KEY PAPERS
in
INFORMATION SCIENCE

edited by
Belver C. Griffith

Published for the
American Society for Information Science
by
Knowledge Industry Publications, Inc.
White Plains, New York

KEY PAPERS IN INFORMATION SCIENCE

edited by Belver C. Griffith

Library of Congress Cataloging in Publication Data
Main entry under title:

Key papers in information science.

Includes index.
1. Information—Addresses, essays, lectures.
I. Griffith, Belver C. II. American Society for Information Science.
Z699.K422 029 79-24288
ISBN 0-914236-50-4

Printed in the United States of America

TABLE OF CONTENTS

Part III - The Structure of Literature and Documents

Part IV - Information Retrieval and Analysis

Part V - Tools and Ideas: The Interface of Information Science and Librarianship

FRONTISPIECE

A Map of Some Major Information Science Writers

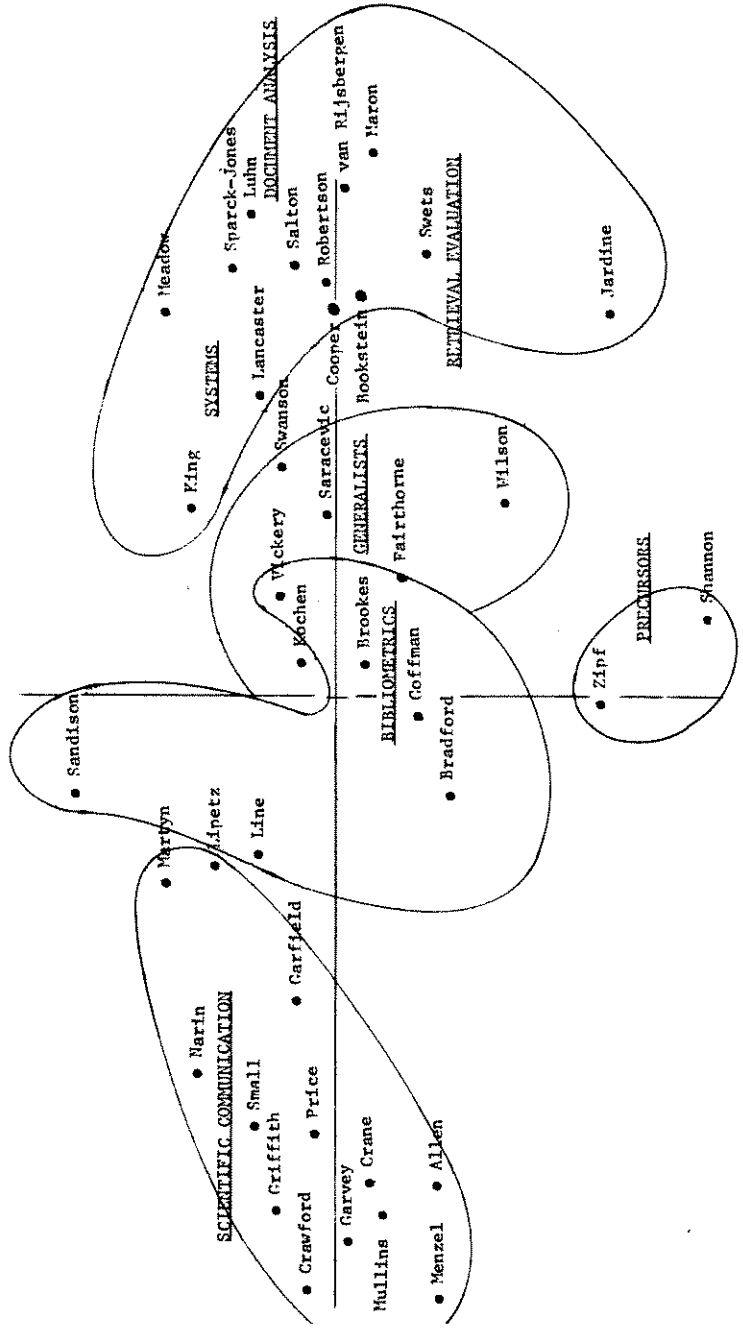
Citations of earlier published works have been used to map the position of documents, as in the paper by Small included in this collection. White's recent work on combining authors to retrieve from the *Science Citation Index* (SCI) and the *Social Science Citation Index* (SSCI) now permits the mapping of authors.

Howard White and the editor generated this figure based upon the co-citations of the entire work of some major writers in information science. Distances between pairs of authors were estimated through a procedure which starts with counts of co-citations of the entire published work of pairs of authors by any citing paper covered in seven years (1971-1978) of SSCI. Raw co-citations across the matrix of 39 authors were then subjected to Pearson product-moment correlation procedure correlating pairs of authors (after a suitable transform was employed to create diagonal scores). The resulting correlation coefficients were subjected to Kruskal's MDSCAL method of non-metric multidimensional scaling. All normal criteria for this procedure indicate that the spatial model fits the data well.

The author list was taken from this volume after first eliminating persons who were too rarely cited to generate reliable results. To that was added a list of well-known writers, including some precursors (e.g. Luhn) whose work is not included herein. While certainly not complete, the list and resulting map tell us where most "missing" persons would go.

The regions were delineated through another clustering procedure, the labels coming from the mappers' imagination. The origin is set by procedure; most generalists, as for example, Brookes and Kochen, lie nearby. The axes are arbitrary, being set to run horizontally between the centers of the scientific communication group and the information retrieval group. The vertical axis seems to represent different variables on the left, center and right sectors of the figure. At the left, lower writers use only behavioral measures to study scientific communication; higher writers, only citation measures. In the center, there is some broad opposition of theory (Shannon and Zipf) and practice (Sandison and those persons involved in system design and evaluation). At the far right, persons who evaluate retrieval seem to oppose persons who develop methods of document analysis.

The figure is a unique attempt of a field to display itself using its own methods.



ACKNOWLEDGEMENTS

My principal acknowledgement is to the researchers and scholars whose efforts justify this volume and whose products are herein embalmed. The dreadful logistics of permissions, originals, etc. were ably managed by Lynn A. Deglin in a several months' orgy of tactfulness, intelligence and attention to detail. A chance debate with Gerard Salton revealed that we had similar views on the value of literature on information retrieval; he very kindly reviewed my selections and made helpful additional suggestions; M. Carl Drott and Michael E.D. Koenig contributed to the introduction to that section. Howard D. White contributed the seminal ideas leading to the development of the technique used in creating maps of authors, as in the frontispiece; we shared the technical work. The idea of an appendix listing works citing the Key Papers was suggested by Roger Needham of the University of Cambridge Computer Laboratory and kindly transmitted by his wife, Karen Sparck Jones; Patricia N. Servi assisted in creating the appendix. Portions of the work, especially those exemplifying and displaying techniques developed by coworkers, Henry Small, M. Carl Drott, Howard White and myself, were funded by PHS Grant #ROI LM00911 07 from the National Library of Medicine.

Belver C. Griffith

DEDICATION

To those persons who have, over the period covered by this volume, instructed most of the rest of us — especially,

Bertram C. Brookes
Robert A. Fairthorne
Manfred Kochen
Derek J. de Solla Price

Reprinted from Information Processing & Management, Vol. 15.

EXPERIMENTS IN RELEVANCE WEIGHTING OF SEARCH TERMS

K. Sparck Jones
Computer Laboratory, University of Cambridge, Corn Exchange Street, Cambridge, England

(*Received* 15 *September* 1978)

Abstract—Following successful initial tests of theoretically-based schemes for relevance weighting of search terms, further experiments were undertaken to validate these results. The experiments were designed to investigate weighting for a large document set, poor matching conditions, heterogeneous data, and limited relevance information, i.e. the use of weighting in more realistic conditions than the initial ones. The results confirm the earlier ones: very striking improvements in retrieval performance were obtained, especially for the theoretically best-founded weighting formula. The experiments illustrate a much more promising application of statistical methods to indexing and searching than any studied hitherto.

This paper describes large scale experiments to test the value of index term weighting based on relevance information, and presents results showing that weighting leads to remarkable improvements in retrieval performance.

RELEVANCE WEIGHTING

Robertson and Sparck Jones[1] discussed the use of relevance weights for index terms, i.e. term weights derived from information about the occurrences of request terms in known relevant documents; presented a statistically-based approach to the computation of such weights; and, with reference to relatively limited experiments, showed (a) that substantial improvements in retrieval performance could be obtained through the use of these weights, and (b) that formulae with certain specific bases work better than others.

This paper describes experiments designed to investigate these methods of relevance weighting more fully. In particular, since it can be proved (see Yu and Salton[2], Robertson[3]) that relevance information in given circumstances must improve retrieval performance in simple term matching, the experiments attempted to show whether favourable conditions are likely to be encountered in practice. It might indeed be argued, though somewhat tentatively, that if the experimental material is sufficiently varied, the tests can be regarded as falsifying rather than confirming ones.

The theoretical basis of our approach to weighting is fully described in [1], and so is simply summarised for convenience here. Specifically, we assume simple binary term document and request characterisations, with the terms treated independently in matching. The information available for weighting a given request term (assuming some documents are known to be relevant to the request) is expressed in the following contingency table:

		relevant: +	relevant: −	
indexed	+	r	$n-r$	n
	−	$R-r$	$N-n-R+r$	$N-n$
		R	$N-R$	N

where for term t in request q: n = the number of documents containing t, r = the number of relevant documents containing t, N = the number of documents in the collection, and R = the number of relevant documents for q.

The exploitation of this information in a probabilistic weighting formula is determined by a choice of an underlying Independence Assumption about the data and an Ordering Principle for

the search output. There are two possible Assumptions I1 and I2, as follows: I1, the distribution of terms in relevant documents is independent, and their distribution in all documents is independent; I2, the distribution of terms in relevant documents is independent, and their distribution in non-relevant documents is independent; and two possible Principles 01 and 02, as follows: 01, that probable relevance is based only on the presence of search terms in documents; 02, that probable relevance is based on both the presence of search terms in documents and their absence from documents. Assumptions and Principles may be combined in four ways giving four weighting formulae F1–F4:

	I1	I2
01	F1	F2
02	F3	F4

Thus the formulae define the weight of an index term as follows:

$$w = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n}{N}\right)} \tag{F1}$$

$$w = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n-r}{N-R}\right)} \tag{F2}$$

$$w = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n}{N-n}\right)} \tag{F3}$$

$$w = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n-r}{N-n-R+r}\right)} \tag{F4}$$

The development of these formulae is outlined in the Appendix to [1]. It was argued in [1] that theoretically F4 should perform best, and this was borne out by experiment in that the other formulae either performed less well, or no better. F1, which is closest to relevance formulae used in other work (see Miller[4], Barker *et al.*[5] and Robson and Longman[6]) performed least well. Results for F2 were similar to those for F1 and F3 to F4. Since F1 and F4 represent the most strongly contrasted formulae, subsequent work has been confined to them.

In searching, documents are ordered by their sum of weights score, but only documents with positive scores are deemed retrieved.† Performance is evaluated using recall and precision: specific results are obtained by averaging by numbers at matching value levels, with subsequent linear interpolation to obtain precision at ten standard recall levels.

It is evident that the primary use of relevance information is in predictive weighting, i.e. in the derivation of weights from given data for use in new searches. In the experiments reported in [1] the value of predictive weighting was shown by the application of weights derived from half of a test collection, the even-numbered documents, in searching the other, odd-numbered half. Thus for 225 requests and 700 out of 1400 documents characterised by relatively long, manually-obtained term lists, requests with relevance weights performed substantially better than unweighted requests; relevance weighting was also much superior to simple weighting by (inverse) term collection frequency alone, itself superior to unweighted matching. For prediction the weighting formulae are modified to allow for uncertainty by the addition of 0.5 to the contingency table components. Thus F1 and F4 respectively become

†F4 requires a scaling constant: see [1].

$$w = \log \frac{\left(\frac{r+0.5}{R+1.0}\right)}{\left(\frac{n+1.0}{N+2.0}\right)} \quad \text{and} \quad w = \log \frac{\left(\frac{r+0.5}{R-r+0.5}\right)}{\left(\frac{n-r+0.5}{N-n-R+r+0.5}\right)}.$$

The formula for collection frequency weighting is the simple one

$$w = -\log\left(\frac{n}{N}\right),$$

implemented as

$$w = -\log\left(\frac{n}{\max n}\right). \qquad \text{(F0)}$$

This formula is not directly related formally to the relevance formulae, but is clearly based on a similar statistical approach.

Relevance weights may, however, also be applied retrospectively to establish optimal performance for the weighting formulae used, and hence a performance yardstick. The retrospective weighting tests also reported in [1] showed that for the relatively small collections used, a high level of performance is possible. SPARCK JONES[7] suggested that such yardsticks have a wider utility in the evaluation of indexing and searching experiments generally. The retrospective application of the relevance weighting formulae requires special treatment of limiting cases where components = 0. These are fully detailed in [1]. Unfortunately retrieval runs with special cases are expensive, so to simplify matters weights computed with the predictive formulae may be applied retrospectively: the performance obtained is still useful as a yardstick, though it tends to be lower than that given by the "pure" formulae.

EXPERIMENTAL OBJECTIVES

The object of the present series of experiments was to test relevance weighting more adequately. They were designed (a) to test our initial results with new data for (i) the performance of relevance weighted terms compared with collection frequency weighted terms (F0), (ii) the performance of relevance weighted terms compared with unweighted terms, and (iii) the performance of terms relevance weighted by F4 compared with terms weighted by F1; and (b), more importantly, to establish the possible practical limits on the utility of relevance weighting. In fact, tests appropriate to (b) would cover (a) as well.

Specifically, tests for the utility of relevance weighting should provide answers for the following questions:

(1) is relevance weighting impervious to scale effects, i.e. is it effective for the large collections encountered in real life?

(2) is relevance weighting useful in poor matching conditions, i.e. when few index keys are available for requests and documents?

(3) is relevance weighting helpful for heterogeneous collections, i.e. where the document files consist of only partly-related subject groupings?

(4) is relevance weighting effective even where little relevance information is available?

Ideally, such questions should be answered for the stringent conditions of modern on-line search systems where files are large and heterogeneous, a search specification may be filled out only slowly, and few documents may be evaluated on each search iteration. But there are many methodological and practical difficulties about adequate, controlled, on-line search experiments, and we thought it would be more useful to start by investigating relevance weighting in less stringent but more realistic conditions than those of our first experiments. The tests reported here therefore model SDI operations, with real data. Their limitations and the direction of future experiments are considered in the conclusion.

EVALUATION PROBLEMS

The main problem about conducting any experiments designed to answer questions (1)–(4) is of obtaining adequate test data. The tests described in (1) used a small collection with requests

and documents indexed in an appropriate way by natural language word stem lists, and for which exhaustive relevance assessments were available. Weighting was therefore simply a natural extension of the coordination searching used in other experiments with the collection (see for example CLEVERDON *et al.*[8] and SPARCK JONES and BATES[9]), and the calculation of weights and evaluation of output presented no problems since the relevance status of every document in the collection was known.

Limited relevance information makes evaluation difficult when new search methods retrieve unassessed documents. If evaluation is geared to previously assessed output a completely false picture of performance for the new methods may be obtained; thus if the assessed output is associated with a specific search method, characterising performance for a new strategy only by its failures compared with the other's successes, and not vice versa, is evidently unfair. The difficulty is compounded, for weighting experiments, if the assessed output is that of Boolean searches. As the objective of weighting is to order output, it is most appropriately compared only with other indexing and searching strategies designed to produce similar output; if unordered output is in question, some neutral method is required of fixing a cutoff to select the highest ranking documents as the output set. Since operational systems tend to produce retrieved sets meeting Boolean specifications, comparing unconventional relevance weight searching with their conventional searching clearly presents problems. MILLER[4] approached this problem by setting the cutoff for the weighted search to give approximately the number of documents retrieved in the earlier Boolean searches, but this technique clearly introduces bias. A similar technique, noted as not wholly satisfactory, was used by EVANS[10]. Abandoning the ordering given by weighting is in any case only plausible if it is so small that ordering within it is unimportant; applying a cutoff to obtain a set for comparison with a large Boolean output is throwing away an advantage of weighting.

