III. Relevance Assessments and Retrieval System Evaluation

M. E. Lesk and G. Salton

Abstract

Two widely used criteria for evaluating the effectiveness of information retrieval systems are, respectively, the recall and the precision. Since the determination of these measures is dependent on a distinction between documents which are relevant on the one hand, and documents which are not relevant on the other to a given query set, it has sometimes been claimed that an accurate, generally valid evaluation cannot be based on recall and precision.

A study was made to determine the effect of variations in relevance assessments on the average recall and precision values used to measure retrieval effectiveness. Using a collection of 1200 documents in information science for test purposes, it is found that large scale differences in the relevance assessments do not produce significant variations in average recall and precision. It thus appears that properly computed recall and precision data may represent effectiveness indicators which are generally valid for many distinct user classes.

1. Introduction

Over the last few years, the interest in the design of automatic information handling systems has steadily increased. At the same time, it has become necessary to devote a good deal of attention to the evaluation of information systems in an attempt to identify those factors which contribute to system effectiveness. Many criteria can be used in such an evaluation process; furthermore, the factors which may be most appropriate in one circumstance may not be in another. In particular, different effectiveness indicators might be generated depending on whether one's viewpoint is the user's, the manager's, or the operator's. The manager, for example, may be most concerned about system costs, whereas the operator may be interested primarily in the characteristics of the equipment used in the process. The user, however, is not normally interested in the equipment, and may be only peripherally concerned with costs. He does, however, want to make certain that the system is responsive to user needs.

Many recent efforts at retrieval system evaluation have been based mainly on user criteria, and while several possible criteria are available - including, for example, the type of presentation of the output, the amount of user effort needed during a search, the time lag between submission of a query and the presentation of search results, and the coverage of the collection being searched - it is generally agreed that the two most important user-oriented measures are the ability of the system to retrieve wanted and at the same time to reject unwanted information. As a result, several of the more recent evaluation studies have used a test methodology based mainly on the computation of the <u>recall</u> and <u>precision</u> values applicable to a set of test queries [1,2,3].

Recall and precision are defined, respectively, as the proportion of relevant material actually retrieved, and the proportion of retrieved material actually relevant. In an ideal system, it may be assumed that everything relevant to a user's query is in fact retrieved (thus producing a recall of 1) while everything not relevant is rejected (producing a precision of 1). In real life, conditions are not so perfect, and it is generally not possible to achieve at the same time both a high recall and a high precision.

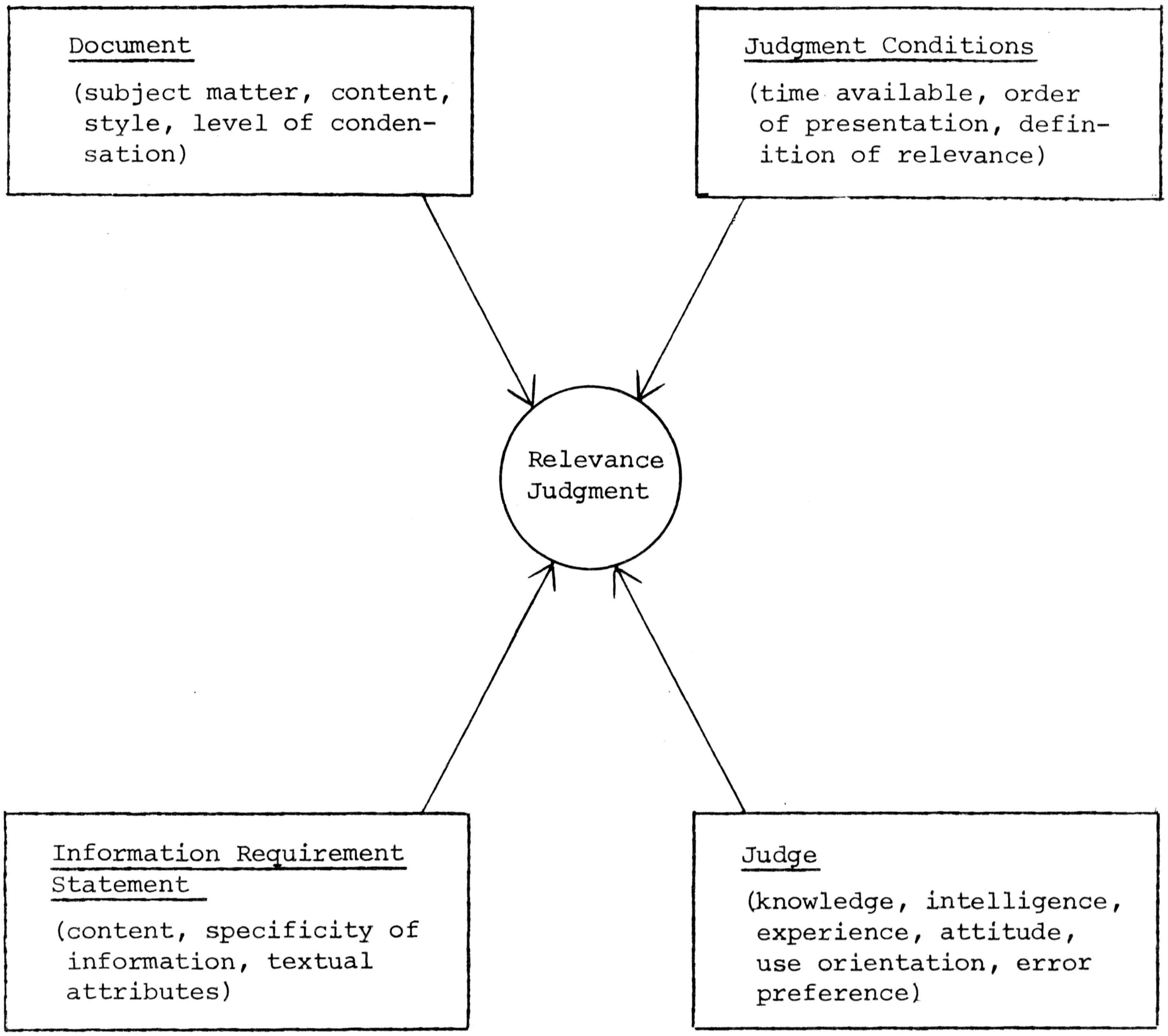
In order to generate recall and precision values, it is necessary first to differentiate retrieved from non-retrieved documents, and second to separate documents termed relevant to a query from those termed nonrelevant. The second of these partitions must obviously depend on a personal judgment either by the author of a given query, or by a system operator, or an outside expert. In any case, once a decision is reached about the relevance of each document to each query, it is possible by examining the set of retrieved and nonretrieved documents to compute unique recall and precision values. Unfortunately, relevance assessments tend to vary depending on who renders the judgment, and the recall and precision values obtained by using these assessments may then turn out to be inherently unstable. This question is further investigated in the remainder of this study.

