevant is re-arranged to show the numbers of search requests that gave a given rank (in three ranges) to the first relevant.



a) Evaluation from Viewpoint of User Satisfaction for High Precision Users.

b) Evaluation from Viewpoint of User Satisfaction for High Recall Users.

Cran-1 Abstracts Stem, Micro Evaluation, Pseudo Cut-off with Final Cut-off at Last Relevant, Average over 21 Requests

Fig. 34

measures provide the same merit between specific and general requests as that shown by the fallout versus recall plot. This means that the normalized measures tend to reflect the merit experienced by a high recall user. Some of the comparisons made in Section X (see part 5c) are thus seen to be recall oriented.

9. The Presentation of Data as Individual Request Merit

Evaluation using averaged measures always suppresses some data, and when the variance of individual requests is large, the arithmetic mean may be a poor measure of merit. For this reason, presentation of results using averages in section III, V, VI, and VII is followed in each case by data on the individual requests. For example, if normalized recall averaged over 12 requests shows one option to be quite superior to another, individual request examination might reveal that 6 of the requests favored the option that was superior or average, 4 favored the other option, and 2 had an identical performance on both options. The tables used to present this type of data usually give both the numbers of requests favoring each option together with those equal on both options; in addition, percentages are produced to aid speedy interpretation. Several ways of computing the percentages are possible, and six methods are illustrated in Fig. 35. The technique adopted for the sections listed is nearly always that of giving percentages for each option ignoring the commonly few cases where both options have equal merit. Percentages including the equal cases are needed when the number of such cases is large, and may be given either in the form of row 3 or 4. The "Superiority Percentage" has the advantage of a single number presentation.

	Number and Percentages of Individual Requests			
	Option I	Option II	Both Options Equal	"Superiority" Percentage
1 Numbers	6	4	2	
2 Percentages ignoring equal cases	60%	40%	-	20%
3 Percentages including equal cases	50%	33%	17%	17%
4 Percentages adding equal cases to both options	67%	50%	-	17%

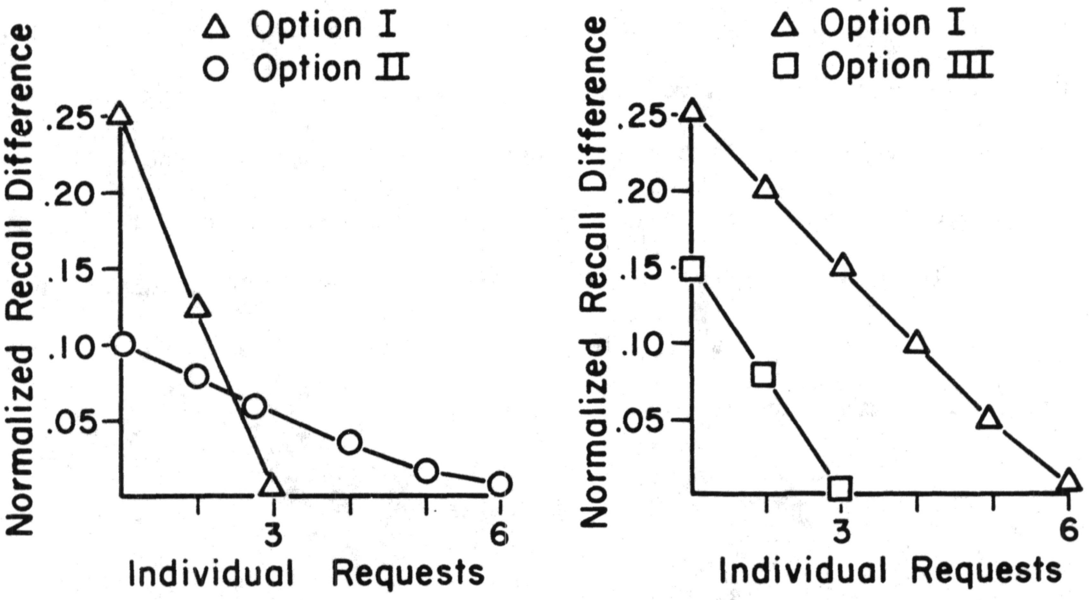
Illustration of Methods of Computing Percentages of the Numbers of Requests Favoring Given Options.

Fig. 35

An extension of this type of comparison is the presentation of the magnitudes of the differences in the merit of individual requests. A set of nine hypothetical request results is given in Fig. 36, comparing three options. A table of the numbers of requests preferring options I and II would show that 66-79 prefer option II, and 33.3% prefer option I. However, since the average normalized recall values given in Fig. 36(a) show that options I and II have almost identical merit, it is clear that the three requests preferring I over II do so by quite large amounts, and the six preferring II over I by smaller amounts. The magnitude difference plot in Fig. 36(b) is designed to show this situation visually. The requests favoring each option are arranged in decreasing order of their performance differences accross the plot, and since the areas underneath both curves are nearly equal, this reflects the fact that both options have nearly identical averages. Further, since the option I curve terminates some way short of the option II curve on the x axis, this indicates that more individual requests favor II. Another comparison is given in Fig. 36(c), where option I is seen to be superior to option II both in the averages and in the individual requests.

	Normalized Recall		
Req.	Option I	Option II	Option III
A	.8125	.8000	.8135
B	.8250	.8000	.8135
C	.8010	.8000	.7810
D	.8000	.8100	.7850
E	.8000	.8080	.7900
F	.8000	.8060	.7950
G	.8000	.8040	.7990
H	.8000	.8020	.8080
I	.8000	.8010	.8150

a) Results of Nine Hypothetical Requests



6 Requests better on Option III (66.7%)

3 Requests better on Option I (33.3%)

6 Requests better on Option I (66.7%)

3 Requests better on Option III (33.3%)

b) Plot of Differences Comparing Options I and II.

c) Plot of Differences Comparing Options I and III.

Illustration of Method Used to Compile Magnitude Difference Plot

Fig. 36

References

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

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

[3] J. A. Swets, Effectiveness of Information Retrieval Methods, Report No. 1499, Bolt Beranek and Newman Inc., Cambridge, Mass., Draft, April 1967.

[4] 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).

[5] J. J. Rocchio, Evaluation Viewpoint in Document Retrieval, Information Storage and Retrieval, Report ISR-9, to the National Science Foundation, Section XXI, Harvard Computation Laboratory, August 1965.

[6] J. J. Rocchio, Document Retrieval Systems — Optimization and Evaluation, Doctoral Thesis, Report ISR-10, to the National Science Foundation, Harvard Computation Laboratory, April 1966.

[7] J. A. Swets, Information Retrieval Systems, Science, Vol. 141, 19 July 1963.

[8] V. E. Giuliano and P. E. Jones, Study and Test of Methodology for Laboratory Evaluation of Message Retrieval Systems, Interim Report ESD-TR-66-405, Decision Sciences Laboratory, Electronic System Division, U.S.A.F., Contract to A. D. Little Inc., August 1966.

[9] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing — Automatic Indexing Methods are Evaluated and Design Criteria for Modern Information Systems are Derived, Report ISR-12, to the National Science Foundation, Section III, Cornell University, June 1967.

[10] S. J. Sillers, Distinguishing Retrieved from Nonretrieved Information: The Cut-off Problem, Information Storage and Retrieval, Report ISR-9, to the National Science Foundation, Section XXII, Harvard Computation Laboratory, August 1965.