These evaluation difficulties are liable to arise when any attempt is made to conduct large scale experiments relevant to operational systems, and in particular to exploit any available test data derived from such systems, since this is typically restricted in some "undesirable" way. It may be possible with such data to carry out some experiments of interest, for example to compare the performance of F1 and F4; but for the reasons just mentioned it may not be possible to relate performance for the relevance weighting techniques with which we are concerned to conventional performance for the systems from which the data was derived, in a systematic and reliable way.

This point has been laboured to ensure that our experimental results are not misinterpreted. The tests reported in this paper are primarily ones designed to provide more information about the value of relevance weighting for search contexts to which it is appropriate, and only secondarily, given the limitations of the available data, to allow some estimates to the extent to which relevance weighting could compete with conventional search techniques.

TEST DATA

The only conveniently available data meeting the requirements imposed by the test questions was that derived from the UKCIS project (BARKER *et al.*[5] and also [11]).† This consists of 27361 documents taken from Chemical Abstracts Condensates and 75 standard Boolean profiles intended for SDI use. The documents are represented by titles, while the profiles are of an exhaustive kind involving words or word strings, with front and/or back truncation. Relevance assessments are for the pooled output of UKCIS searches on different document fields and with different profile formulations, representing an average of 195.6 documents assessed per profile. For the relevance weighting tests the Boolean profile structure was abandoned, and in addition NOT terms were rejected. This gave simple term (fragment, word or string) lists for the profiles, averaging 18.3 terms.

It will be evident that the material is subject to the limitations discussed earlier in providing only limited relevance information, and those relating to Boolean search performance. It is, however, adequate for direct studies of relevance weighting, and specifically, meets the requirements of our initial questions. Thus the document set is large, if not enormous. (The request set is also a respectable size, meeting a general experimental requirement.) The

†Made available by courtesy of UKCIS.

document set size means that predictive experiments on an acceptable scale can be conducted, in an attempt to answer question 1. As the set is large, subsets of varying sizes and predictive merit can be extracted for tests to answer question 4. Further, the document set consists of two major blocks of items from CAC, representing two major subject area groupings of CA: this provides material for question 3. (The file size is in any case large enough for considerable topic variation.) Finally, the original user need statements are available for use in answering question 2: when processed "straight" they provide an average of 5.7 search terms per question, for comparison with the much fuller profiles. Obtaining test data without any limitations is outside the scope of the individual project: the UKCIS material meets the essential test requirements, and is also derived from an operational service. It is therefore adequate for the purpose of validating past experiments and of beginning to relate them to operational environments. Details of the test data are given in Table 1.

EXPERIMENTS

The general pattern of the tests was as follows. In each case two parts of the document set were used: the first, the weight generation subset, provided the information for computing the term relevance weights; the second, the weight application subset, was searched using the weights for the request terms derived from the documents in the first subset. In the artificial conditions of the experiment it was assumed that the entire weight generation subset was evaluated to provide relevance information: i.e. all the known relevant documents in the subset were deemed identified relevant. This treatment of the collection allows the following performance comparisons:

(a) between relevance weighted term searching and simple unweighted term searching for the application set;

(b) between relevance searching and collection frequency weighted searching for the application set;

(c) between relevance weighted term searching based on different formulae, i.e. F1 and F4; and

(d) between the predictive relevance weighted searching and the corresponding retrospective relevance weighted searching, the latter using the application set as generation set.

In a given test, therefore, the range of comparisons is over terms, collection frequency weights, predictive relevance weights using F1 and the same using F4, and retrospective relevance weights. As indicated earlier, performance is represented by recall/precision figures giving a conventional graph. As significance tests are not easily applied to such results, it is assumed, crudely, that an area difference of 5% is significant. In earlier papers a difference of more than 5% was labelled noticeable, and one of more than 10% material. We are dealing here with rather larger performance differences, and also with a wide range of performance comparisons, while we may attach more importance to a smaller percentage improvement at high precision than a larger one at low because the absolute performance improvement is greater. We can accordingly only characterise performance differences informally, and will use three degrees of difference, saying (a) is better than (b), or much better, or strikingly better, than (b), and defining these degrees ostensively by reference to the test recall/precision results. A convenient notation for the degrees of difference is '⟩', '⟩⟩' and '⟩⟩⟩', while the forms of searching may be named T (terms), C (collection frequency weights), R (relevance weights), hence R1 and R4 and Y (yardstick), hence Y1 and Y4, or Y1p and Y4p if the predictive formula is applied retrospectively. Our experimental objective may therefore be summarised as being to show that in four defined contexts, at least R4⟩R1⟩C⟩T.

The individual tests were as follows.

Experiment 1: *scale effects*

Following the model of the experiments reported in (1), the document set was divided into even and odd-numbered subsets. The Even set was used for generation and the Odd for application. Relevant data details are given in Table 1. The recall and precision results obtained for the different runs are given in Table 2.1 and illustrated in Fig. 1. They cover terms, T, collection weights, C, predictive relevance weights R1 and R4, yardstick "pure" retrospective relevance weights for both formulae, i.e. Y1 and Y4, and Y1p and Y4p. Comparison across the

Table 1. Data details.

No. profiles			75
Av. terms per profile			18.3
No. documents			27361
No. word stems in titles			17537
Av. stems per document			6.6
No. different documents relevant to some profile			3380
No. relevant postings			3739
Av. relevant per profile			49.9
Av. documents assessed per profile by UKCIS			195.6
Experiment 1			
No. documents		Even	13680
		Odd	13681
No. relevant postings		Even	1837
		Odd	1902
Av. relevant per profile		Even	24.5
		Odd	25.4
Experiment 2			
Av. word stems per request			5.7
Documents as Experiment 1			
Relevant " " "			
Experiment 3			
No. documents (CAC-1)		First	11613
(CAC-2)		Last	15748
No. relevant postings		First	1444
		Last	2295
Av. relevant per profile		First	19.3
		Last	30.6
Experiment 4			
No. documents		Sixteenth	1711
		Eighth	3421
		Quarter	6841
		Half	13681
	Search	Quarter	6840
No. relevant postings (estimated as percent whole collection for generation sets)		Sixteenth	234
		Eighth	467
		Quarter	935
		Half	1870
	Search	Quarter	902
Av. relevant per profile		Sixteenth	3.1
		Eighth	6.2
		Quarter	12.5
		Half	24.9
	Search	Quarter	12.0

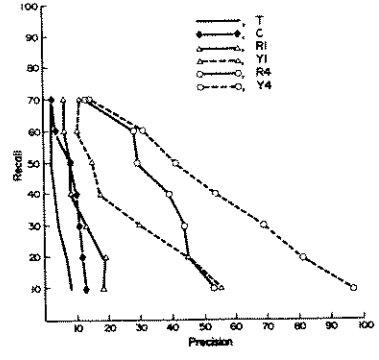


Fig. 1. Experiment 1.

Table 2. Retrieval performance.

		PRECISION							
2.1 Experiment 1	G	-	-	Even	Even	Odd	Odd	Odd	Odd
	A	Odd	Odd	Odd	Odd	Odd	Odd	Odd	Odd
		T	C	R1	R4	Y1	Y4	Y1p	Y4p
RECALL	100	0	0	0	0	0	0	0	0
	90	0	0	0	0	0	0	0	0
	80	0	0	0	0	0	0	0	0
	70	2	2	6	13	11	14	6	14
	60	2	3	6	28	10	31	6	28
	50	2	8	8	29	15	42	8	36
	40	3	10	8	39	17	54	9	42
	30	4	11	13	44	30	69	15	54
	20	7	12	19	45	45	81	22	60
	10	8	13	18	53	55	97	23	62
2.2 Experiment 2	G	-	-	Even	Even	Odd	Odd	Odd	Odd
	A	Odd	Odd	Odd	Odd	Odd	Odd	Odd	Odd
		T	C	R1	R4	Y1	Y4	Y1p	Y4p
RECALL	100	0	0	0	0	0	0	0	0
	90	0	0	0	0	0	0	0	0
	80	0	0	0	0	0	0	0	0
	70	0	0	0	0	0	0	0	0
	60	0	0	0	0	0	0	0	0
	50	0	0	0	0	0	0	0	0
	40	0	0	0	0	0	0	0	0
	30	1	2	2	4	5	8	2	6
	20	1	6	8	16	13	29	8	22
	10	5	14	14	30	38	58	19	46
2.3 Experiment 3	G	-	-	First	First	Last	Last	Last	Last
	A	Last	Last	Last	Last	Last	Last	Last	Last
		T	C	R1	R4	Y1	Y4	Y1p	Y4p
RECALL	100	0	0	0	0	0	0	0	0
	90	0	0	0	0	0	0	0	0
	80	0	0	0	0	0	0	0	0
	70	2	2	3	0	6	8	3	7
	60	2	4	3	16	12	43	7	41
	50	2	7	6	20	12	52	7	46
	40	3	10	6	23	11	56	6	51
	30	3	12	7	23	20	65	9	57
	20	6	13	8	36	42	79	18	62
	10	7	13	12	41	51	92	20	70
2.4 Experiment 4	G	-	-	Six.	Ei.	Quar.	Half	Se.Q.	Se.Q.	Ei.	Se.Q.	Se.Q.
	A	Se.Q.	Se.Q.	Se.Q.	Se.Q.	Se.Q.	Se.Q.	Se.Q.	Se.Q.	Se.Q.	Se.Q.	Se.Q.
		T	C	R4	R4	R4	R4	Y4	Y4p	R1	Y1	Y1p
RECALL	100	0	0	0	0	0	0	0	0	0	0	0
	90	0	0	0	0	0	0	0	0	0	0	0
	80	0	0	0	0	0	0	0	0	0	0	0
	70	2	3	0	8	12	13	24	18	4	12	6
	60	2	5	13	27	31	26	36	34	5	13	6
	50	2	7	17	32	33	33	54	43	6	17	7
	40	3	11	38	37	41	43	63	50	7	24	10
	30	4	12	45	47	46	45	74	51	10	35	15
	20	7	12	48	48	51	49	87	59	13	51	20
	10	8	15	53	46	51	52	98	62	13	61	22

Key

Indexing :

T = terms

C = collection frequency weights

R1 = predictive use of relevance weights, formula F1

R4 = " " " " " " " formula F4

Y1 = yardstick retrospective use of relevance weights, formula F1

Y4 = " " " " " " " formula F4

Y1p = yardstick using predictive form of formula F1

Y4p = " " " " " " " formula F4

Document sets :

G = generation set for relevance weights

A = application set for searching

strategies finds R4 strikingly superior to R1, the latter the same as C, with C much better than T, i.e. R4⟩⟩R1 = C⟩⟩T. We note also that Y4⟩⟩R4 and Y1⟩⟩R1, but interestingly that R4⟩⟩Y1, i.e. that predictive relevance weighting by formula F4 is much better than yardstick formula F1. Table 2.1 also shows that when the yardsticks are set by the predictive versions of the formulae they are inferior to the "pure" ones, and indeed that Y1p is much the same as R1.

Experiment 2: *poor matching*

For this test the initial need statement texts were processed to give request statements analogous to those represented by the profiles, but much briefer. As noted earlier, the profiles are long and carefully worked out, and so provide high-class matching material even when treated as term lists rather than Boolean formulations. The alternative brief requests were obtained by simple automatic processing: stop words were eliminated from the texts, and the remaining words replaced by word stems defined by a stem dictionary derived from the document title texts. Some of the characteristic fragmentation of the profiles is thus repeated, but less full profiles are obtained indirectly rather than directly: hand reduction of the original profiles to obtain a strictly comparable poorer version for matching proved too difficult. The content of the original statements and profiles is, however, fundamentally the same, so the use of the text-derived requests for this experiment seems acceptable. They are detailed in Table 1. Weighting and searching are based on the even and odd-numbered document sets, as in Experiment 1. The results are given in Table 2.2 and illustrated in Fig. 2. The strategy comparison here shows essentially the same pattern as Experiment 1, i.e. R4⟩⟩R1 = C⟩⟩T, though absolute levels of performance are much lower. The weighting improves precision noticeably, though the recall ceiling is low and overall performance is poor. Again the yardsticks define much better performance than prediction achieves; but in this case predictive weighting using formula 4 is the same as yardstick F1. The relationship between the forms of yardstick is as in Experiment 1.

Experiment 3: *heterogeneous documents*

Here the regular profiles were used but the documents were divided into First and Last sets representing the major division of CAC by subject groupings. In general the documents known to be relevant to a request fall predominantly into one set, but most profiles have some relevant documents in the other. Predictive weighting from one set to the other should therefore constitute a rather stronger test than the even/odd case. The experiment should properly be done both ways round, but the effort involved would be considerable and it was thought adequate to predict from the first set of documents 1-11613 (CAC-1) to the last documents 11614-27361 (CAC-2). Data details again appear in Table 1, and retrieval results in Table 2.3 and Fig. 3. Comparative performance in this experiment is similar to that of Experiment 1,

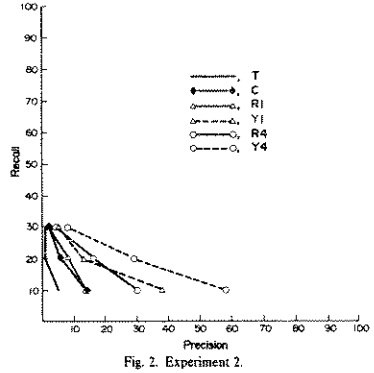


Fig. 2. Experiment 2.

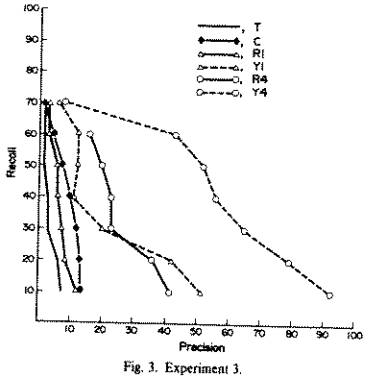


Fig. 3. Experiment 3.

specifically R4⟩⟩R1 = C⟩⟩T. Yardstick performance is again much better than predictive, while R4 is in this case the same as Y1. Absolutely, R4 performance in Experiment 1 is better than in Experiment 3.

Experiment 4: little information

For this experiment the generation document set was progressively reduced. The application set, called Search Quarter, contained every fourth document, 4,8,12,16, etc.† The largest generation set, called Half, contained 1,2,5,6,9,10, etc.; this was divided to obtain Quarter, containing 1,5,9, etc. Quarter was divided to form Eighth, containing 1,9, etc. and Eighth was divided to obtain Sixteenth, containing 1,17, etc. The implication of this reduction for the amount of relevance information input to weight computation is indicated in Table 1 by the average number of relevant documents per request (figures estimated from the whole collection). Thus we see that for Sixteenth only 3–4 relevant documents are available. Relevance weighting for this experiment was confined to formula F4 for all the sets, with F1 only for Eighth. Performance is shown in Table 2.4 and Figs. 4a and 4b, leading to the conclusion that the reduction in source has no effect on precision and only lowers the recall ceiling slightly. Fig. 4a shows R4 for Sixteenth and Half virtually the same. The overall picture otherwise is as for Experiment 1, i.e. R4⟩⟩C⟩⟩T, while comparisons for generating set Eighth including predictive weighting using formula F1 give the same result as Experiment 1, with R1 ≈ C. Predictive formula F4 is better than retrospective F1, while the "pure" yardsticks are much better than the predictive formula versions.