2. The Relevance Problem

In a recent study of the relevance judging process, Cuadra and Katter recognize four main types of variables which potentially affect the outcome of a relevance judgment [4]: First, the type of document being judged, including its subject matter, level of difficulty, level of condensation, style, and so on; next the conditions under which the judgments must be rendered, that is the time available, the order of presentation and size of the document set, the type of task specification, and so on; then, the statement specifying the information requirement which determines relevance; and, finally, the type of judge used to render the judgments, that is, his experience, background, attitude, and so on. These variables are summarized in the chart of Fig. 1. Additional variables may enter into the process if the judgment to be rendered is not expressible as a simple yes/no decision.

Because of the obvious complexity of the judgment process, numerous authors have stated that stable relevance judgments cannot possibly be obtained from individual informants. Fairthorne, for example, has suggested that individual relevance judgments should be replaced by global judgments representing a consensus of ideas by several independent judges [5]. O'Connor and Doyle have pointed out that the expression of a user's information need can take many different forms, and that it is not possible in consequence simply to claim that "document A is relevant to query B" without appropriate qualifying statements [6,7,8]. Taube has drawn the conclusion that recall and precision are not concepts which can be properly defined or used in retrieval systems evaluation [9].

A number of studies have also been conducted to show that different sets of relevance judgments are actually obtained under different judgmental conditions. Thus, distinctions are made between "motivated" and "unmotivated" judges [10], and between judgments based on an examination of full compared with partial document excerpts [11]. Furthermore, in the two most extensive studies of the judgment process by Cuadra and Katter [4, 12] and Rees [13, 14] respectively, a


Document
(subject matter, content, style, level of condensation)
Judgment Conditions
(time available, order of presentation, definition of relevance)
Relevance Judgment
Information Requirement Statement
(content, specificity of information, textual attributes)
Judge
(knowledge, intelligence, experience, attitude, use orientation, error preference)

Variables Related to Relevance Judgments

Fig. 1

(adapted from Cuadra and Katter [4])

large number of factors are varied and the effect on the resulting relevance judgments is observed.

The conclusion is sometimes drawn from studies such as the preceding that the existing methodology in systems evaluation must be revised, and that evaluation results based on recall and precision are unreliable and must be viewed with great caution. Cuadra and Katter state in particular [4]:

> "the first and most obvious implication is that one cannot legitimately view 'precision' and 'recall' scores as precise and stable bases for comparison between systems or systems components, unless ... (appropriate controls are introduced)"

Rees voices similar misgivings in a somewhat different context [13]:

> "the lack of replication (that is experimental control permitting duplication of the experiments) of the results of either the SMART [3] or the Cranfield studies [1] must necessarily introduce a note of caution to the existence of 'rules' and generalizability of results".

While these sentiments appear at first to be perfectly justified, since the subjectiveness and variability of individual relevance judgments cannot obviously be contested, the jump which is necessary to reach the conclusion that recall and precision results are unreliable because relevance judgments are unstable has never been adequately proved or substantiated. Indeed, there exists some evidence that such a conclusion cannot be drawn from the available evidence. Giuliano and Jones, for example, made a small study using a panel of three relevance judges.

Their findings are summarized as follows [2]:

> "for purposes of comparing retrieval performance curves for two or more search options, it does not appear to matter much whether the curves are for any one of the single judges, or whether they are the averaged curves for a panel of three judges; the differences are primarily ones of scale, and the relative positions of the curves for the different search options tend to be the same in all cases".

Rees and Schultz also find that the judgmental groups used in their study agree substantially as to the relative positioning (i.e. ordering in decreasing order of relevance to a search request) of the documents, although the judges tend to assign to the documents different numerical ratings [14].

The experimental evidence cited above may indicate that, contrary to expectations, recall and precision results do not vary as widely as the relevance judgments used generate them. Several reasons can be cited further to support such an opinion:

a) recall and precision data are normally given as <u>averages</u> over many search requests; these averages may not be sensitive to small variations in the results for individual queries;

b) recall and precision data depend mainly on the relative positions of relevant and nonrelevant documents, when the documents are arranged in decreasing or increasing relevance order; individual changes in the composition of the relevant and nonrelevant document sets may have only a minor effect on the ordering of the sets as a whole;

c) disagreements among relevance judges may affect mostly the borderline cases, while preserving a general consensus for a large set of items difinitely termed either relevant or nonrelevant; such borderline cases normally receive a low position in the relevance ordering, and their effect on the recall and precision values may be expected to be negligible;

d) recall-precision results are often given as relative differences between sets of different search and retrieval methods; the recall and precision results may vary in such a way that differences between methods are preserved even though the values for the individual methods may change.

These questions are further examined in an experiment to be described in the remaining sections of this study.

3. The Experiment

The evaluation procedures incorporated into the SMART document retrieval system lend themselves to a pairwise comparison of the effectiveness of two or more processing methods. Specifically, a number of evaluation parameters are computed for each of the processing methods under investigation. A comparison of the corresponding measures for two or more methods can then be used to produce a ranking of the methods in decreasing order of retrieval effectiveness.

The following evaluation measures are generated by the SMART system for each processing run [3]:

a) a recall-precision graph reflecting the average precision value at ten discrete recall points (from a recall of 0.1 to a recall of 1.0 in intervals of 0.1);

b) two global measures, known as normalized recall and normalized precision, which together reflect the overall performance level of the system;

c) two simplified global measures, known as rank recall and log precision, respectively.

The experiments described in the present report were designed to determine the degree of sensitivity of the SMART evaluation output to variations in the relevance assessments and used to compute the evaluation measure. If the recall-precision output obtained by SMART turns out to be unstable because of the instability of the relevance judgments used, then an extrapolation of the results to other user populations and different retrieval environments may not be possible. On the other hand, if the evaluation output remains stable, then the significance of the results appears to be confirmed.

A collection of 1268 abstracts in the field of documentation and library science comprising about 131,500 English text words (the 'Ispra' collection) was used for experimental purposes. The collection includes most of the articles published in 1963 and 1964 in American Documentation and several other journals in the information retrieval area. Eight different persons were used to generate a total of 48 different search requests in the documentation field; each person was familiar with the library science field, either being a librarian himself or a student in library science, and each one was asked to produce six requests that might actually be asked by library science students. To aid in the query generation, a detailed and carefully drawn set of instructions was distributed to the group of query authors. The main criteria proposed for the query formulation are reproduced in Table I.

Each query was expected to represent a real information need, and had to be expressed in grammatically correct, and hopefully unambiguous English. As usual for queries processed by the SMART system, positive formulations were required, and the queries were to be generated independently from the document collection; in particular, no "source" document was to be used for the formulation of any of the queries.