SUMMARY OF RESULTS

The results for Experiment 1 show the same relative picture, but for a much larger file, as the original tests of (1). They show relevance weighting clearly superior to simple term or collection weight matching; F4 strikingly better than F1, indeed predictive F4 comparable to retrospective F1; and predictive F4 achieving a remarkable improvement in performance. In Experiment 2 the same relative results are obtained, though the absolute level of performance is much lower. Since the average number of plain matching terms per relevant document is a mere 1.3, the weighting is doing a great deal of work. Performance for the relevance weights in Experiment 3 is inferior to that for Experiment 1, but still improves performance very considerably. The results for Experiment 4 are very interesting: performance does not decline as the size of the generation set is reduced, so the improvement over term matching even for Sixteenth is very marked: i.e. relevance information, even when very limited, is of great value.

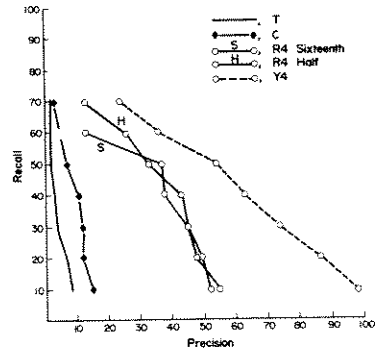


Fig. 4(a). Experiment 4.

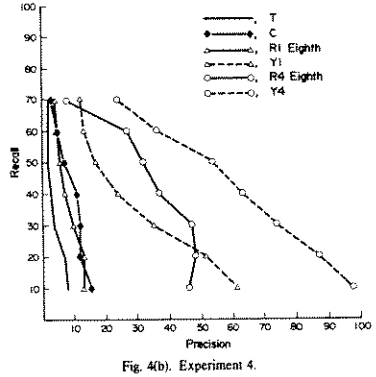


Fig. 4(b). Experiment 4.

DISCUSSION

(a) *Straightforward performance of relevance weights*

All the tests show that relevance weighting in a coordination matching environment is very effective, much more effective than previously investigated statistical indexing/searching techniques. Tests with other, though smaller collections, confirm the general picture: for example with 39 profiles and 2542 documents used for a variety of coordination and weighting tests by EVANS[10]. One question which may be asked about the testing procedure is whether results would be different if another evaluation technique was adopted. For example, what happens if ranking is explicit and averaging across requests is by matching ranks rather than matching scores? Ranking experiments are very expensive, present methodological problems, and are perhaps of doubtful operational relevance, so some trials only have been conducted. A test with the even and odd-numbered Cranfield documents used for the original experiments of (1) shows overall results similar to those reported above, but much smaller absolute performance differences (a 15–20% improvement in precision for predictive formula 4) and no difference between F1 and F4. Further work in this area is needed.

(b) *Operational implications of relevance weights*

(i) Comparison with conventional Boolean search performance is very difficult for the reasons discussed earlier. The output of the original UKCIS alternative Boolean searches was

all assessed, and recall relative to the pooled output of all the searches obtained. The UKCIS workers estimated absolute recall for Boolean searches on titles at about 40% as opposed to the relative value of about 60%[11]. The problem for comparing our coordination output with the Boolean is first, what proportion of the hypothetical total relevant documents be deemed retrieved in the much larger output of the coordination searches, and second, what distribution over the levels should be assumed for the extra relevant documents. Various plausible totals of "real" relevant retrieved can be suggested under the first head, but the second problem is much more intractable. For unweighted terms distributing the extra documents in parallel with the known ones is a possibility, but with relevance weighting negative weights make the situation too complicated. Only a very limited comparison between our test results and the original UKCIS operational ones can therefore be made. Specifically, we adjust the given Boolean search point to its estimated real value, i.e. reduce recall from 60% to 40%. A worst assumption adjustment of coordination searching would add extra documents at the lowest level only, raising the recall ceiling but not affecting recall/precision graph shape. The net result for the even/odd case would be very similar precision at 40% recall. Thus the original relative recall of 64% and precision of 44% become recall 40% and precision 44%, compared with 39% precision at recall 40% for predictive relevance weights with formula F4. Precision at recall 40% for F4 weights derived from the Sixteenth generation set in Experiment 4 is 38%.

Such games should be treated with some reserve: they are intended chiefly to suggest that a proper experimental comparison between unconventional weighting techniques and conventional searching would show the former competitive with the latter. In the meantime it cannot be stressed too strongly that the performance figures obtained in our experiments are not absolute, and are likely to be in some degree affected by the restricted relevance information available for evaluation. It should also be noted that the Boolean profiles used by UKCIS were finely honed on the basis of experience over searches, and therefore present an unusually high standard for comparison.

The performance of relevance weights must be influenced by the number and quality of the given search terms. The experiments reported were to some extent unrealistic since it is not surprising that relatively good performance can be obtained with a long and carefully formulated term list. From this point of view the substitution of weights for Boolean structure does not represent much effort saving. On the other hand, we must reckon on many searches with far less carefully developed input. If coordination performance here is not very good, neither is Boolean (see, for example, Aitchison *et al.*[12]). The important point here, as Experiment 2 shows, is that relevance weights improve poor coordination performance considerably, though how competitively with poor Boolean searching our data cannot show.† Miller's relatively limited experiments[4, 13] show even weighting by F1 competitive, so F4 would presumably be superior to Boolean. It is hoped that tests can be carried out to compare relevance weights with simple Boolean searches, or subsearches, preferably in an on-line environment. This would allow better estimates of the value of relevance weights, balancing their contribution to performance against the effort required to generate them, in relation to the relative contribution to conventional Boolean performance of term choice and request structure. As some effort is involved in choosing terms, in providing structure, and in deriving weights, we need to know how they are most efficiently combined for best performance with least effort.

(ii) The tests reported assume that all the documents in the generation set are scanned for relevance. This is clearly unrealistic. Even Sixteenth generation set in Experiment 4 represents some 1711 documents. The question is how much would be lost, in input information for relevance weight computation, if data for far fewer documents was available. Or rather, since the crux is relevance information, how few documents scanned provide enough relevance information. As the experiments in input reduction show, 3–4 relevant documents provide valuable information. It appears, using Search Quarter of Experiment 4 as an example, that simple term matching retrieves 2.7 relevant documents for 39.3 scanned (at matching level 2), while collection weighting retrieves 3.0 relevant for 24.2 scanned (at level 9) or 4.6 for 37.7 scanned (at level 8). We may thus hope to obtain enough relevance information from a far smaller number of documents inspected than the whole set, in this case 6840. In itself, the

proportion of relevant thus obtained is small (though bearing in mind the data properties, it is probably misleadingly small) but as the means to a future large return it looks quite different. At any rate, such figures show that obtained productive relevance information would not be the effort the experimental design assumed.

This point is clearly relevant to current on-line search procedures where relatively few documents may be inspected to provide relevance information. Tests are currently under way to evaluate relevance weighting in a simulation of such an environment.

CONCLUSION

As noted earlier, though the tests reported here are confined to one collection, similar performance patterns have been obtained with other data. The questions asked initially have been satisfactorily answered for the test environment. The results show quite clearly that relevance weighting, expecially formula F4, is very useful, leading to really substantial performance improvements over simple term matching. The levels of performance attained are higher than those achieved by other statistical techniques applied to term coordination matching; and they are comparable with those achieved by conventional Boolean searches, and perhaps can be obtained with less effort.

REFERENCES

[1] S. E. Robertson and K. Sparck Jones, Relevance weighting of search terms. *J. ASIS* 1976, **27**, 129–146.

[2] C. T. Yu and G. Salton, Precision weighting—an effective automatic indexing method. *J. ACM* 1976, **23**, 76–88.

[3] S. E. Robertson, *A Theoretical Model of the Retrieval Characteristics of Information Retrieval Systems*, Ph.D. Thesis, University of London, 1976.

[4] W. L. Miller, A probabilistic search strategy for MEDLARS. *J. Docum.* 1971, **27**, 254–266.

[5] F. H. Barker, D. C. Veal and B. K. Wyatt, Towards automatic profile construction. *J. Docum.* 1972, **28**, 44–55.

[6] A. Robson and J. S. Longman, Automatic aids to profile construction. *J ASIS* 1976, **27**, 213–223.

[7] K. Sparck Jones, A performance yardstick for test collections. *J. Docum.* 1975, **31**, 266–272.

[8] C. W. Cleverdon, J. Mills and M. Keen, *Factors Determining the Performance of Indexing Systems*, 2 vols., College of Aeronautics, Cranfield, 1966.

[9] L. Sparck Jones and R. G. Bates, *Research on Automatic Indexing* 1974–76, 2 vols., Computer Laboratory, University of Cambridge, 1977.

[10] L. Evans, *Search Strategy Variations in SDI profiles*, Report No. 75/21, Institute of Electrical Engineers, London, 1975.

[11] F. H. Barker, D. C. Veal and B. K. Wyatt, *Retrieval Experiments Based on Chemical Abstracts Condensates*, Research Report No. 2, UKCIS, Nottingham, 1974.

[12] T. M. Aitchison, A. M. Hall, K. H. Lavelle and J. M. Tracy, *Comparative Evaluation of Indexing Languages, Part II: Results.* Project Inspec, Institute of Electrical Engineers, London, 1970.

[13] W. L. Miller, *The Evaluation of Large Information Retrieval Systems with Application to MEDLARS*, Ph.D. Thesis, University of Newcastle, 1970.

Reprinted from AMERICAN DOCUMENTATION
January 1969, Vol. 20, No. 1

Effectiveness of Information Retrieval Methods*

Results of some 50 different retrieval methods applied in three experimental retrieval systems were subjected to the analysis suggested by statistical decision theory. The analysis validates a previously-proposed measure of effectiveness and demonstrates its several desirable properties. The examination of a wide range of data in relation to this one metric provides a clear and general assessment of the current state of the retrieval art.

JOHN A. SWETS

Bolt Beranek and Newman Inc.
Cambridge, Massachusetts

A desirable measure of retrieval performance would have the following properties. First, it would express solely the ability of a retrieval system to distinguish between wanted and unwanted items—that is, it would be a measure of "effectiveness" only, leaving for separate consideration factors related to cost or "efficiency." Second, the desired measure would not be confounded by the relative willingness of the system to emit items—it would express discrimination power independent of any "acceptance criterion" employed, whether the criterion is characteristic of the system or adjusted by the user. Third, the measure would be a single number—in preference, for example, to a pair of numbers which may covary in a loosely specified way, or a curve representing a table of several pairs of numbers—so that it could be transmitted simply and apprehended immediately. Fourth, and finally, the measure would allow complete ordering of different performances, indicate the amount of difference separating any two performances, and assess the performance of any one system in absolute terms—that is, the metric would be a scale with a unit, a true zero, and a maximum value. Given a measure with these properties, we could be confident of having a pure and valid index of how well a retrieval system (or method) was performing the function it was primarily designed to accomplish, and we could reasonably ask questions of the form, "Shall we pay X dollars for Y units of effectiveness?"

* This work was supported by the Advanced Research Projects Agency under ARPA Order No. 627, Amendment 2, and under Contract No. AF19(628)-5065 with the Air Force Cambridge Research Laboratories.

In a previous article I reviewed 10 measures that had been suggested prior to 1963, and proposed another (*1*). None of the 10 measures, and none that has come to my attention since then, has more than two of the properties just listed. Some of them, including those most widely used, have the first two properties, and some of the others have the last two properties. The measure I proposed, one drawn from statistical decision theory, has the potential to satisfy all four desiderata. At the time it was proposed, however, the decision-theory measure had not been applied to any empirical retrieval results, so that its assumptions about the form of retrieval data had not been tested. In the present paper we examine this measure in relation to test results obtained from three experimental retrieval systems with some 50 different retrieval methods. With minor qualifications, the data are uniformly consistent with the assumptions of the decision-theory measure and quite clearly demonstrate its usefulness. A substantive outcome of the extensive analysis in terms of this measure is a clear appraisal of the current state of the retrieval art. The analysis shows in precise terms how much room for improvement is left by current retrieval techniques. The room for improvement, as we shall see, is large.

Before proceeding to a review of the decision-theory measure and to an examination of the data, let us consider briefly the domain of the measure and a disclaimer about the scope of this paper.

The measure is most clearly applicable to retrieval systems that deal in documents or messages, and it is applied here to systems of this type. Less clearly perhaps, but

as well, the measure can be applied to information systems that handle facts or give answers to ordinary English questions. In both cases queries are addressed to a system and the system's responses to the queries must be evaluated. Whether the response is a set of documents or a fact selected or deduced from a collection of writings is immaterial. Appropriate text must be isolated in either case, to constitute the response or to supply the base from which the response is drawn. The data represented by the decision-theory measure are entries in a two-by-two contingency table: Just as documents suited or unsuited to a need may be retrieved or not retrieved, so facts that correctly or incorrectly answer questions may be presented or withheld. For some relatively simple fact systems, of course, such as airline-reservation systems, discrimination or correctness is not a problem; the reference here is to fact systems in which the facts to be retrieved are not all neatly isolated and in which the questions are not all anticipated in detail.

This measure, like those used most often in the past, is most directly applicable when the entire information store is known and when, in particular, the number of items appropriate as responses to each query is known. This condition is frequently satisfied in experimental systems, which usually contain no more than a few thousand items. If the measure is to be applied to stores large enough to make impractical a complete knowledge of them, three alternatives exist for estimating the required number. One is to select, by some heuristic process or by fiat, that subset of the full store likely to contain almost all of the items appropriate to a given set of queries, and to examine the subset in detail. A second alternative, used in one instance in the following, is simply to sample the large store and to extrapolate from the sample. A third alternative, used in another instance in the following, is to preselect certain items from the store and to design test queries specifically to retrieve those items.

Application of the decision-theory measure assumes that the "relevance" of any item in the store to a given query or user's need can be determined. As the reader will know, or can imagine, the definition of relevance is generally regarded in the retrieval field as a very thorny problem, and even the concept itself has at times come under attack. However that may be, the definition of relevance is an issue separate from the measure under consideration, and is not discussed here. Nor is the concept defended here; I take it for granted that it is essential to the evaluation of retrieval performance and that sooner or later we shall come to terms with it. For our present purposes, we can accept the definitions of relevance adopted by the investigators who collected the data we shall examine, just as we accept for the present purposes other experimental procedures they have followed. It will become clear, by the way, that the decision-theory measure can be applied when judges use several, rather than two, categories of relevance, and that it uses to full advantage the output of a system that ranks or otherwise scales all items in the store according to their degree of relevance to the query at hand.

• Decision-Theory Measure

A good way to begin in reviewing the decision-theory measure is to consider a measure more familiar in the retrieval context and to note the differences between the two. The measure used far more than any other (*8*) consists of two quantities termed the "recall ratio" and the "precision ratio." Like other measures that attempt to assess only retrieval effectiveness, this measure can be described by reference to the relevance-retrieval contingency table shown in Fig. 1.

The recall ratio is defined as $a/a+c$, the number of items both relevant and retrieved divided by the number of items relevant. This ratio, then, is the proportion of relevant items retrieved, and it may be taken as an estimate of the conditional probability that an item will be retrieved given that it is relevant. The precision ratio (formerly called the "relevance ratio") is defined as $a/a+b$, the number of items both relevant and retrieved divided by the number of items retrieved. This ratio is the proportion of retrieved items deemed relevant, and an estimate of the conditional probability that an item will be relevant given that it is retrieved.

Now, if a system's effectiveness is characterized by two numbers, a value of the recall ratio and a value of the precision ratio, we know relatively little about the system, for one reason because we don't know how the two quantities relate to each other. What does it mean, for example, to say that a system yielded a recall ratio of 0.70 and a precision ratio of 0.50? If System A performs this way, and System B yields a recall ratio of 0.90 and a precision ratio of 0.40, is System B more or less discriminating than System A? That is, is a gain of 0.20 in recall and a loss of 0.10 in precision good or bad? Of course, should System B show a gain in both recall and

	r	$\bar{r}$	
R	a	b	$a+b$
$\bar{R}$	c	d	$c+d$
*	$a+c$	$b+d$	$a+b+c+d$

FIG. 1. The relevance-retrieval contingency table: r and $\bar{r}$ denote, respectively, relevant and irrelevant items; R and $\bar{R}$ denote, respectively, retrieved and unretrieved items; a, b, c, and d represent frequencies of occurrence of the four conjunctions.

precision over System A, we know B's effectiveness is superior to A's, but, in general, the measure consisting of this pair of quantities will give only a partial ordering of different systems or of different methods employed by one system.