Following receipt of the query formulations from each of the eight authors, the texts of the document abstracts comprising the collection were distributed, and each author was asked to assess the relevance of each document abstract with respect to each of _his_ six queries. Dichotomous relevance judgments were to be used, asserting either the relevance or the nonrelevance of each item for each query. Furthermore, the relevance criterion to be used was a strict one, in the sense that relevance of a document was to be specified only

> "if it is directly stated in the abstract as printed, or can be directly deduced from the printed abstract, that the document contains information on the topic asked for in the query".

Since each query presumably represented an information need, an abstract would thus be called relevant if the author felt that given the abstract he would with great probability wish to consult the complete document.

After receipt of the relevance judgments from each of the authors (the A judgments), a second, independent set of relevance judgments (the B judgments) was obtained by asking each person in the test group to judge for relevance six additional queries originated by six different people, not including himself. The same relevance criteria were used for the second relevance judgments as for the original ones, the only difference being that the A judgments were rendered by query authors, whereas the B judgments are nonauthor judgments. In order to preserve independence, the B judges were not informed of the A judgments previously obtained, nor was there any interaction between assessors either before or during the judging process.

For each of the 48 queries, a set of four different document sets thus became available, each consisting of the items termed relevant by a different set of people as follows:

A set: relevance assessed by query author;
B set: relevance assessed by outside subject expert;
C set: relevance asserted by either A or B assessor;
D set: relevance asserted by both A and B.

The situation is summarized in Table 2.

A measure of agreement in the relevance judgments can be obtained for the query set from the material of Table 3. For each query, the number of items is given for sets A and B, respectively, as well as the total number of distinct items (set C), and the total number of items common to both sets A and B (set D). Each query number listed in Table 3 is coded in such a way that the number ij is assigned to the query authored by person i, with the second (B) relevance judgment being obtained from person j.

The agreement among the relevance sets is measured as usual

Positive Criteria for Query Formulation	Negative Criteria for Query Formulation
1. Generate queries of real interest to a potential researcher or student	Avoid "exotic" topics and doubtful subject matter
2. Formulate queries in clear, coherent, properly punctuated, grammatically correct sentences	Avoid metaphors, jokes, and allusions
3. Use from 50 to 100 words and up to 3 sentences to formulate queries	Do not submit queries corresponding to the contents of a specific document; do not rephrase specific document contents
4. Use positive formulations stating what subject areas are actually wanted	Avoid negative formulations and clauses introduced by "except", "other than", "not", etc.
5. Use homogeneous query formulations representing a single topic	
6. Use only common abbreviations	

Principal Criteria for Query Formulation

Table 1

Group of Judges	Explanation
A	Original group of query authors. Each person in A group made relevance judgments for his six queries
B	Nonauthor judges. Each person in B group made relevance judgments for six queries corresponding to six different authors from A group
C	Document is relevant to a given query if either the A judge or the B judge termed it relevant
D	Document is relevant to a given query if both A and B judges termed it relevant

Relevance Judgment Groupings

Table 2

by dividing the total number of common items by the total number of distinct items $(A \cap B)/(A \cup B)$. The numeric values are given in column 6 of Table 3. An average agreement score is given for each author in column 7 of Table 3. This score is seen to vary from a high of 0.53 for author 6 to a low of 0.11 for author 8. The overall agreement for all 8 authors is seen to be slightly higher than thirty percent (0.3074). This figure is believed to be typical of the consistency obtainable under independent conditions from different assessors. It also agrees with comparable figures contained elsewhere in the literature.

It remains to show how such a relatively low consistency level is reflected in the evaluation output. This is described further in the next section.

4. Experimental Results

Three of the principal automatic language analysis procedures incorporated into the SMART system are used with the Ispra collection under study. The methods known as word form, word stem, and thesaurus, respectively, may be described as follows:

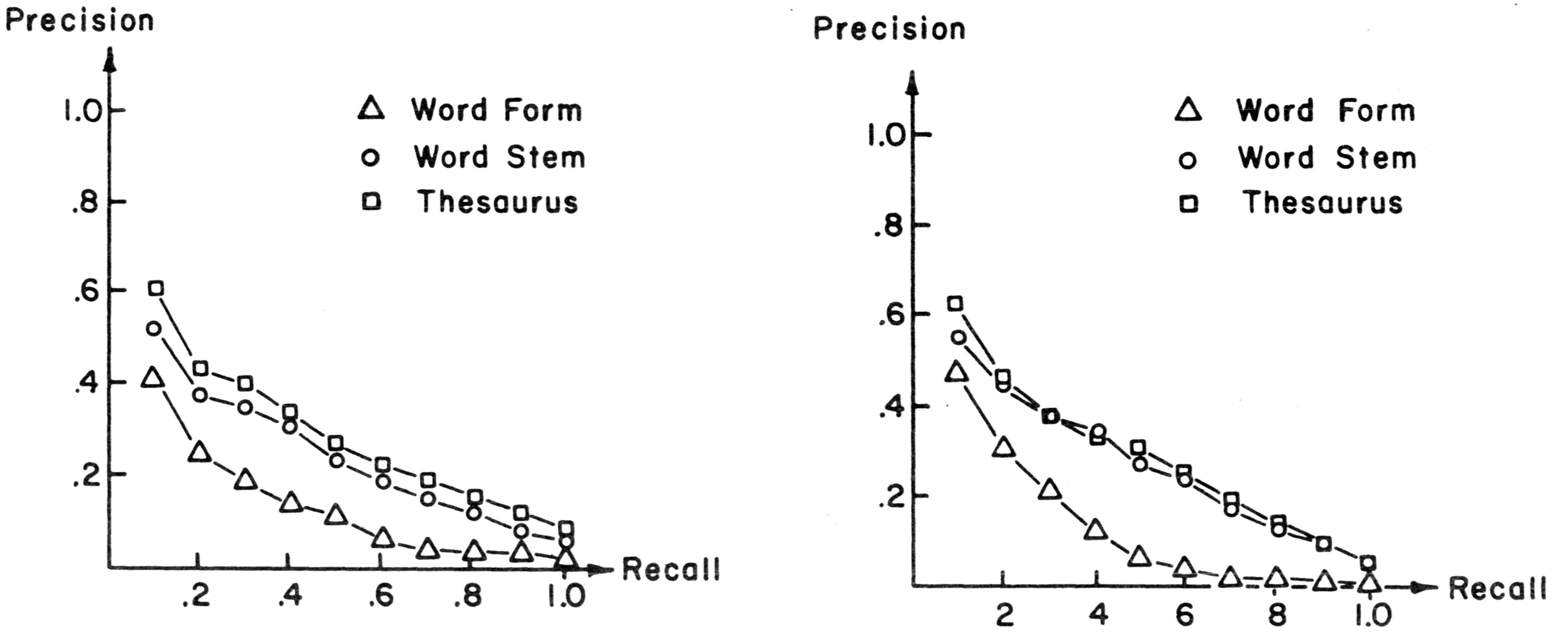
a) word form: texts of document abstracts and queries are reduced by removal of common words and final 's' endings, and weights are assigned to the remaining word forms; the reduced texts are then matched to obtain document-query correlation coefficients;

b) word stem: texts are treated as above, except that com-

Query Number	Number of Relevant				Agreement	Average Agreement with Author
First Judge Second Judge	A	B	C (A∪B)	D (A∩B)	$\frac{A\cap B}{A\cup B} = \frac{D}{C}$	
12	17	18	26	9	0.3462	Author 1 0.3611
13	3	4	4	3	0.7500	
14	19	6	21	4	0.1905	
15	22	25	37	10	0.2703	
16	7	9	14	2	0.1429	
17	9	13	15	7	0.4667	
21	20	25	37	8	0.2162	Author 2 0.1757
23	32	8	39	1	0.0256	
25	20	7	25	2	0.0800	
26	19	8	23	4	0.1739	
27	14	17	23	8	0.3478	
28	7	16	19	4	0.2105	
31	4	3	5	2	0.4000	Author 3 0.1838
32	6	18	22	2	0.0909	
34	27	20	45	2	0.0444	
35	27	27	35	19	0.5429	
36	34	8	41	1	0.0244	
37	5	6	11	0	0	
41	14	8	17	5	0.2941	Author 4 0.4298
42	19	20	26	13	0.5000	
45	8	41	42	7	0.1667	
46	10	8	12	6	0.5000	
47	8	10	10	8	0.8000	
48	25	33	44	14	0.3182	
51	10	11	16	5	0.3125	Author 5 0.2167
52	72	16	78	10	0.1282	
53	31	14	42	3	0.0714	
54	13	9	19	3	0.1579	
56	34	25	47	12	0.2553	
58	33	22	40	15	0.3750	
61	6	10	11	5	0.4545	Author 6 0.5297
63	23	49	61	11	0.1803	
64	12	12	17	7	0.4118	
65	7	6	9	4	0.4444	
67	11	10	11	10	0.9091	
68	7	9	9	7	0.7777	
71	14	28	31	11	0.3548	Author 7 0.4557
72	9	14	15	8	0.5333	
73	19	11	23	7	0.3043	
74	9	22	24	7	0.2917	
75	6	5	6	5	0.8333	
78	12	22	24	10	0.4167	
81	22	6	24	4	0.1667	Author 8 0.1067
82	37	32	54	15	0.2778	
83	34	4	35	3	0.0857	
84	21	10	28	3	0.1071	
86	18	0	18	0	0	
87	17	8	25	0	0	
Totals	853	713	1260	306	Overall Average	0.3074

Agreement of Relevance Judgments

Table 3

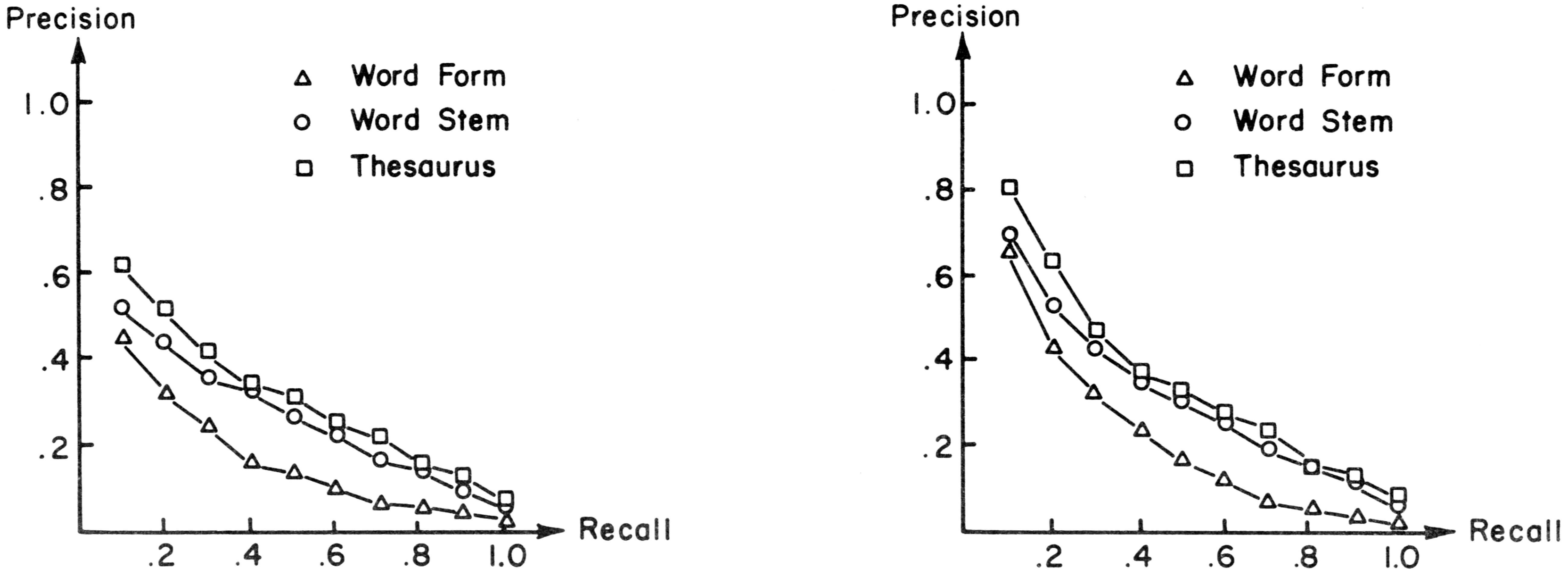


a) A Judgments

b) B Judgments

Recall-Precision Curves for Ispra Abstracts
(averages for 48 queries, 1268 documents)
(cosine correlation, numeric vectors)

Fig. 2



c) C Judgments

d) D Judgments

Recall-Precision Curves for Ispra Abstracts
(averages for 48 queries, 1268 documents)
(cosine correlation, numeric vectors)

Fig. 2
(contd)