The problem here is that System A's recall of 0.70 and precision of 0.50 represents only one of the many balances between the two ratios that it can achieve. This balance might have occurred when an item had to satisfy five descriptors specified in a query in order to be retrieved. If this requirement is changed, so that now an item has only to satisfy any two of the query's five descriptors, it is very likely that more items will be retrieved, and that recall will go up and precision will go down. But we must know exactly how recall and precision will covary, along with variation in the acceptance criterion, if uncertainties are to be avoided in attempting to rank different systems or methods.

A solution to this problem, one that is sometimes adopted, is to test each system with several acceptance criteria and to present as the measure of a system's effectiveness the empirical curve so generated. Extensive tests have shown (*3*) that the empirical curve will resemble in form the curve shown in Fig. 2. If System A yields the curve shown while System B yields another curve everywhere above and to the right of the one shown, it is clear that B is superior to A.

However, these curves do not tell us, in general terms, by how many units B is superior to A. (We can determine that B's precision is greater than A's by some specific percentage at some specific value of recall, but this number varies widely as a function of the value of recall selected.) Nor can we tell from the curves how good either system is in absolute terms. And, of course, it is

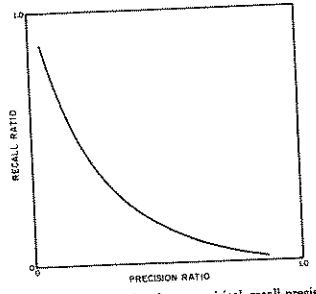


FIG. 2. Idealized example of an empirical recall-precision curve, fanned out by varying the acceptance criterion. For lenient criteria, recall is high and precision is low. Progressively more stringent acceptance criteria increase precision at the expense of recall.

relatively awkward (We might say that a large "bandwidth" is required) to transmit and receive a full curve.

A measure that retains the basic information inherent in the recall-precision curve, and at the same time overcomes the drawbacks of using a curve as a measure, would be attained if there is a way to represent completely an empirical curve of this general sort by a single number on a scale with a unit, a true zero, and a maximum. The thrust of my earlier article was that statistical decision theory offers a way—indeed, several ways. Whether or not we can take advantage of one of them, or to what extent, depends upon the form of retrieval data when analyzed by decision-theory techniques, and that form is the concern of this paper.

Though a way might be found to completely characterize any empirical recall-precision curve by a single number on the type of scale desired, decision theory suggests using the curve that results when another variable is substituted for precision. The variable to be substituted, in the terms of Fig. 1, is $b/b+d$. This quantity is the number of items both irrelevant and retrieved divided by the number of items irrelevant, or the proportion of irrelevant items retrieved, and is an estimate of the conditional probability that an item will be retrieved given that it is irrelevant.

As in my earlier article (*1*), I refer to the retrieval of an irrelevant item as a "false drop." Also for consistency, the retrieval of a relevant item is termed a "hit," so instead of the term "recall ratio" I use "the conditional probability of a hit." Some of the notation used here differs from that of the previous article. Here, as seen in Fig. 1, lower-case letters, r and $\bar{r}$, designate relevant and irrelevant items, while upper-case letters, R and $\bar{R}$, designate retrieved and unretrieved items. The two conditional probabilities of principal interest are here denoted $P(R|r)$ and $P(R|\bar{r})$. In the present notation, the curve we shall consider has $a/a+c$ or $P(R|r)$ on the ordinate and $b/b+d$ or $P(R|\bar{r})$ on the abscissa. This curve is a comparison of two "operating characteristics," as the term is used in statistics, and is called here the "relative operating characteristic," or *"ROC."*

One consideration in choosing the two variables used in decision theory, which are derived from the two columns of the relevance-retrieval contingency table, is that they contain all the information in the table; the remaining quantities of the table ("misses" and "correct rejections") are, respectively, their complements. The recall and precision ratios are derived from a column and a row of the table and do not serve to specify the remainder of the table.

A related, but more salient, consideration is that using the two variables of decision theory permits us to draw upon several models of the retrieval process which stipulate different forms that empirical *ROC* curves may take. That is, each of several available models developed within decision theory precisely specifies a given form for a theoretical *ROC* curve. Or rather, each model specifies

a family of *ROC* curves having an index of effectiveness as the parameter. Conveniently, the *ROC* curves of all but one of the models devised to date are straight lines or very nearly straight lines when plotted on linear normal-deviate, or "probability," scales. A single number is adequate as an index of effectiveness, because it is sufficient to generate the entire curve, under those models that assume some fixed relationship between the degree of effectiveness and the slope of the curve. Generality is gained at the cost of a second parameter in one model that permits a variable relationship between effectiveness and slope. Still another model gives a one-parameter fit to data without regard to the slope, or, for that matter, without regard to the general form of the *ROC* curve, but this number is not sufficient to regenerate the curve from which it is taken. We turn now to a description of these alternative models, and then to the retrieval data that will enable us to choose from among them the one or ones that will be useful.

The General Decision Model

Though the assumption is not essential to their application, I shall assume in describing the alternative decision-theory models that for each query submitted to a system, the system in some manner assigns an index value (call it z) to each item in the store to represent the degree of relevance of the item to the query. Plotting separately for irrelevant and relevant items the probability of assignment of each value of z yields two probability density functions. One form the two density functions might have is depicted in Fig. 3. The left-hand function is associated with irrelevant items, $f(z|\bar{r})$, and the right-hand function is associated with relevant items, $f(z|r)$.

If, as suggested in the figure, any given value of z might be assigned by the system to an item that is relevant or to an item that is irrelevant (as judged by a user or other umpire), then, as shown, some criterion value of z, denoted z_c, should be adopted, such that items assigned values greater than z_c are retrieved while items assigned values less than z_c are not retrieved. The areas under the two density functions to the right of z_c represent the probabilities of retrieving irrelevant and relevant items. They are the coordinates of the *ROC* curve, $P(R|r)$ and $P(R|\bar{r})$.

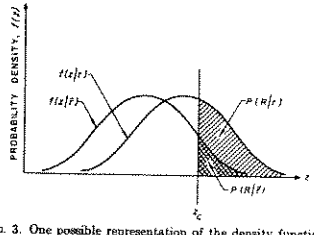


Fig. 3. One possible representation of the density functions for relevant and irrelevant items. The abscissa is the index of relevance, z, assigned by the system to each item. An acceptance criterion is labeled z_c.

Any given separation between the two density functions represents a stable retrieval system, with some particular capacity to distinguish between relevant and irrelevant items or some particular degree of effectiveness. For a fixed separation between the density functions, variation in the acceptance criterion z_c will result in a particular *ROC* curve. Another system or method, with greater or lesser ability to discriminate relevant and irrelevant items, will yield a different *ROC* curve as the acceptance criterion is varied.

The exact form of an *ROC* curve, it is clear, depends upon the shapes of the density functions that underlie it. Various measurement models are generated by hypothesizing density functions of different shapes.

Gaussian, Equal-variance Model

The density functions shown in Fig. 3 are Gaussian and of equal variance. Given the separation shown, variation in the acceptance criterion will trace the *ROC* curve labeled $E=1$ in Fig. 4. The measure E is defined as the difference between the means of the two density functions divided by their common standard deviation. If the separation is increased so that the difference between the means is twice as great as that shown in Fig. 3, then criterion variation will produce the *ROC* curve labeled $E=2$ in Fig. 4.

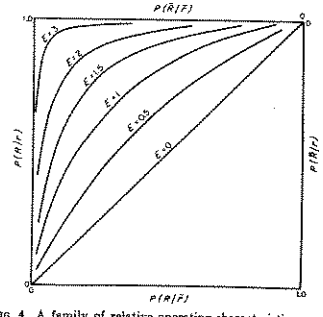


Fig. 4. A family of relative-operating-characteristic curves, based on Gaussian density functions of equal variance, with values of the parameter E. Labels on the upper and right-hand scales indicate that the full relevance-retrieval contingency table can be recovered from the plot.

We see that empirical data obtained from a test of a retrieval system could be plotted in the space of Fig. 4. If the data points followed the contour of one of the curves shown, or one of the intermediate curves not shown, the label on that curve would completely describe the effectiveness of the system: Knowing the single number permits reconstruction of the entire curve.

It is more convenient to plot data fitted by the *ROC* curves of Fig. 4 on probability scales, that is, on axes scaled linearly for the normal deviate, for then these *ROC* curves are straight lines with unit slope, as shown in Fig. 5. The measure E for any curve can be read from the normal-deviate scales; one simply subtracts the value on the right-hand scale from the value on the top scale corresponding to any point on the curve. In Fig. 5, E is also scaled along the negative diagonal.

It can be seen that for practical purposes E has a maximum of approximately 5.0; though the axes could be extended to show higher values of E, effectiveness is not really at issue for retrieval systems yielding a hit probability greater than 0.99 and, simultaneously, a false-drop probability less than 0.01. There is the additional fact that reliable estimation of such extreme probabilities demands a sample of excessive size.

Gaussian, Unequal-variance Model

If the density functions are Gaussian, but of unequal variance, the *ROC* curves on the scales of Fig. 5 will be linear with slopes other than unity. In particular, the slope of the *ROC* curve is equal to the ratio of the standard deviation of $f(z|\bar{r})$ to the standard deviation of $f(z|r)$.

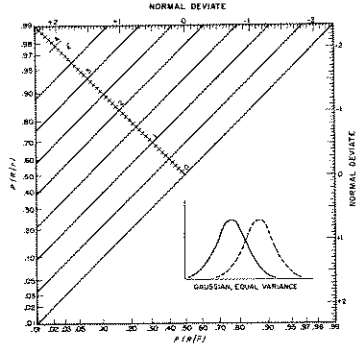


Fig. 5. The relative-operating-characteristic curves of Fig. 4 on probability scales, that is, on axes scaled linearly for the normal deviate. Density functions inserted at lower right identify the basis of these *ROC* curves in Gaussian, equal-variance density functions.

For density functions of unequal variance, E must be redefined, for it was previously defined in terms of a standard deviation common to the two functions. Note that for *ROC* curves of non-unit slope, the value of E obtained by subtracting a normal-deviate value on the right scale from one on the top scale is not constant along the curve. The definition adopted here consists in normalizing the difference between the means of the two density functions by their average standard deviation; this definition is reflected by measuring E at the intersection of the *ROC* curve and the negative diagonal of the *ROC* space.

Now, at least two alternatives are open to us. If we find that the slopes of empirical *ROC* curves vary without regard to E (measured at the intercept of the negative diagonal), two parameters will be needed to fit the curve. Reconstruction of the curve will require reporting the value of the slope, s, in addition to the value of E. It could turn out, on the other hand, that s bears some fixed relation to E, for example, that s increases regularly as E increases. This would be the case if the ratio of the increment in the mean of $f(z|r)$ to a decrement in its standard deviation were a constant. If this constant were a stable property of a given retrieval system, it could be reported once, and then the single value of E would be sufficient to describe the various curves the system produces as a result of changes in one or another independent variable.

Exponential Model

Simply as an illustration of further modelling possibilities, consider hypothesizing that the density functions are exponential in form, as shown at the lower right in Fig. 6. Then, again, the *ROC* curve is essentially linear on probability scales and can be described by a single parameter. The parameter $K=\sqrt{k}$ is defined in the figure; for $k>1.0$, the *ROC* curves have the property that s decreases regularly as the effectiveness (K) increases.

Distribution-free Model

If, after looking at data, hypothesizing some particular form of the density functions, and hence of the *ROC* curve, seems too strong a procedure, we can resort to a measurement scheme that leaves these forms unspecified and free to vary. We can take as the measure of effectiveness the percentage of the area of the *ROC* space that falls beneath any empirical *ROC* curve, when plotted on linear scales (as in Fig. 4). This measure, call it A, will vary from 50% for a curve that follows the positive diagonal, representing equal hit and false-drop proportions or no discrimination, to 100% for a curve that follows the extreme left and top coordinates of the graph, representing a hit proportion of 1.0 at a false-drop proportion of 0.0 or perfect discrimination. The measure

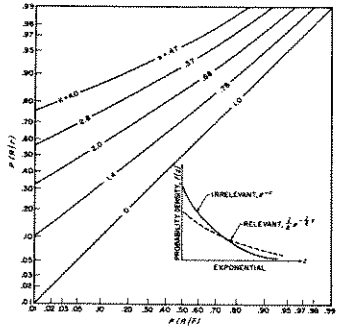


FIG. 6. A family of *ROC* curves based on exponential density functions, plotted on probability scales.

A, though a simple summary measure of effectiveness, does not permit reconstruction of the empirical curve from which it is drawn. It has the property useful for conceptual purposes that the value of A is equal to the percentage of correct choices a system will make when attempting to select from a pair of items, one drawn at random from the irrelevant set and one drawn at random from the relevant set, the item that is relevant. As demonstrated elsewhere (*4*) this equality holds for *ROC* curves of any form.

• Data

The three sets of data we shall examine were collected, respectively, at the Computation Laboratory of Harvard University by Gerard Salton (now at Cornell University) and Michael Lesk; under the Aslib project at Cranfield, England, by Cyril Cleverdon and Michael Keen; and at Arthur D. Little, Inc., by Vincent E. Giuliano and Paul E. Jones. These data were originally presented in technical reports published in late 1966 (*3, 5, 6*).

Salton and Lesk and Giuliano and Jones kindly made their raw data available to me so that I could calculate the hit and false-drop proportions. Cleverdon and Keen presented these quantities in their report. Though they are not responsible for the outcome, one or more of the authors of each report discussed with me the problem of measurement and commented on a draft of this paper. Their cooperation was essential, and I am pleased to acknowledge their very helpful advice and criticism.

Plots of data following are identified by the various terms for independent variables used in the original reports, to make possible cross references, but the terms are not defined here. Similarly, our present purposes do not require a description of the procedures of the three sets of experiments. However, a brief characterization of the scopes of the studies will be helpful in evaluating the general conclusions drawn here.

At Harvard, the questions asked experimentally include these: "Can automatic text processing methods be used effectively to replace a manual content analysis; if so, what parts of the documents [titles, abstracts, full text] are most appropriate for incorporation into the analysis? Is it necessary to provide vocabulary normalization methods to eliminate linguistic ambiguities; should such normalization be handled by means of specially constructed dictionaries, or is it possible to replace thesauruses by statistical word association methods? What dictionaries can be used most effectively for vocabulary normalization? Is it important to provide hierarchical subject arrangements, as is done in library classification systems; alternatively, should syntactical relations between subject identifiers be preserved? Does the user have an important role to fulfill in controlling the search procedure?" (*5*, pp. I-3, I-4). The experimental retrieval system, which operated on an IBM 7094 computer, was fully automatic in most applications; content-analysis procedures incorporated into the system processed documents and queries in natural language with no prior manual analysis. Stores of items used consisted of four collections of documents in three subject fields: documentation, aerodynamics, and computer sciences.

Experiments at Cranfield were based on manual analysis of documents. They were conducted to examine several different index languages (some languages using single terms, others based on concepts, and others based on a thesaurus); the exhaustivity of indexing; the level of specificity of index terms; a gradation of relevance assessments; and the amount of intelligence applied in formulating search rules. Two collections used consisted of documents in aerodynamics and aircraft structures.

The experiments at Arthur D. Little, Inc., evaluated manual and automatic indexing; length of the query; coordinate retrieval methods; and retrieval methods based on statistical word associations, with and without human intervention. The system operated on an IBM 1401 computer, with fully automatic indexing in most applications. All items in the file were abstracts of reports in the aerospace field.

Data from the three sources lead to the same conclusions about the usefulness of a decision-theory measure, so the analyses of the three sets of data will be presented with little evaluative comment prior to a general discussion of results. Each of the *ROC* plots is made on probability scales. Most of the plots summarize the results of one method of retrieval used with a given system; a few of them summarize the results of a single query used with a given method. The first question we ask is whether or not the plots of data are adequately fitted by straight lines. If they are, then we are interested in the slopes of the lines.

Harvard-Cornell Data

All of the data I obtained from the Harvard-Cornell project are presented here; this set includes almost all of the data collected under the project before June of 1966, the major exception being some collected toward the end of that time in tests permitting iterative searches under the user's control.

The system at Harvard, called "SMART," assesses the relevance of each item in the store to each query addressed to the system. Print-outs of data containing the relevance index for each item are, of course, extensive, and are not usually obtained; therefore we can not examine directly the shapes of the density functions. The standard print-out lists, for each query, the code number of every item relevant to it and the rank value of each of these items in a list ordered (by the system) according to degree of relevance. Data in this form permit adopting, for purposes of analysis, each of several arbitrary acceptance criteria according to the total number of items considered as retrieved. That is, $P(R|r)$ and $P(R|\bar{r})$ are calculated in turn, for example, for the 5 items ranked highest, the 10 items ranked highest, the 15 items ranked highest, and so forth, terminating at an arbitrary point.

To gain a relatively stable sample, results are combined for all queries used with a single method. One can pool results before calculating $P(R|r)$ *and* $P(R|\bar{r})$, or alternatively, can calculate these quantities for each query and take their average. The first of these procedures was followed in the analyses reported here.

Figure 7 shows the results for the collection of items in the subject field of documentation (called the ADI collection), under each of six retrieval methods. As in subsequent figures, in order to conserve space, only a portion of the *ROC* space is shown for each plot; the last panel in the figure reproduces the lines of the previous panels on the full *ROC* space. These lines, in all cases, were fitted to the data by eye.

The data are quite adequately fitted by straight lines in every instance. Indeed, according to standards acquired through experience in other fields (for example, human signal detection and recognition memory) where the decision-theory measure has proved to be useful (*4*), the fits are very good.

A small staircase effect can be discerned in the data. This effect may be the result of having a relatively small sample (containing an average of five relevant items for 35 questions); the procedure used in analysis for defining acceptance criteria forces each successive point a certain distance to the right, and a low density of relevant items would produce irregular upward movement. In any case, the effect is not large enough to be of much concern. We can see also some variation in the slopes of the lines; we shall consider the significance of this variation after all the data have been examined.

Figure 8 shows the results of seven retrieval methods applied to a collection of items on aerodynamics borrowed by the Harvard-Cornell group from the Cranfield project. Again, the straight-line fits exceed reasonable aspirations, and a variation in slopes appears.

Figure 9 represents one of two collections in the subject area of computer science, called IRE 1, and six retrieval methods. Figure 10 shows the second IRE collection and 10 methods. Figure 11 shows the second IRE collection with a different set of 10 methods.

With the IRE collection we notice a tendency, at higher values of E, for the slopes to be greater than unity. The slopes in Fig. 9 range from 0.95 to 1.12, in Fig. 10 from 0.98 to 1.40, and in Fig. 11 from 1.20 to 1.56. With the ADI collection (Fig. 7) the slopes range from 0.83 to 0.99, and with the Cranfield collection (Fig. 8), from 0.76 to 1.00.

We can't help but observe the substantive result of this analysis that the differences in effectiveness among the various methods are small relative to the differences among collections. The range in E for the six methods applied to the ADI collection is 0.20 (from 0.90 to 1.10); for the seven methods used with the Cranfield collection, 0.35 (from 1.45 to 1.80); for the six methods used with the IRE 1 collection, 0.40 (from 2.00 to 2.40); for the first 10 methods used with the IRE 2 collection, 0.55 (from 1.95 to 2.50); and for the second group of 10 methods used with the IRE 2 collection, 0.30 (from 2.10 to 2.40). These ranges, on the order of 0.50 or less, can be compared with the range over all collections of 1.60, keeping in mind the scale range of about 5.00 from chance performance to very good performance. The Harvard-Cornell and Cranfield investigators are inclined to believe that the dependency of effectiveness on the collection results both from differences in the "hardness" of the vocabularies of the three subject fields and from the use of different procedures with the three collections for establishing relevance (*7*).

Cranfield Data

The study at Cranfield has been actively pursued for several years, and the last report contains an enormous amount of data. I have plotted only a fraction of the results; however, I am not aware of any particular bias in my casual sampling, and all the plots prepared are included here.

The Cranfield data are distinguished from the Harvard data in being based on a larger file (in most cases 1,400 items, as compared with the largest Harvard collection of about 400 items), and on more questions (approximately 220, as compared with the Harvard maximum of about 40). One consequence is the appearance of lower false-drop proportions, proportions that fall off the graph paper (Codex Graph Sheet No. 41,453) used in the preceding figures. So we use another graph paper (Keuffel and Esser Co. No. 47 8062) that ranges down to a proportion of 0.0001. Though the graphs have on them scales of the normal deviate, these scales, unfortunately, are not given on the Keuffel and Esser paper available commercially.

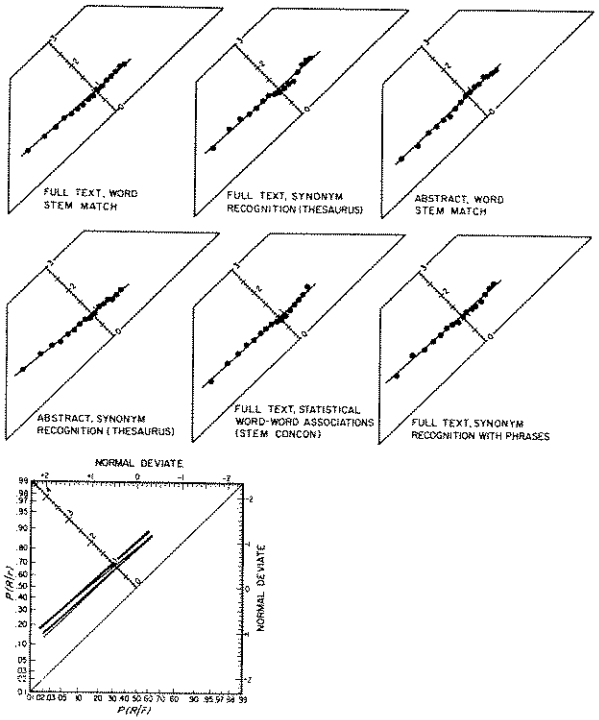


FIG. 7. ADI collection: 6 methods. 82 items, 35 queries. 170 relevant + 2,700 irrelevant = 2,870 total. Criteria: 3, 6, 9, . . ., 47 retrievals. Harvard-Cornell: Salton and Lesk.

In the Cranfield system, a manual one, the relevance of every item to every query is determined by judges, but the system itself does not rank items according to their degree of relevance to the query. Various acceptance criteria are obtained by establishing different "levels of coordination," that is, by varying the requirements on the number of query terms an item must satisfy in order to be retrieved.

Figure 12 shows the results of five retrieval methods that vary in the "recall device" they employ. The slopes are quite uniform, slightly greater than unity, and not many of the points fall off the fitted lines. Essentially the same comments apply to Fig. 13, which shows two levels of indexing exhaustivity for two sets of recall devices. Likewise for Fig. 14, which illustrates the effects of requiring different degrees of relevance for retrieval to be effected. The left panel results when all four categories of judged relevance satisfy the retrieval criterion; moving to the right, the relevance requirement is strengthened, so that in the last panel we have the results when only those items with the highest degree of relevance are retrieved. Figure 15 shows some results obtained with a smaller collection when retrieval is based on only titles

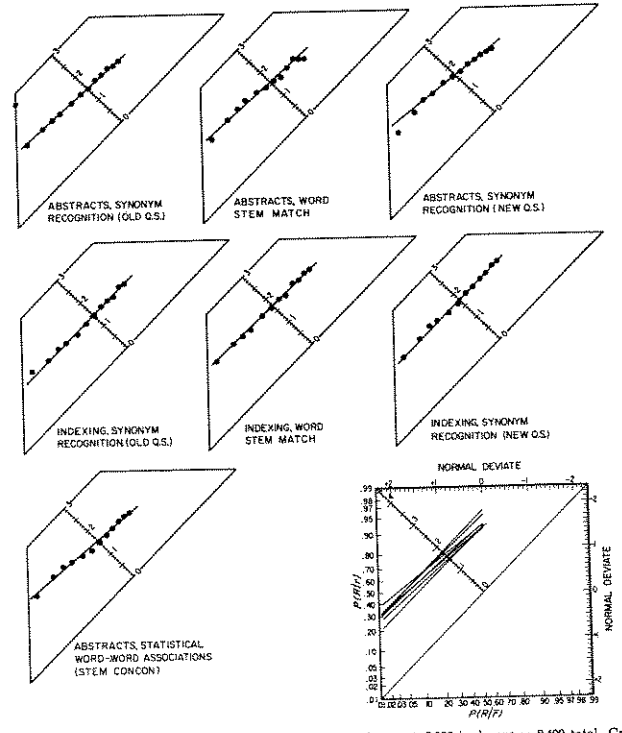


FIG. 8. Cranfield collection: 7 methods. 200 items, 42 queries. 198 relevant + 8,202 irrelevant = 8,400 total. Criteria: 5, 10, 15, 20, 30, 40, . . ., 100 retrievals. Harvard-Cornell: Salton and Lesk.

and abstracts, or on titles only, and the fits are about as good as before.

In Fig. 15 values of E range from 1.33 to 1.70, and values of the slope range from 0.80 to 0.95. In the three figures preceding, E ranges from 1.58 to 1.86, and s lies between 1.08 and 1.18.

ARTHUR D. LITTLE, INC., DATA

Like the Harvard system, the system constructed at Arthur D. Little, Inc. (ADL) assigns an index value to each item according to its relevance for each query. Again, however, the system did not produce a print-out of data in full enough form to enable us to look directly at the density functions supposed to underlie the *ROC* curves.

The ADL system was used with a still larger store, effectively 4,000 items. I have based arbitrary acceptance criteria, again, on the number of items considered as retrieved. The terminal criterion, in this case, was determined by the ADL investigators; they proceeded through the items according to their rank to judge the relevance of each, and stopped when it seemed that relevant items were turning up on a random basis. In order to determine the recall ratio, or hit proportion, the total number

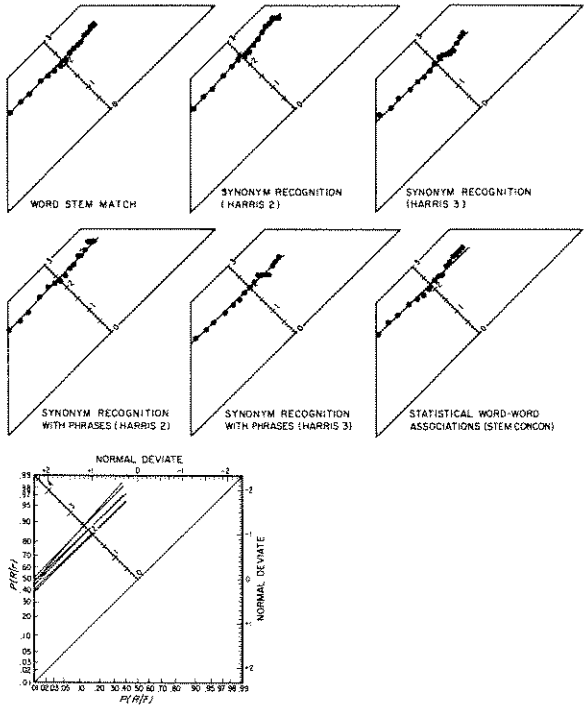


FIG. 9. IRE 1 collection: 6 methods. 405 items, 17 queries. 186 relevant + 6,699 irrelevant = 6,885 total. Criteria: 10, 15, 20, 30, 40, . . ., 150 retrievals. Harvard-Cornell: Salton and Lesk.

of relevant items for each query had to be established. These numbers were estimated at ADL from a sample of 400 items drawn from the store of 4,000 items.

Included in the following figures are almost all the data, and all the major data, collected at Arthur D. Little, Inc. A difference between these and foregoing plots is that most of these are based on single queries. The data points, surprisingly, do not show much greater scatter about a line, but substantially greater variation in the slopes is evident.

Figure 16 shows the associative retrieval method applied to four queries which consisted of abstracts ("full text queries"). Also shown is the same method applied to briefer forms of the same queries. In the latter case ("CBU queries") the queries consisted of critical word strings selected from the abstracts, designated as "content-bearing units." The full *ROC* plots show the pooled results for queries 1, 3, and 4 for each type of query. Query 2 was excluded from the pooled results because the range of acceptance criteria available for it was relatively limited, and various means of pooling queries with different ranges of acceptance criteria proved unsatisfactory. If the curves of the last two plots are extrapolated to the negative diagonal, values of E are obtained (ap-

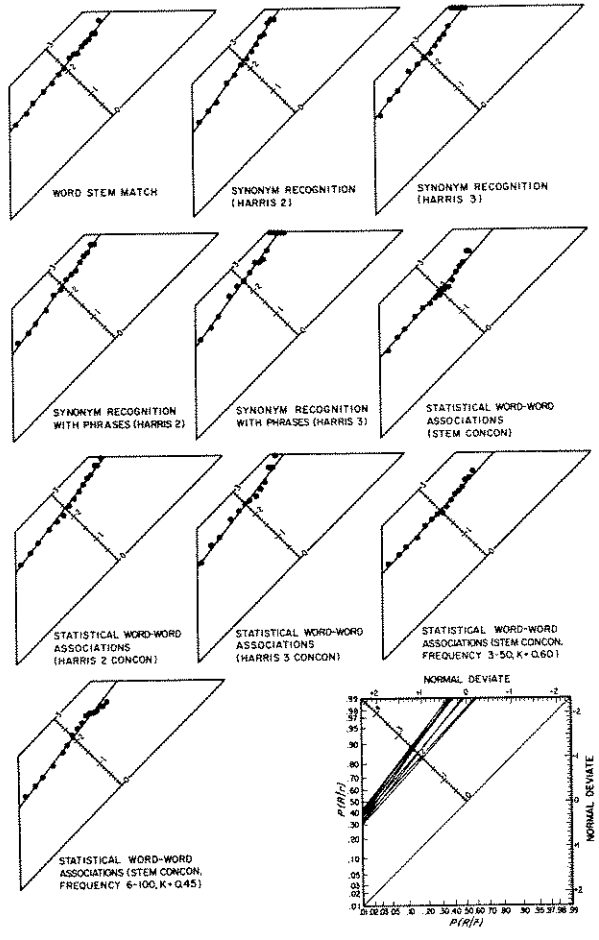


FIG. 10. IRE 2 collection: first set of 10 methods. 380 items, 17 queries. 181 relevant + 6,279 irrelevant = 6,460 total. Criteria: 10, 15, 20, 30, 40, . . ., 150 retrievals. Harvard-Cornell: Salton and Lesk.