Recall	Average Precision Values					
	A	B	A-B	C	D	D-C
0.1	.405	.467	-.062	.448	.664	.216
0.3	.196	.206	-.010	.235	.322	.087
0.5	.102	.081	+.021	.128	.165	.037
0.7	.039	.028	+.011	.058	.070	.012
0.9	.023	.018	+.005	.029	.023	-.006

a) Word Form Process

Recall	Average Precision Values					
	A	B	A-B	C	D	D-C
0.1	.514	.524	-.010	.503	.693	.190
0.3	.363	.375	-.012	.376	.434	.058
0.5	.243	.283	-.040	.266	.308	.042
0.7	.162	.182	-.020	.167	.196	.029
0.9	.095	.093	+.002	.056	.111	.055

b) Word Stem Process

Recall	Average Precision Values					
	A	B	A-B	C	D	D-C
0.1	.612	.627	-.015	.604	.801	.197
0.3	.406	.398	+.008	.433	.485	.052
0.5	.293	.307	-.014	.315	.319	.004
0.7	.204	.199	+.005	.213	.227	.014
0.9	.112	.106	+.006	.113	.119	.006

c) Thesaurus Dictionary

Average Standard Recall and Precision Values
for 3 Analysis Methods and 4 Types of Relevance Judgments

Table 4

plete suffixes are removed from text words to reduce the texts to weighted word stems; the query-document matching process remains the same as in process a);

c) thesaurus: each word stem produced by procedure b) is looked up in a thesaurus providing synonym recognition, and the resulting weighted concept identifiers assigned to queries and documents are compared (instead of word forms or word stems).

The output produced by the SMART system consists of superimposed recall-precision graphs exhibiting averages over a complete query set for several processing methods [3]. The method which generates the curve closest to the upper right-hand corner of the graph (where recall and precision are equal to 1) exhibits the best performance. Under normal circumstances, an evaluation of performance for a variety of processing methods does not require a detailed comparison of the actual recall and precision values, but only an examination of the ranking of the corresponding recall-precision curves. Thus, to show that the performance measures are insensitive to changes in the relevance judgments, it is sufficient to observe a consistent ranking of the recall-precision graphs obtained from the several processing methods. The data for the four types of relevance judgments (A, B, C, and D) are shown averaged over the 48 queries in Fig. 2.

The following conclusions may be drawn from the output of Fig. 2:

a) all four sets of relevance judgments produce the same ranking of the processing methods; in particular, the word form process is always much less powerful than the other two procedures, and the thesaurus process is slightly better than the word stem match;

b) the main difference in the output produced by the A and B judgments is the somewhat closer agreement between word stem and thesaurus runs for the B judgments than for A;

c) the best results in terms of recall and precision are obtained for the D judgments which represent the agreement between both A and B relevance judges; for low recall, the precision is about 20 percent higher for D than for A, B, or C.

While it is clear from the output of Fig. 2 that the SMART evaluation output does not vary with variations in the relevance judgments, it may be of interest to examine the data in somewhat more detail. Table 4 contains the numeric values corresponding to the curves of Fig. 2. The average precision is given for each of five recall points for the four curves of Fig. 2. In addition, the numeric precision difference is given at these same recall points between the A and B curves (in column 4 of Table 4), and between the C and D curves (in column 7 of the table). It may be seen that the maximum difference between the averaged A and B output occurs for the word form process at very low recall (precision difference of 0.06). The normal precision difference for the two sets of relevance judgments is about 1 to 2 percent. For the thesaurus run, which exhibits the best performance, the maximum precision difference is only 0.015 at low recall, with a normal difference of less than one percent.

From the output of Table 4 and Fig. 2, it appears reasonable to conclude not only that the performance methods are ranked in the

		Query Numbers								
		12-17	21-28	31-37	41-48	51-58	61-68	71-78	81-87	All
Word Form	A	.011	.023	.016	.028	.034	.013	.021	.018	.021
	B	.013	.014	.015	.026	.016	.019	.023	.010	.017
	A-B	-.002	.009	.001	.002	.018	-.006	-.002	.008	.004
Word Stem	A	.077	.056	.087	.162	.103	.271	.063	.038	.108
	B	.085	.037	.115	.165	.057	.256	.065	.048	.104
	A-B	-.008	.019	-.028	-.003	.045	.015	-.002	-.010	.004
Thesaurus	A	.133	.078	.091	.254	.125	.156	.134	.067	.120
	B	.122	.079	.114	.236	.052	.108	.162	.074	.118
	A-B	.011	-.001	-.013	.018	.073	.048	-.028	-.007	.002

a) Average Rank Recall Differences (3 Analysis Methods)

		Query Numbers								
		12-17	21-28	31-37	41-48	51-58	61-68	71-78	81-87	All
Word Form	A	.172	.243	.279	.481	.342	.363	.437	.215	.209
	B	.289	.228	.176	.406	.258	.499	.307	.236	.300
	A-B	-.117	.015	.103	.075	.084	-.136	.130	-.021	-.091
Word Stem	A	.511	.416	.358	.659	.385	.760	.663	.256	.469
	B	.489	.432	.524	.632	.432	.768	.566	.299	.518
	A-B	.022	-.016	-.166	.027	-.047	-.008	.097	-.043	-.049
Thesaurus	A	.673	.507	.391	.717	.472	.739	.764	.366	.557
	B	.641	.562	.504	.659	.445	.655	.720	.533	.590
	A-B	.032	-.055	-.113	.058	.027	.084	.044	-.167	-.033

b) Average Normalized Precision Differences (3 Analysis Methods)

Average Rank Recall and Normalized Precision for Each of 8 Query Authors

Table 5

same order, no matter which of the four sets of relevance judgments is used, but also that the actual performance differences resulting from differences between author and nonauthor judgments are negligible. This point can also be made by looking at the individual performance differences for each of the 8 query authors as shown in Table 5.

Table 5 exhibits the average rank recall (in Table 5(a)) and average normalized precision (Table 5(b)) for the six queries originated by each of the eight authors. In each case, the average obtained by using the author relevance judgments is shown (case A), followed by the average for the same six queries using the nonauthor judgments, followed finally by the difference of the measures between A and B. It may be seen once again that the processing methods are ranked identically by 7 out of 8 authors from the best method (thesaurus) to the worst (word form). Only for the queries of author 6 (nos. 61-68) does the word stem process produce slightly better results than the thesaurus method; however, the word form process is inferior even for that author. When the B relevance judgments are used, the same ranking is again obtained for 7 out of 8 query sets. For queries 61-68, the word stem process is again superior to the thesaurus, while for queries 31-37 and 51-58, the B judgments produce approximately equal performance for word stem and thesaurus. The differences in rank recall and normalized precision obtained for the two sets of relevance judgments (A and B) are shown in row 3 of Table 5 for each dictionary. The differences are again exceedingly small.

In the next section, performance results are given for individual queries, and an attempt is made to explain why the relatively large differences in the relevance judgments do not lead to substantial differences in the performance parameters.

5. Judgment Consistency and Performance Measures

In order to explain why the average recall and precision data previously exhibited are relatively insensitive to differences in the relevance assessments, it is necessary to look at the performance characteristics for some individual queries.