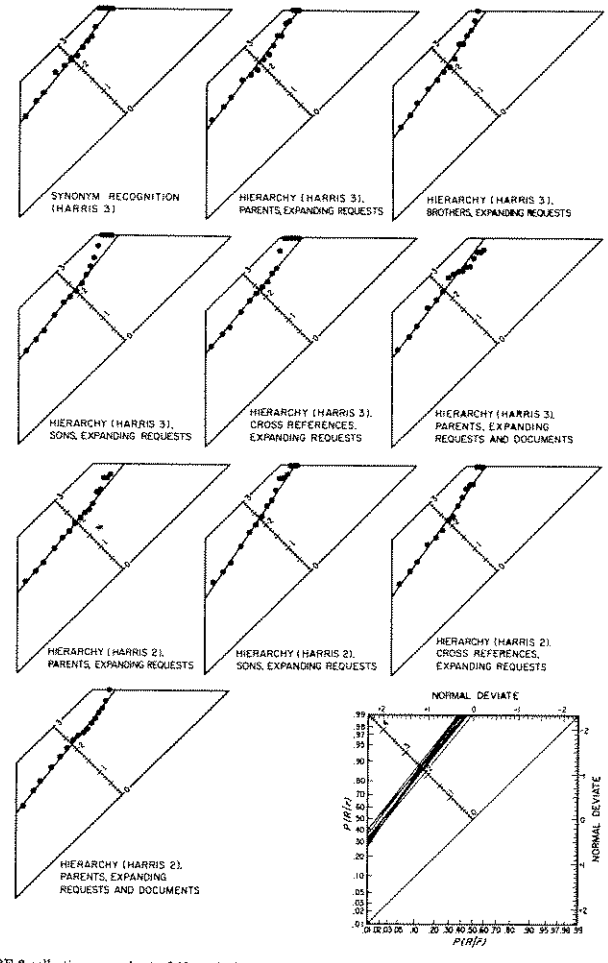


FIG. 11. IRE 2 collection, second set of 10 methods. 380 items, 17 queries. 178 relevant + 6,282 irrelevant = 6,460 total. Criteria: 10, 15, 20, 30, 40, . . ., 150 retrievals. Harvard-Cornell: Salton and Lesk.

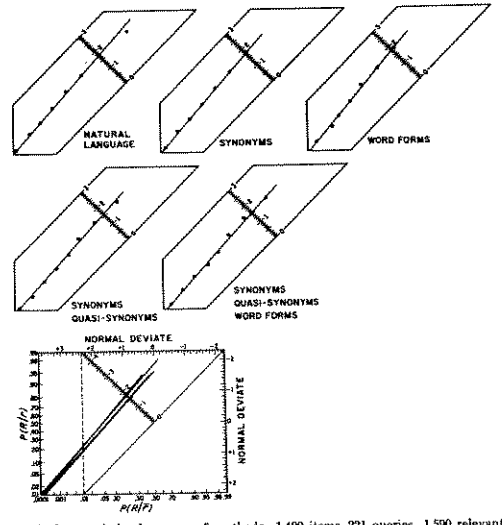


FIG. 12 Recall devices, single-term index language: 5 methods. 1,400 items, 221 queries. 1,590 relevant + 307,810 irrelevant = 309,400 total. Criteria: levels of coordination. Cranfield: Cleverdon and Keen.

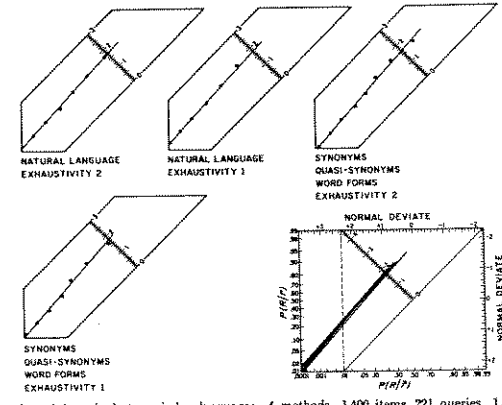


FIG. 13. Indexing exhaustivity, single-term index language: 4 methods. 1,400 items, 221 queries. 1,590 relevant + 307,810 irrelevant = 309,400 total. Criteria: levels of coordination. Cranfield: Cleverdon and Keen.

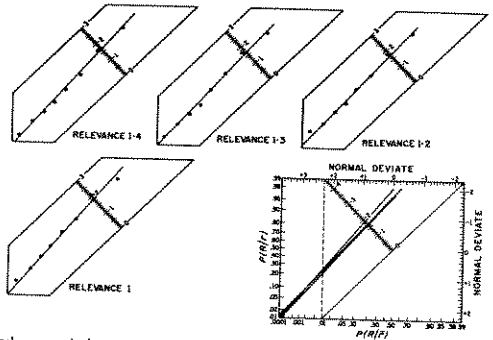


FIG. 14. Document relevance, single-term index language, natural language: 4 methods. 1,400 items, 50 queries.

Relevance 1–4: 361 relevant + 69,639 irrelevant = 70,000 total.
Relevance 1–3: 297 relevant + 69,703 irrelevant = 70,000 total.
Relevance 1–2: 155 relevant + 69,845 irrelevant = 70,000 total.
Relevance 1: 95 relevant + 69,905 irrelevant = 70,000 total.
Criteria: levels of coordination. Cranfield: Cleverdon and Keen.

proximately 1.30 and 2.20) that lie in the range of empirical values noted earlier. The slopes of the lines (approximately 1.00 and 1.50) are also in the range of empirical values noted earlier.

Figure 17 shows three different retrieval methods used with the short queries. Figure 18 shows another method and reproduces the fitted lines for the four methods of Figs. 17 and 18 for each query. There is a tendency for the slopes to depend more upon the query than the method. Averaging over methods, the slopes range from about 1.00 for query 4, through approximately 1.45 for queries 2 and 3, to about 1.75 for query 1. Average slopes for the four methods lie between 1.28 and 1.52. The average values of E associated with the four methods, by extrapolation, range from 1.60 to 2.10.

We may note that the highest value of E, 2.10, is obtained with the method called "selected associations," shown in panels (i) through (l) of Fig. 17. It can be seen that, in fitting straight lines to the data obtained with that method, data points falling below the line at the lower false-drop probabilities were virtually ignored in the case of two queries (queries 1 and 4). Clearly, if we were to restrict our interest to low false-drop probabilities—say, if we were to consider only the left-most half-dozen or so points—then the slopes for that method would be steeper, and the values of E estimated would be higher. In fact, if the four queries are pooled with only the left-most nine points included, the resulting value of E is close to 3.0 (and the resulting slope is about 1.8). The "selected-associations" method is one of two methods tried at ADL with user intervention between iterative searches. The other method in which adjustments were made between iterations is the one called "reweighted associative," shown in Fig. 18; in that case all the data points are quite well taken into account in fitting lines, and an $E=1.90$ is obtained.

• Conclusions

The consistent linearity of the empirical relative-operating-characteristic curves confirms that a decision-theory measure can be used to reflect solely the effectiveness of a retrieval system, and effectiveness unconfounded by variation in the acceptance criterion. The apparently irregular variation in the slopes of the curves presents a slight complication relative to achieving a measure that is a single number, but not enough of a complication to impair seriously the usefulness of a decision-theory measure.

Two numbers—E measured at the negative diagonal of the ROC space, and the slope, s—give an accurate description of the curve representing constant retrieval effectiveness over varying acceptance criteria. Two numbers are not as convenient as one, but these particular two give a considerably more economical description of the performance curve than available previously and can be reported in cases where conveying information about the full curve is desirable.

The data at hand indicate, however, that for most purposes conclusions about effectiveness can be drawn from

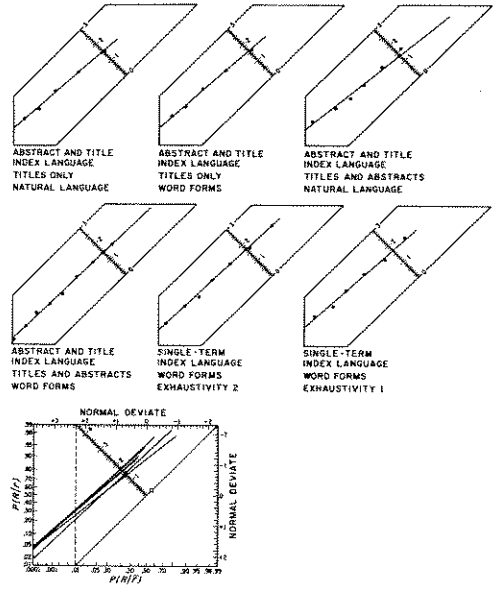


FIG. 15. Abstracts and titles: 6 methods. 200 items, 42 queries. 198 relevant + 8,202 irrelevant = 8,400 total. Criteria: levels of coordination. Cranfield: Cleverdon and Keen.

the value of E alone, without regard to s. In short, there is little point in concern over small differences in s when differences in E are small. We have seen that when values of s are based on more than a few queries they do not vary enough to obscure a substantial difference in E.

What constitutes a "substantial difference" in E, or a difference of practical significance? An approximate answer derived from the present data is that a difference in E in the neighborhood of 0.30 to 0.50 is a reasonably significant one. Thus, for example, in the Harvard data based on the IRE collections (Figs. 9, 10, 11), a difference between two methods of that magnitude corresponds to a factor of about two in the false-drop probability. (By way of illustration, it can be seen in Fig. 9 that at a hit probability of 0.90 the extreme methods show false-drop probabilities of approximately 0.25 and 0.13; at a hit probability of 0.70 the extreme false-drop probabilities are about 0.13 and 0.07; at a hit probability of 0.50 the extreme false-drop probabilities are about 0.02 and 0.01.) It seems unlikely that a smaller experimental difference would have much practical import.

As discussed earlier, if it should seem worthwhile to have a measure that is both a single number and sensitive to variation in slope, the distribution-free measure A could be used. Let us use the measure A now to get a different view of the observed differences among methods in the present sample, a view that will help us judge how small a difference in E is practically significant. A, it will be recalled, is the proportion of the area of the ROC space that lies beneath an ROC curve plotted on linear scales (as in Fig. 4), and is equal to the probability of choosing between two items, one drawn at random from the relevant set and the other drawn at random from the irrelevant set, the item that is relevant. Assume for the purpose at hand that all of the ROC curves in our sample are of unit slope; this approximation introduces a distortion that is negligible relative to the point of interest here, and permits a conversion from E to A by

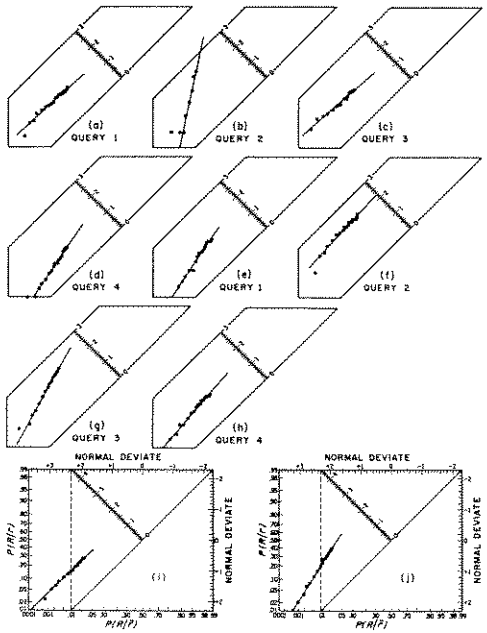


FIG. 16. (a)–(d): Fully automatic associative, 4 full-text queries. (e)–(h): Fully automatic associative, 4 CBU queries. 4,000 items. Number relevant: Query 1, 80; Query 2, 30; Query 3, 70; Query 4, 100. (i), (j): Average of Queries 1, 3, and 4. (i): full-text queries; (j): CBU queries. Criteria: 5, 10, 15, 20, 30, 40, . . ., retrievals. Arthur D. Little, Inc.: Giuliano and Jones.

means of published tables (*8*). For the Harvard data, values of A, or values of the probability of a correct choice in a two-alternative forced-choice test, denoted $P_2(C)$, range from 0.74 to 0.78 for the ADI collection (Fig. 7), from 0.85 to 0.90 for the Cranfield collection (Fig. 8), and from 0.92 to 0.96 for the IRE collections (Figs. 9, 10, 11). For the Cranfield data, $P_2(C)$ ranges from 0.87 to 0.91 for the large collection (Figs. 12, 13, 14) and from 0.83 to 0.89 for the small collection (Fig. 15). For the data collected at Arthur D. Little, Inc., the range of the four "CBU" methods (Figs. 17, 18), averaged over the four queries, is from 0.87 to 0.93. It might be argued, again, that the differences between extreme methods for any collection, of 0.04 to 0.06, are real differences, but it seems unlikely that differences of less than 0.04 in $P_2(C)$ have material implications.

These values of $P_2(C)$ lying between 0.74 and 0.96 indicate that present retrieval methods leave considerable room for improvement. (Said otherwise, these values of $P_2(C)$, considered along with the competence and diligence with which the experiments here represented were pursued, indicate that information retrieval is a very difficult problem.) On the face of it, choosing the single relevant item from a collection of two items is not a demanding task, and we should hope that our retrieval systems would make the correct choice almost every time, say, with a probability of 0.99 or greater. A more compelling impression, however, of the current state of the retrieval art is gained by taking pairs of hit and false-drop probabilities from the empirical *ROC* curves and converting these probabilities to raw numbers.

Consider an *ROC* curve with $E=2.5$ and $s=1.3$. This

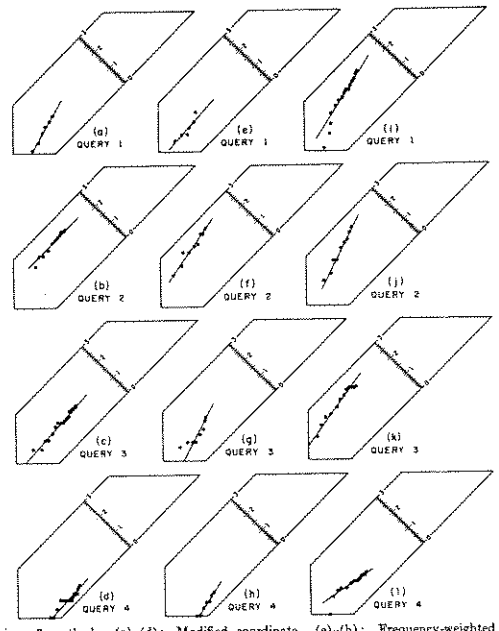


FIG. 17. CBU queries: 3 methods. (a)–(d): Modified coordinate. (e)–(h): Frequency-weighted coordinate. (i)–(l): Selected associations. Number of items, number relevant per query, and criteria as in Fig. 16. Arthur D. Little, Inc.: Giuliano and Jones.

curve is close to the best of the curves seen in the foregoing, and exceeded by none of them. It passes through the points $P(R|\mathrm{r})$ and $P(R|\mathrm{r})$ having coordinate values of (0.001, 0.12), (0.01, 0.42), and (0.10, 0.88). Assume a file of 3,000 items and a group of queries to each of which 10 of the 3,000 items are relevant. Now, if we will settle for retrieving, on the average, only 1 of the 10 relevant items per query, we will also receive 3 false drops each time. If we desire 4 of 10 relevant items, we will have to winnow the 4 from 30 irrelevant items. If we should aspire to 9 of 10 relevant items, we would have to examine more than 300 items in response to each query to find the 9.

These noise-to-signal ratios are dramatically large. The ratio mounts rapidly even for a file as small as 3,000 items: from 3 to 7 to 33 for the three acceptance criteria of the example. For a file of 10,000 items the corresponding noise-to-signal ratios are 10, 25, and 100 plus. It is with these ratios in mind that I earlier suggested dismissing small differences in E and ignoring small variations in s.

The decision-theory analysis can be seen to set the stage clearly for identifying an important advance in retrieval technique. The best of the performances sampled here, in the vicinity of $E=2.5$ and $s=1.3$, gives a false-drop probability of approximately 0.10 for a hit probability of 0.90. Assuming the same slope, and taking the same hit probability, an $E=3.0$ corresponds to a false-drop probability of 0.05, and an $E=3.8$ corresponds to a false-drop probability of 0.01. An $E=4.0$ means a false-drop probability of 0.005, or reception of 15 unwanted items along with 9 of the 10 wanted items from a file of 3,000. An $E=4.5$ means a false-drop probability of 0.001, or reception of 3 unwanted items along with 9 of the 10 wanted items from a file of 3,000.