Consider first the data of Table 6 giving normalized recall and normalized precision figures averaged over the 48 queries for the four sets of relevance assessments. It may be seen that with the sole exception of the word form normalized recall, the highest performance is obtained in each case using the D judgments followed by B, A, and finally C. The D judgments, however, represent the agreement in the relevance assessments between authors and nonauthors, and the corresponding relevance sets are therefore produced under reasonably stringent conditions (at least two independent people must agree before an item is termed relevant). On the other hand, the C judgments are produced by relatively free criteria, since an item is called relevant if either one of two independent judges calls it relevant. It appears then from the output of Table 6, that the D judgments which are designed to select those documents most certainly relevant to each query, also select those documents most efficiently retrieved by the computer system. That is, the query-document pairs which are most closely and unarguably related are exactly the pairs on which the retrieval performance is best.

This result is confirmed by the output of Table 7, where those

Group	Normalized Recall			Normalized Precision		
	Word Form	Stem	Thes	Word Form	Stem	Thes
A	0.5289	0.8340	0.8858	0.2090	0.4690	0.5570
B	0.5249	0.8452	0.8904	0.2473	0.5104	0.5850
C	0.6234	0.8249	0.8777	0.1804	0.3655	0.4885
D	0.5493	0.8759	0.9212	0.3492	0.5777	0.6403

Normalized Recall and Precision Averages
(3 Analysis Methods and 4 Types of Relevance Judgments)
(Cosine Correlation, Numeric Vectors)

Table 6

		Performance Measure					
		Top 12		Middle 24		Bottom 12	
		B better	A better	B better	A better	B better	A better
Relevance Judgment Consistency	Top 12	4	5	1	2	0	0
	Middle 24	1	2	6	11	4	0
	Bottom 12	0	0	0	4	5	3

Correlation of Judgment Consistency with Performance

Table 7

Performance Measure: $$\frac{(\text{NP for A})+(\text{NP for B})+(\text{NR for A})+(\text{NR for B})}{4}$$

Relevance Judgment Consistency: $$\frac{\text{No. of relevant in D}}{\sqrt{(\text{No. of relevant in A})(\text{No. relevant in B})}}$$

queries are selected which exhibit the best agreement between the relevance assessors. Such queries represent relatively unambiguous, closely related query-document sets. It is found, just as with the D judgments, that these closely related pairs produce the best retrieval performance.

Table 7 contains a plot of performance effectiveness versus consistency in the relevance judgments. Specifically, a single performance measure and a single measure of relevance consistency are computed for each query as follows:

$$\text{performance measure} = \frac{(\text{NP for A}) + (\text{NP for B}) + (\text{NR for A}) + (\text{NR for B})}{4}$$

and

$$\text{judgment consistency} = \frac{\text{Number of relevant items in D}}{\sqrt{(\text{No. of relevant in A})(\text{No. of relevant in B})}} \quad ,$$

where NP and NR are normalized precision and normalized recall, respectively. The 48 queries are then arranged into three groups for performance (the top 12, the middle 24 and the bottom 12), and into three groups for relevance judgment consistency. Table 7 shows how relevance consistency correlates with performance.

It may be seen that performance is best for those queries with the best relevance consistency. Indeed, 9 of the 12 queries in each top group are also in the other top group. Contrariwise, not a single query from the bottom 12 in judgment consistency is in the top 12 for performance, and vice versa, not a single query from the bottom 12 in performance is in the top 12 for judgment consistency.

The performance indicators of Table 7 are further subdivided into queries for which the B judgments provide the better performance and those for which the A judgments are superior. In the former case, the nonauthor judgments proved more useful than the author judgments, indicating possibly that these queries are ambiguous or poorly formulated. It may be seen from the table that this is the case for a total of 5+7+9 = 21 queries out of 48, of which 9 are ranked in the bottom 12 for performance.

It is now possible to explain why the recall-precision output is basically invariant for the collection under study, even though the agreement among relevance judgments is relatively low:

a) on the one hand, the performance is best for those queries with the best consistency in the relevance judgments;

b) on the other hand, the recall and precision measures are most sensitive to documents (both relevant and irrelevant) retrieved early in the search, that is, documents with low rank.

The conclusion is then obvious that although there may be a considerable difference in the document sets termed relevant by different judges, <u>there is in fact a considerable amount of agreement for those documents which appear most similar to the queries and which are retrieved early in the search process</u>, (assuming retrieval is in decreasing correlation order with the queries). Since it is precisely these documents which largely determine retrieval performance, it is not surprising to find that the evaluation output is substantially invariant for the different sets of relevance judgments.

The situation is illustrated by a typical query (number 12) in Fig. 3. The first row of Table 3 shows that for this particular query, the number of relevant items identified by A was 17, while the B judge identified 18 relevant documents. The total number of distinct relevant items was 26 of which 9 were chosen in common by the A and B judges. The agreement score is 0.3462. The ranks of all 26 relevant documents are given in Fig. 3(a), with the common items being shown underlined. It may be seen that of the 8 relevant items with the lowest rank (from rank 1 to rank 25) there was agreement between the judges for 6 items; on the other hand, of the 8 relevant items retrieved with highest rank (ranks 178 to 832) there was not a single agreement between the A and B judges. The two recall-precision graphs for query 12 are shown in Fig. 3(b); they are seen to be remarkably similar, reflecting the fact that for the top 25 documents retrieved, the differences in relevance judgments between the A and B judges are very small indeed.

In conclusion, it can be stated that, if the relevance assessments obtained from the query authors used in the present study are typical of what can be expected from general user populations of retrieval systems, then the resulting average recall-precision figures appear to be stable indicators of system performance which do in fact reflect actual retrieval effectiveness.