A belief of several people working in the retrieval field is that a very significant advance in retrieval effectiveness will be achieved in the near future by "on-line" systems,

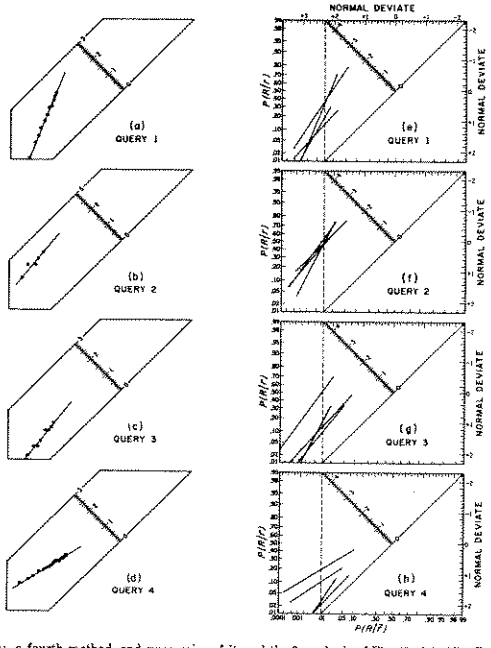


FIG. 18. CBU queries: a fourth method, and summaries of it and the 3 methods of Fig. 17. (a)–(d): Reweighted associative. (e)–(h): results of 4 methods for each query. Number of items, number relevant per query, and criteria as in Fig. 16. Arthur D. Little, Inc.: Giuliano and Jones.

in which the user is given immediate feedback and enabled to progressively refine the search prescription over successive trial searches. It will be informative to apply the decision-theory analysis in experiments on on-line procedures. Will we see values of E in the vicinity of 3.0, or 3.5? Might we even find values of E about 4.0—or will present knowledge of language forms impose a barrier at a lower level of effectiveness?

References

1. SWETS, J. A., Information Retrieval Systems, *Science*, 141:245–250 (1963).
2. BOURNE, C. P., *Annual Review of Information Science and Technology*, Interscience, New York, 1966, Chapter 7, pp. 176–179.
3. CLEVERDON, C., and M. KEEN, *Factors Determining the Performance of Indexing Systems*, vol. 2, *Test Results*, Association of Special Libraries and Information Bureaux, Cranfield, England, 1966.
4. GREEN, D. M., and J. A. SWETS, *Signal Detection Theory and Psychophysics*, John Wiley, New York, 1966, pp. 45–51.
5. SALTON, G., M. LESK, et al., *Information Storage and Retrieval*, Scientific Report No. ISR-11, Department of Computer Science, Cornell University, 1966.
6. GIULIANO, V. E., and P. E. JONES, *Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems*, Interim Report, Arthur D. Little, Inc., Cambridge, Mass., 1966.
7. LESK, M., and M. KEEN, Personal communications, 1967.
8. SWETS, J. A. (Ed.), *Signal Detection and Recognition by Human Observers*, John Wiley, New York, 1964, pp. 682–683.

Part V: Tools and Ideas

This is a miscellany which extends and/or explodes basic theory in the field, and the inclusion of these papers is intended to give a broader perspective to information science. There are quite good things which lie nearly beyond the fringe, generally applying a technique to the analysis of a problem in librarianship.

Reprinted from the Journal of the American Society for Information Science, Vol. 27, No. 5, 1976.

Availability Analysis

A recently proposed technique for determining the availability of items in a circulating collection is expanded and applied. We present a discussion of the theoretical justification and a careful analysis of the sources of error and the reliability of the results. We report the results of a new longitudinal study using this technique and apply it retrospectively to a number of previous studies to which it is applicable. The resulting data indicate certain striking regularities in the performance of large university libraries, as well as significant differences.

The implications of this tool, both for management and for the scientific analysis of various aspects of library performance, are discussed.

Paul B. Kantor
School of Library Science
Case Western Reserve University
Cleveland, OH 44106

• Introduction

In evaluating the effectiveness of an information service, there are important quantitative and qualitative aspects to the meaning of the phrase "service to the users." The quantitative aspects have been discussed at some length in an earlier paper (*1*) where we have argued that the *Total Contact Time* is probably the best single quantitative measure of service rendered in the specific field of book circulation and in-library use.

The present work is directed to the analysis of the qualitative aspects of service.

There are three important qualitative dimensions to the service rendered: relevance, accessibility and availability. The last of these factors is particularly amenable to mathematical analysis. In particular, the probability of satisfaction (on author-title searches) can be resolved into several conditional probabilities which are separately meaningful and which respond essentially independently to changes in policy.

The outline of this paper is as follows. In Section 1, we review the basic ideas of this type of availability analysis. In Section 2, we present some new data on a "before and after" study conducted at Case Western Reserve University. In Section 3, we essay a retrospective analysis of a number of important frustration studies which have been carried out by other authors prior to the development of this analytical tool. In Section 4, we assemble the various results, comment and look toward the future.

• Availability Analysis

Let us suppose for the moment, that there is only one reason preventing library users from finding the titles they seek. Specifically, we assume that either the book is circulating or the user will be satisfied. If we denote the number of inquiries by W, the number of Dissatisfactions which are caused by Circulation by DC and the number of user Satisfactions by S we have:

$$W = S + DC$$

For example, if $W = 100$ and $DC = 20$, then $S = 80$. We can also say that the probability that a book is not circulating, p_C, is 80 percent.

Now let us make a more realistic assumption. Even if the book is not circulating, the user does not necessarily find it. Suppose, to be specific, that 10 of the 80 items not in circulation have been misshelved (Dissatisfaction caused by Library Performance). Thus, the number of satisfactions is reduced to 70. However, performance at the level of circulation remains 80 percent. To describe the additional dissatisfaction we introduce a second performance factor p_L, which measures the probability that a book which is *in* is also correctly shelved. Thus p_L = 70/80 = 87.5 percent. Of course, we could also define the fraction of all books sought which are correctly shelved, which is 70 percent. Let us see why that does not make sense.

Suppose for some reason (a change in circulation policy) a new study shows that (on the average) 12 out of 100 books are circulating. Then of each 100 books, there are now 88 which are in the library. Presumably the eight additional books are no easier or harder to shelve than the other 80, and so (on the average) 7/8 X 8 = 7 of them will be in the correct place, while one of them will be out of place. Thus the second performance factor becomes 77/88 = 7/8, which is the same. Of course, if we were comparing the number of satisfactions to the number of inquiries we would now have 77 percent instead of 70 percent.

Obviously the correct way to look at it is that there are two factors p_C and p_L, and one of them can remain constant while the other one changes:

Before: 8/10 X 7/8 = 70%

After: 88/100 X 7/8 = 77%.

Finally, there could be a change in the accuracy of reshelving, without a change in circulation. Suppose it rises to 90 percent. Then the overall satisfaction is given approximately by the product

$$8/10 \times 9/10 = 72\%.$$

If we think about it more carefully, we see that a change in the accuracy of the reshelving will produce some change in the probability p_C. Roughly, if more books are reshelved correctly all the time, then more of them will have been found by previous borrowers and will be circulating at the time of the study. Detailed analysis shows that this effect is, in general, less than a 1 percent in p_C. Thus, for *practical purposes* p_C and p_L are independent.

Now that the principle is clear we can easily see how to deal with the four principal categories of dissatisfaction:

DA = dissatisfaction due to the books never having been acquired.

DC = dissatisfaction due to the books having been in circulation.

DL = dissatisfaction due to the books having been misshelved, lost, etc.

DU = dissatisfaction due to the user's error.

An actual frustration study (*1, 2*) results in six numbers, the number of inquiries (W); the number of satisfactions (S); and DA, DC, DL, DU.

The corresponding performance measures are denoted by p_A, p_C, p_L and p_U. They are calculated using the branching diagram shown in Fig. 1:

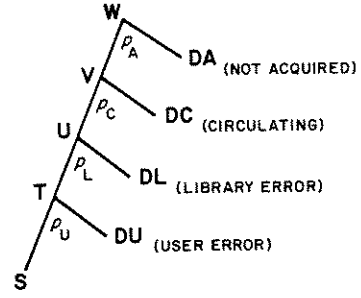


Fig. 1. The principal categories of dissatisfaction in correct order.

and the formulae,

T = S + DU	p_U = S/T
U = T + DL	p_L = T/U
V = U + DC	p_C = U/V
W = V + DA	p_A = V/W

The performance measures combine multiplicatively to determine the total probability of satisfaction p_S:

$$p_S = p_A\, p_C\, p_L\, p_U$$

We have discussed elsewhere (*1*) the relevance of such an analysis for deciding how to improve library procedures. That discussion can be summed up by a rule of thumb: attack the lowest factor first, and improve it until it is no longer the lowest.

• A Longitudinal Study of Availability

The techniques described here have been applied to a pair of studies at the Freiberger Library of Case Western

Reserve University, a nontechnical library containing about 1,000,000 items. In fact, the analysis technique was developed between the occurrence of the two studies, so that the earlier data represents a retrospective analysis of data collected by students under the guidance of Professor Tefko Saracevic. The later data were collected by students studying the technique of availability analysis.

The study itself uses two instruments. One is a form (Fig. 2) handed to each user as he approaches the author-title catalogue. It is designed to obtain the information which *we* need (author, title, and call number) in a mode which actually facilitates the user's search, by providing him with "scratch paper." It has a column marked "can't find", a column for reporting the "cause code", and some additional columns related to a subproject not reported here.

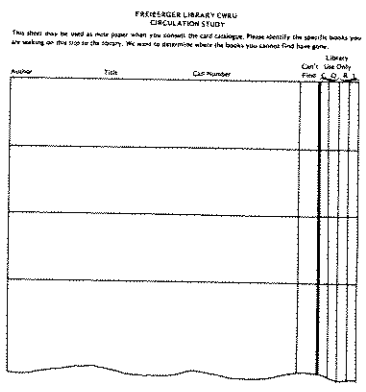
With the form, the following statement is made to the user: "We are studying how well this library satisfies its users. Would you use this as 'scratch paper' and put a check in this column (point to it) for each book that you can't find? Just drop it off here when you leave. Thank you." (The study was aided by the fact that one person can easily monitor the A–T catalogue and also the exit, to remind participants to drop off the forms when leaving.) The data requested at the bottom of the form can be used for a differential study, but it requires a larger sample.

The second instrument is a check list* to guide the investigators in tracking down the books which the user is unable to find. The major headings are:

- C1. Not owned by the library,
- C2. Incorrect call number,
- C3. Book located in a special area identified on catalog card,
- C4. Book properly shelved,
- C5. Book misshelved,
- C6 Book recently used or in use in library,
- C7 Book in preshelving,
- C8 Book located in (or listed in) an area not identified on card catalog,
- C9 Book circulating,
- C10. Other.

The specific subheadings should be drawn up by someone familiar with the specific library being studied. At least two investigators are required since one continues to monitor the catalogue and exit, while the other(s) immediately investigate those books which the user could not find.

*We are indebted to Dr. William Shaw for preparing the list.


FREIBERGER LIBRARY CWRU
CIRCULATION STUDY

This sheet may be used as note paper when you consult the card catalogue. Please identify the specific books you are seeking on the slip on the library. We want to determine where the books you cannot find have gone.

Author	Title	Call Number	Can't Find	Library Use Only C	O	R	L

Fig. 2. The data collection form handed to patrons.

It took about 12 hours (24 person hours) to collect data on 312 titles during April 1975. The categories C1–C10 were assigned to the four performance categories in the following fashion:

$$DA = C1$$
$$DU = C3 + C4 + C2$$
$$DL = C5 + C7 + C8 + C10$$
$$DC = C9 + C6$$

Multiple copies were treated as follows. If there are two copies, one in each of two different categories, we counted 1/2 a dissatisfaction in each. This results in logically consistent counting in all cases.

The numbers were:

$$S = 203$$

C1 - 30	C2 - 4	C3 - 0	C4 - 8	C5 - 4
C6 - 6	C7 - 5	C8 - 7 1/2	C9 - 18 1/2	C10- 26

Thus:

$$DA = 30 \quad DC = 24\ 1/2 \quad DL = 42\ 1/2 \quad DU = 12$$

These are summarized by the branching diagram shown in Fig. 3.

When using these performance measures as a guide in policy planning, it is important to be aware of the limits

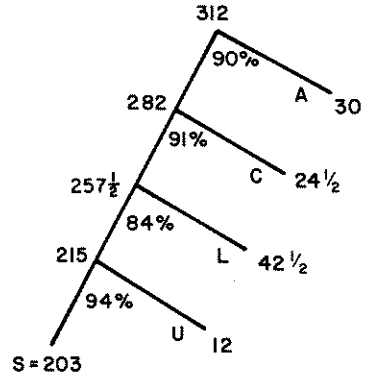


Fig. 3. Results of the April 1975 study at Freiberger Library.

of significance of the figures obtained. Small variations with time, or differences between institutions, are not likely to be significant, while large ones can be very meaningful. To explain what we mean by "small" and "large" we must discuss the sources of error and uncertainty.

There are basically two sources of uncertainty, which we will refer to as statistical error and systematic error.

Statistical error arises because we do not determine the fate of every item search which is made at the library. We use a sample, and there are well understood procedures for determining the probable deviation of the sample value from the true value. We quote the standard error, determined on the assumption that the branches of the diagram are independent and that individual searches are uncorrelated Bernoulli trials. This results in the simple formula

$$\delta p = [p\,(1\text{-}p)/N]^{1/2}$$

N = the number of cases to which p applies.

The systematic error arises from two sources. The first is simply mistakes in counting the number of dissatisfactions. These errors are minimized by being as careful as we can and double checking all cases in which no positive identification could be made. We believe that the remaining errors of this type are negligible.

The second source of systematic error arises from the difficulty of assigning the observed categories C1. . . .C10 to the theoretical categories *DA*. . . .*DU*. In most cases there is no problem, but there are some points which cannot be resolved by theory alone.

One example will serve to illustrate this idea. Reserve books are somewhere in limbo between library errors and "not in the collection." However, they cannot be circulating, so they cannot be entered on the branching diagram at an intermediate position. If a great many reserve books are encountered in the study, it is probably best to treat them as a separate collection as we have done below in analyzing the data collected by Meier. Even in that case, however, the library may be frustrating the reader by *concealing* the true location of the book. For this reason, unless the number of such items is very large we simply include them in the category *DL*.

Systematic uncertainties can be of the order of 2 to 3 percent.

Our experience shows that data can be collected at a rate somewhere between 8 and 20 items per person-hour. This depends, of course, on the rate at which requests arrive, the size of the library, its performance factors and so on. However, these represent reasonable limits. The task of checking up on causes cannot be done much faster than about ten items per hour. Remember that more than 50 percent of the items are found and do not require further work. If the rate at which requests arrive is too low, it becomes inefficient to use this study tool at all.

If it is important to know more about the details of the performance picture (for example, in order to make improvements in policy), it becomes necessary at this point to attack the problem from the other direction. For example, suppose it has been determined that library procedures are the limiting performance factor. The next step is not *statistical* analysis of the categories (C) involved, it is *management* analysis of policies, working procedures and perhaps personnel. Usually such analysis will point up inconsistencies or bottlenecks which have grown up in the system, perhaps over the years, that can be eliminated by establishing new procedures or (sometimes) simply by following, rather than ignoring, old ones. The function of availability analysis in this example is to tell us that library procedures (rather than collection size, user skills, or circulation policy) are what need to be studied further. As always, analytical techniques can serve to guide the study of policy, but they cannot replace it.

In our tabulated results only the statistical error is quoted. By the reasoning outlined above we estimate that systematic uncertainty is of the order of 2 to 3 percent. Combining these we conclude that the results of our availability analysis are reliable to within about

4 or 5 percent. Thus, a change or difference of more than 5 percent is to be regarded as significant, while a smaller change in any particular performance measure may simply be due to the effects of systematic and statistical uncertainties.

The statistical formula can be used to estimate the sample size needed. For example, to achieve the value $\delta = 2\%$ for a (typical) value of $p = 85\%$ requires a sample of size

$$N = (85\%) \times (15\%) / (2\%)^2 = 320$$

• Comparison With Previous Performance

An earlier study done at the same library, in the Spring of 1973 (*3*) can now be analyzed in exactly the same fashion. The principal features of that analysis are shown in Fig. 4:

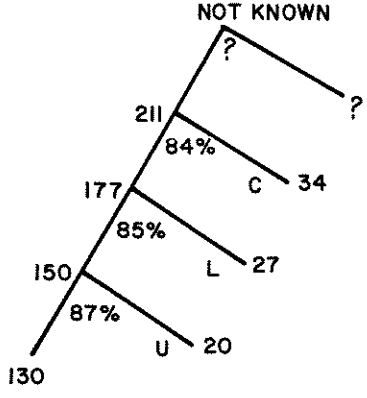


Fig. 4. Results of the Spring 1973 study at Freiberger Library.