6. Machine Search Effectiveness

It has been said elsewhere [17] that the retrieval effectiveness obtained with the automatic text processing methods incorporated into the

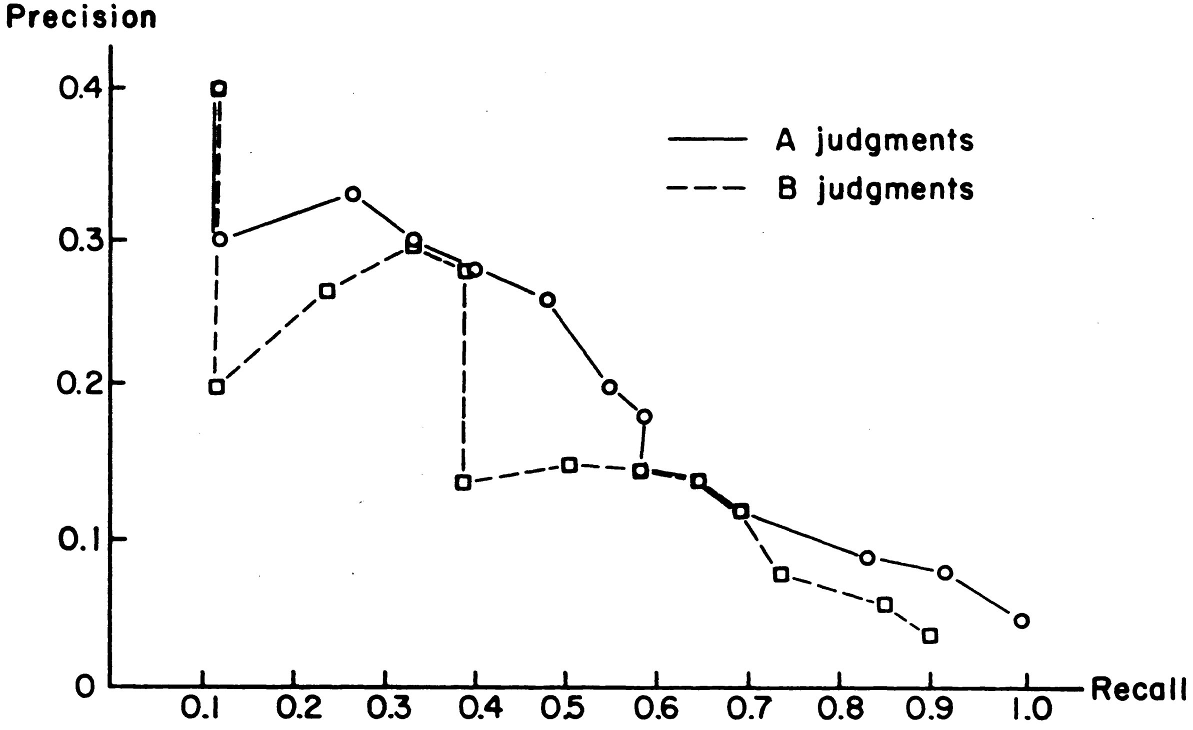
SMART system appears to be roughly equivalent to the effectiveness obtainable with presently operating manual, or semi-automatic retrieval systems. It may be of interest to ask how this presently achievable performance compares with the performance of the best possible imaginable delegated search system. The differences between present performance and such an optimum delegated search system may then give an indication of the amount of improvement in performance which may eventually result from future developments.

It is not completely unreasonable to assume that the best possible delegated search system is one where a subject expert completely reads through an entire document collection and ranks each document in decreasing similarity order with a given search query. Such a system, which for obvious reasons is not operationally implementable, should in theory be superior to any search system based on indexing or on other reduced document representation. The set of B searchers used in the present experiment can then be assumed to constitute such an ideal search system, since they in fact were asked to search through the complete document collection for each query.

A comparison has been made between the amount of material "retrieved" by the B searcher (that is the number of documents termed relevant by B), and the number of relevant items retrieved by the machine search using the same cut-off as the B searcher. In both cases, relevance is determined by using the author (A) judgments as criteria. Specifically, "optimum recall" and "optimum precision" figures are computed for each query by evaluating the performance of the B searcher (in comparison with the A relevance judgments) as follows:

RANK	1	2	9	11	15	18	20	25	30	44	54	60	61	69	72	88	105	133	178	179	196	199	277	298	322	832
A	x	x	x	x	x		x	x	x	x	x			x		x	x	x	x		x			x		
B	x	x		x	x	x	x	x			x	x	x	x	x			x		x		x	x		x	x
A∩B	x	x		x	x		x	x			x			x				x								

a) Ranks of Relevant Documents
for Query 12
(agreement score 0.3462)



b) Recall Precision Graph for Query 12
(Cutoff after 5, 10, 15, 20, 25, etc. retrieved items)
Sample Performance (Query 12)

Fig. 3

$$\text{optimum (B) recall} = \frac{\text{no. of relevant in } (A \cap B)}{\text{no. of relevant in } A}$$

$$\text{optimum (B) precision} = \frac{\text{no. of relevant in } (A \cap B)}{\text{no. of relevant in } B}$$

These optimum recall and precision figures are then compared with the machine performance, using the thesaurus dictionary for analysis purposes. To permit a fair comparison, the number of items retrieved in the machine search must be the same as the number of items "retrieved" by the B searcher; the cutoff in the machine search is therefore set at the number of relevant items identified by B. The machine recall and precision are defined as follows:

$$\text{machine recall} = \frac{\text{no. of relevant in A retrieved before cutoff}}{\text{no. of relevant in A}}$$

$$\text{machine precision} = \frac{\text{no. of relevant in A retrieved before cutoff}}{\text{total no. of relevant in B (cutoff)}}$$

The output of Table 8 shows that the overall machine search results are about 25 percent lower on the average than the 'B' results. For some query sets (for example 41-48, and 71-78), the results are approximately equivalent, and for five queries out of the set of 48, the machine performance is in fact better than that of the B searcher. It is seen in Table 8 that the average recall and precision for the B searcher are about 0.46, whereas the comparable machine figures are 0.32. These figures once again demonstrate that the improvements obtainable by refinements in the search and analysis techniques (from 0.32 to 0.46) are relatively modest, in comparison with the desirable perfect system where recall and precision are close to 1. The gap between the complete human search and a perfect search (from 0.46 to 1) appears to be due to ambiguities inherent in the query formulation process and to the difficulties of reconciling the user's view of a subject area with the subject analysis provided for a given document collection. This latter gap may well never be bridged by any search and retrieval system likely to come into existence in the foreseeable future.

To summarize, several sets of relevance judgments are used in the present study in conjunction with a document collection of over 1200 items in library science and documentation. The retrieval results in terms of recall and precision obtained for the various relevance sets are substantially identical, even though the overall agreement among the relevance assessments is only about thirty percent. This fact is explained and the conclusion is drawn that there appears to be no reason to reject previously published evaluation results for manual or automatic searches, because of uncertainties and instabilities in the computation of the performance measures. It is pointed out again that the absolute performance achievable under present conditions, or likely to be achieved in the future, is much lower than the theoretically desirable optimum.