The statistical errors can be calculated in the same fashion. The results are summarized in Table 1.

We note that two of the three performance measures, for which a change might be observed, show significant change. The crucial question to ask is: "Can we identify policy changes which would account for this?"

Table 1.

Category	Performance 1973	1975	
A	not known	90 ± 1.7%	
C	84% ± 2.5%	91.3 ± 1.4%	*
L	85 ± 2.7%	83.5 ± 2.4%	
U	87 ± 2.7%	94 ± 1.6%	*

With regard to circulation the answer is definitely yes. Between the two studies the library changed from a semester loan policy to a four week loan policy. This should, of course, decrease the dissatisfactions caused by circulation. In fact, using very reasonable approximations we can account not only for the existence of the change, but also for its magnitude.*

With regard to the *user* performance, p_U, we cannot be so certain. There has in fact been an orientation program introduced between the two studies. We do not have a mathematical means of predicting the effect of such a program, nor at the present time do we have any other "before-and-after" studies on such programs. The sign of the change is certainly as we would expect. But, in all honesty, we cannot say whether the observed 7 percent increase can be reasonably attributed to the orientation program.

In summary, the data presented here provides encouraging support for the hypothesis stated in Section 1, that the parameters described will characterize the availability aspect of library performance and will respond independently to changes in policy. This is the second such longitudinal study to be reported. The first was done at the Sears (Technical) Library of Case Western Reserve University by Saracevic, Shaw and the author and has been described in some detail elsewhere (*2*). The results of that study are also in agreement with the hypothesis of this analysis, and are summarized in the third and fourth columns of Table 2.

*If requests arrive uniformly through the semester at a rate r, then the probability that a book is *not* out (p) is determined by the average loan period (t) (in the one copy approximation) by the formula $p = 1/(1 + rt)$. We do not precisely know t, but a good estimate is $t = 50$ days for semester loan and $t = 20$ days for 4 week loan (*1*). For technical reasons discussed elsewhere, (*1*) this p cannot be compared directly with p_C, but must be compared with an adjusted figure defined by

$$p = S/(S + C9),$$

whose value is 79% (1973) and 92% (1975).

Assuming that there was no significant change in the average demand, the theoretical result would predict a rise from 79 percent (given) to 90.4 percent (calculated). This is obviously in excellent agreement with the 92 percent actually observed.

• Retrospective Analyses

A variety of studies have been carried out over the past 20 years, each of which tells something about the availability of library materials. Some are amenable to retrospective analysis using the branching diagram technique. The crucial ingredient is obviously a record of the *number of satisfactions*, as well as an analysis of the number of failures or dissatisfactions attributable to each cause. Those studies which did not record the number of satisfactions cannot be analyzed in our terms. The point is that the apportionment of the dissatisfactions to the various causes is not a set of independent parameters and, if any one of them changes, all the rest will change also. That is why the additional effort required to record the number of satisfactions is a justified expense.

The earliest study we found which is amenable to this type of analysis is one conducted by Meier and reported in 1957 (*4*). He studied the closed stack library of a large university, and thus the category of "user error" does not apply. Since the study began with the receipt of filled out book slips, no items were included unless they were listed in the card catalogue. This means that the category DA, and the corresponding measure p_A are not included in this study. An enormous number of requests (4614) were analyzed, of which 58 percent were delivered immediately (*S*). The remaining 42 percent were divided as follows: 12 percent elsewhere (that is, located in other parts of the university library system); 24 percent were in the reserve collection; 32 percent were circulating; and 32 percent were "not found." Finally, of those which were not found, a portion (about 2/3rds) were looked for, and 82 percent of them were located. From this array of data we can conclude the value of p_C without ambiguity. The value of p_L (which in some sense represents all the remaining dissatisfactions) invites further analysis.

The category of items which were "not found" presumably corresponds to two sources of dissatisfaction: (1) books or items which the library has lost track of; or (2) items which were in or near their proper place but were missed by the staff members who looked for them. This human source of error corresponds to the kind of user errors which we have defined, and gives us some idea of how well the users might be expected to perform by comparing them with library staff. We can analyze this possibility by supposing that the 84 percent figure (Fig. 5) represents items which could be located in their proper places. This results in a lower bound on the human performance factor of the library staff. The numerical results are obtained from the diagram shown in Fig. 5.

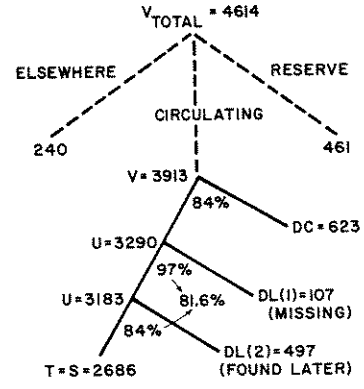


Fig. 5. Results of the study by Meier (see text). It was necessary to treat the data as belonging to three distinct collections with the same access point.

The analysis has been somewhat complicated by the fact that there are essentially three different collections with access through the same central card catalogue: the circulating collection, the reserve collection and the collection of materials stored elsewhere. With regard to the first, we can define the performance measures in the usual fashion. With regard to the other two, we do not have enough additional information to consider them.

In the present retrospective study, we must content ourselves with an analysis of the circulating collection alone. The results are: $p_C = 84$ percent, $p_L = 81.6$ percent and a lower bound on the human error of 84 percent. The "missing books" performance factor is 97 percent, reflecting the enormous advantages, *in this regard*, of a closed stack system.

A second study amenable to the present availability analysis was conducted by Buckland and colleagues at the University of Lancaster Library (*5*). They identified 18 distinct causes of failure. A total of 188 items were obtained, and 132 were identified but not obtained. We have assigned the 18 causes of failure as follows:

DA = not owned or on order.
DA = on loan, second copy also on loan.

DL = missing, missing and on order, binding, awaiting photocopy, awaiting binding, temporarily removed.

DU = on shelves, on short loan shelves, in stack, kept at service desk.

The resulting analysis yields the branching diagram of Fig. 6. (In the case of double copies, we count each single item as contributing 1/2 dissatisfaction, etc.) The analysis in this case is complicated slightly by the fact that, with the type of questionnaires which were used, there were 32 cases of dissatisfaction in which they could not identify the item sought. Therefore, only 132 of the 165 dissatisfactions could be analyzed. We dealt with this by assuming that the other 33 were also item searches, and they would have been distributed in the same way as the cases which were analyzed. This means that for each branch of the diagram, the analyzed cases represent only 132/165ths (which equals 4/5ths) of the true number of dissatisfactions. They are corrected by multiplying by 5/4ths, as shown in Fig. 6.

Urquhart and Schofield (*6*) report on a study in which users were asked to place slips on the shelves where the books they were looking for should have been. They did not directly measure the number of satisfactions.

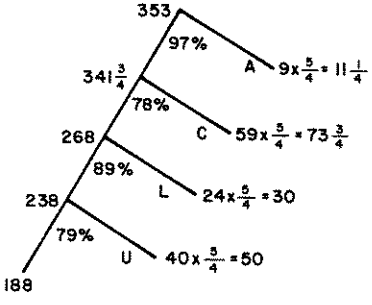


Fig. 6. Results of the study by Buckland (see text). The correction factor 5/4 has been applied because only 4/5ths of the cases in that study were analyzed in detail.

When information about the number of inquiries is not available at any corner of the branching diagram (that is, *W* or *V* or *U* or *T*), it is not possible to calculate any of the performance measures. Thus the data which these authors obtain cannot be used to provide performance measures.

A similar study by Seymour and Schofield (*7*), in which detailed information was reported on the type of user and the source of his information, also could not be analyzed because the number of satisfactions is not reported.

N.K. Kaske (*8*) collected data for the University of Oklahoma Library of a related type under the heading "Collection Status Report." His sample of volumes was drawn in a significantly different fashion. To determine all factors except misshelving he used a random sample drawn from the shelf list. This means there was no user participation (hence no *DU* category) and inactive volumes are included more often then they would be in a study of actual users. Thus the value called p_C in what follows is probably an overestimate of the actual p_C as experienced by users. The number of misshelved volumes was determined directly, that is, the shelves were sampled to determine what fraction of all volumes are shelved out of sequence. The results shown in Fig. 7 are for his first study, December 3, 1972.

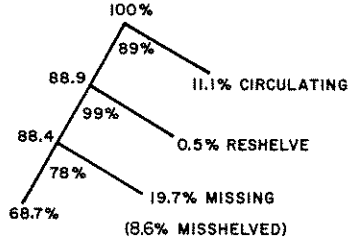


Fig. 7. Results of the study by Kaske (see text) done on December 3, 1972. Books in the reshelving areas were counted separately.

Roughly speaking, the 78 percent is a library performance figure and the 89 percent is an upper bound for the circulation performance. Sample size is about 600 volumes.

The Kaske study was repeated at two later dates, January 14, 1973 and April 15, 1973, with results summarized in Table 2. As Kaske points out, the circulation factor responds to changes in academic activity. The high value (94 percent) was measured during the first week in a term. We are also able to see that the library performance factor remains essentially constant at 79 ± 2%. (The "significant differences" which Kaske noted are due to circulation or to the change in the number of

Table 2. Performance Measures

Case Western Reserve University

	FREIBERGER '73 '75	SEARS '72 '74	MEIER 1961	BUCKLAND	KASKE I	KASKE II	KASKE III
p_A	--- 90	88 91	----	97	---	---	---
p_C	84 91	77 87	84	78	≤89	≤94	≤89
p_L	85 84	89 86	81.6	89	78	81	79
p_U	87 94	80 82	----	79	----	----	----
p_S	~56* 64	48 56	----	53	----	----	----

*assuming $p_A = 90\%$

volumes misshelved. From the *users* point of view there is no significant change in library performance.)

Additional studies, partially amenable to analysis in that they present the product of several performance factors together, have been presented by Tagliocozzo and Kochen (*9*). In connection with a study of catalogue use, they studied the probability that users would find books in the stacks. This corresponds to the product $p_C p_L p_U$ and yielded results of 58 percent at the University of Michigan Medical Library and 69 percent at the Undergraduate Library. If these are combined with a typical figure of 90 percent for the performance of acquisition, we would arrive at satisfaction probabilities of 52 percent and 62 percent respectively.

Their studies of the card catalogue failure rate provide a direct measure of the single parameter p_A. This was determined at four libraries: the University of Michigan Undergraduate Library, the General Library, the Medical Library and the Ann Arbor Public Library. The results, together with the approximate size of the collection in each library, are summarized in Table 3.

Table 3. Values of p_A Obtained From Catalog Use Studies

	Tagliocozzo and Kochen				Lipetz
	Gen. Lib.*	Undergrad.*	Medical*	Ann Arbor Pub. Lib.	Yale
p_A	93.9%	72.5%	89.6%	65.2%	90%
Size of Lib.	1.6M	156K	175K	155K	4M

*Libraries at the University of Michigan included in the study.

Finally, the large study by Lipetz (*10*) of the Yale Library catalogue yields the value of p_A for that library directly. He found that in 10 percent of all searches "the document definitely or probably exists, but was not in the catalogue at the time of the search". This leads directly to the value $p_A = 90$ percent recorded here in Table 3.

• Summary and Conclusions

All of our numerical results are summarized in Tables 2 and 3. In particular, the changes observed in those libraries which were studied twice correlate well with the known changes in policy. This provides important inductive verification for our hypothesis, originally presented on the basis of deductive arguments, that these parameters tend to remain constant unless they are altered by specific changes in policy.

With regard to the Table 2 as a whole, the first observation to make is that the numbers are generally similar to each other (all but one lie in the range of 75 percent to 95 percent). Since the libraries also have many common characteristics--all are large libraries associated with universities in English language nations--we venture the hypothesis that this range of parameter values is indeed typical for such institutions.

Taking 87 percent as a typical value for any entry in the Table 2, we can say that "typically" the probability of satisfaction (that is, the availability of items) is given by

$$p_S = 87\% \times 87\% \times 87\% \times 87\% = 57\%.$$

If, on the other hand, there is no circulation whatever, the availability is given by

$$p_S = 87\% \times 100\% \times 87\% \times 87\% = 66\%.$$

That is to say, even if circulation is completely removed as a source of dissatisfaction, only about two-thirds of the items sought will be promptly found at a "typical" library.

Having noted that there is a general similarity, we then note that there are variations which lie well outside the range of experimental uncertainties. This fact, combined with the two longitudinal studies, suggests that the variations are caused by significant differences in policy.

Thus, three directions for future study are opened.

1. Studies of this kind should be performed at a variety of libraries–such as special and technical libraries, public libraries and reference libraries. These studies will determine what range of values of performance measures is typical of the corresponding class of libraries.

2. The existence of statistically significant differences should be used as the starting point in the search for general relations between policy and effectiveness and, ultimately, guidelines for the selection of cost-effective (optimal) policies.

3. The use of this technique provides a quantitative method for answering a host of troublesome questions, such as "How much does an orientation program really improve user performance?" "Do users have more trouble with the Dewey or Library of Congress classification?" "How much does a change in loan period affect availability?"; etc.

Acknowledgments

This work grew out of studies done at the Complex Systems Institute of Case Western Reserve University, directed by William Goffman. I am indebted to T. Saracevic for making the results of his earlier study available; to the students in LS 509 (O. Mansur, T. Murthy, T. Kochtanek, M. Kinnucan and A. Quijano) for data collection; and to W. Shaw for numerous discussions and criticisms which have led, among other things, to the discussion of the limits of this technique, given in Section 2. The enlightened cooperation and interest of Ann Drain (Head Librarian at the Freiberger Library) and James Jones (Director of the University Libraries at Case Western Reserve University) were essential to the work described here.

At the suggestion of the editor and referees, this paper was condensed from a more detailed manuscript in which several troublesome points dealing with the logical foundations, the mathematical interrelations and the sources of error in Availability Analysis are examined. The complete paper is available from the author upon request.

References

1. **Kantor, P.B.** 1976. "The Library as an Information Utility in the University Context: Evolution and Measurement of Service." *Journal of the American Society for Information Science.* 1976; 27:100-112.
2. **Saracevic, T.; Shaw, W.; Kantor, P.B.** "Causes and Dynamics of User Satisfaction in an Academic Library." *College and Research Libraries.* (to be published).
3. **Saracevic, T.** Personal communication.
4. **Meier, R.L.** 1963. "Information Input Overload: Features of Growth in Communications–Oriented Institution." *Libri.* 1963; 13: 1-44.
5. **Buckland, M.K.; Hindle, A.; Mackenzie, A.G.; Woodburn, I.** 1970. *Systems Analysis of a University Library; Final Report on a Research Project.* University of Lancaster Library Occasional Papers, No. 4. 1970.
6. **Urquhart, J.A.; Schofield, J.L.** 1971. "Measuring Readers' Failure at the Shelf." *Journal of Documentation.* 1971 December; 27: 273-286.
7. **Seymour, C.A.; Schofield, J.L.** 1973. "Measuring Reader Failure at the Catalog." *Library Resources and Technical Services.* 1973; 17(1): 6-24.
8. **Kaske, N.K.** 1973. "Effectiveness of Library Operations: A Management Information Systems Approach and Application." Library Science Dissertation. The University of Oklahoma. 1973.
9. **Tagliocozzo, R.; Kochen, M.** 1970. "Information Seeking Behavior of Catalog Users." *Information Storage and Retrieval.* 1970 December; 6(5), 363-381.
10. **Lipetz, B.** 1972. "Catalog Use in a Large Research Library." *Library Quarterly.* 1972 January; 42: (1) 129-139.