	12-17	21-28	31-37	41-48	51-58	61-68	71-78	81-87	All
Optimum Search	0.51	0.50	0.35	0.34	0.59	0.25	0.54	0.68	0.46
Machine Search (Thesaurus)	0.36	0.19	0.17	0.43	0.16	0.40	0.48	0.11	0.32
Difference	0.15	0.31	0.18	-0.09	0.43	-0.15	0.06	0.57	0.14

a) Average Optimum and Machine Recall

	12-17	21-28	31-37	41-48	51-58	61-68	71-78	81-87	All
Optimum Search	0.52	0.49	0.45	0.38	0.54	0.35	0.47	0.52	0.46
Machine Search (Thesaurus)	0.40	0.24	0.17	0.33	0.23	0.31	0.35	0.27	0.32
Difference	0.12	0.25	0.28	0.05	0.31	0.04	0.12	0.25	0.14

b) Average Optimum and Machine Precision

Optimum (B) Recall	$\frac{\text{Relevant in D}}{\text{Relevant in A}}$
Optimum (B) Precision	$\frac{\text{Relevant in D}}{\text{Relevant in B}}$
Machine Recall	$\frac{\text{Relevant in A before cutoff}}{\text{Relevant in A}}$
Machine Precision	$\frac{\text{Relevant in A before cutoff}}{\text{Relevant in B (cutoff)}}$

Comparison of Optimum with Machine (Thesaurus) Performance

Table 8

References

[1] C. W. Cleverdon, J. Mills, and E. M. Keen, Factors Determining the Performance of Indexing Systems: Vol. 2, Test Results, Cranfield, England 1966.

[2] V. E. Giuliano and P. E. Jones, Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, Report ESD-TR-66-405, A. D. Little, Inc., Cambridge, August 1966.

[3] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1968, pp. 8-36.

[4] C. A. Cuadra and R. V. Katter, Experimental Studies of Relevance Judgments: Final Report, Report TM-3520, Vol. 1: Project Summary, Vol. 2: Description of Individual Studies, System Development Corp., Santa Monica, June 30, 1967.

[5] R. A. Fairthorne, Implications of Test Procedures, in Information Retrieval in Action, Western Reserve University Press, Cleveland, 1963, pp. 109-113.

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

[7] J. O'Connor, Some Questions Concerning Information Need, American Documentation, Vol. 19, No. 2, April 1968.

[8] L. B. Doyle, Is Relevance an Adequate Criterion in Retrieval System Evaluation, Proceedings of the 26th Annual Meeting, American Documentation Institute, Chicago, October 1963.

[9] M. Taube, A Note on the Pseudomathematics of Relevance, American Documentation, Vol. 16, No. 2, April 1965, pp. 69-72.

[10] E. D. Dym, Relevance Predictability - Investigation, Background and Procedures, in Electronic Handling of Information: Testing and Evaluation, A. Kent, et.al. editors, Thompson Book Co., Washington, p. 175-185.

[11] D. L. Shirey and M. Kurfeerst, Relevance Predictability - Data Reduction, in Electronic Handling of Information: Testing and Evaluation, A. Kent et.al. editors, Thompson Book Co., Washington, pp. 187-198.

References (contd)

[12] C. A Cuadra and R. V. Katter, Opening the Black Box of Relevance, Journal of Documentation, Vol. 23, No. 4, December 1967.

[13] A. M. Rees, Evaluation of Information Systems and Services, in Annual Review of Information Science and Technology, Vol. 2, C. Cuadra, editor, Interscience Publishers, New York, 1967.

[14] A. M. Rees and D. G. Schultz, A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching, Final Report to the National Science Foundation, Center for Documentation and Communication Research, Case Western Reserve University, October 1967.

[15] G. Salton, et. al., Scientific Reports on the SMART System to the National Science Foundation, Nos. ISR-11, ISR-12, ISR-13, Department of Computer Science, Cornell University, Ithaca, New York, June 1966, June 1967, and January 1968.

[16] G. Salton, Search and Retrieval Experiments in Real-Time Information Retrieval, Proceedings IFIP Congress '68, Edinburgh, August 1968.

[17] G. Salton, A Comparison Between Manual and Automatic Indexing Methods, Technical Report No. 68-11, Computer Science Department, Cornell University, March 1968.

Appendix

It is shown in this appendix that the ranking in decreasing order of performance for several processing methods stays constant under conditions of considerable generality, assuming that the performance order is defined by the usual recall and precision measurements.

A perfect relevance judge can be characterized by the fact that he will call a relevant document in fact "relevant" with a probability equal to 1.0, while calling a nonrelevant item "relevant" with a probability equal to 0.0. A somewhat slipshod judge who makes random errors in judgment can be characterized by probabilities p_r and p_{nr}, where p_r is the probability that he will call a relevant document "relevant", and p_{nr} is the probability that he will call a nonrelevant document "relevant". As p_r decreases from 1.0, and p_{nr} increases from 0.0, the judge becomes increasingly inaccurate.

Consider a retrieval system, operating with procedure a, at a recall R_a and precision P_a, measured by the perfect judge. Call the number of documents retrieved n, the total number of documents in the collection N, and the total number of relevant items G. The performance of this system can be evaluated using judgments made by a slipshod judge. The total number of relevant retrieved in $P_a \cdot n$; of these $P_a \cdot n \cdot p_r$ are called relevant. Furthermore, $(1-P_a) \cdot n$ nonrelevant are retrieved; of these $p_{nr} \cdot (1-P_a) \cdot n$ are called relevant. The apparent precision using the slipshod judge for evaluation is therefore:

$$P_a' = \frac{(P_a n p_r + (1-P_a) n p_{nr})}{n} = P_a(p_r - p_{nr}) + p_{nr}.$$

Similarly, the apparent recall turns out to be

$$R'_a = \frac{(R_a G(p_r - p_{nr}) + np_{nr})}{G(p_r - p_{nr}) + Np_{nr}} \quad .$$

It is obvious that if $p_r > p_{nr}$, P'_a is monotonically increasing with P_a; also R'_a is linearly increasing with R_a. That is, any judge with $p_r > p_{nr}$ ranks retrieval methods in the same order of performance as does the ideal judge; and any two inaccurate judges for whom this criterion is satisfied therefore produce identical performance rankings. Furthermore, if $p_{nr} << p_r$, the above transformations are equivalent to changes of scale, and even percentage changes in the measures will be accurately reproduced. In practice, p_{nr} is expected to be very small, so that this condition is likely to be fulfilled.

Of course, the statistical expectations derived here, do not imply that random variations may not produce individual queries for which an inaccurate order is apparently indicated. If a reasonable number of queries is averaged, however, the above results are expected to hold.

It should be noted that the constraints placed on the relevance judgments are very weak: it is necessary only that a judge be more likely to label relevant material as "relevant" than he is to label nonrelevant material as "relevant". It does not matter if only a small fraction of the relevant material is identified as relevant, so long as less than that fraction of the nonrelevant material is called relevant. It does not matter either if the total number of documents called relevant is greater than the true number relevant, less than the true number, or equal to it;

nor does it matter if the majority of the documents labeled relevant are actually nonrelevant. Because of the weakness of the constraints, it is therefore most unlikely that any misranking of the performance of retrieval methods can result from inferior relevance judgments.