Information Processing & Management Vol 28, No. 5, pp. 619–627, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + 00

VARIATIONS IN RELEVANCE JUDGMENTS AND THE EVALUATION OF RETRIEVAL PERFORMANCE

ROBERT BURGIN
School of Library and Information Sciences, North Carolina Central University, Durham, NC 27707, U.S.A.

(*Received* 19 *June* 1991; *accepted in final form* 21 *January* 1992)

Abstract—The relevance judgments used to evaluate the performance of information retrieval systems are known to vary among judges and to vary under certain conditions extraneous to the relevance relationship between queries and documents. The study reported here investigated the degree to which variations in relevance judgments affect the evaluation of retrieval performance. Four sets of relevance judgments were used to test the retrieval effectiveness of six document representations. In no case was there a noticeable or material difference in retrieval performance due to variations in relevance judgment. Additionally, for each set of relevance judgments, the relative performance of the six different document representations was the same. Reasons why variations in relevance judgments may not affect recall and precision results were examined in further detail.

INTRODUCTION

Standard methods of evaluating the performance of information retrieval systems make use of relevance judgments. Typically, in retrieval experiments each document in the test collection is evaluated by a judge or set of judges as relevant or not relevant to each query in the test set. Recall and precision measures can then be computed for the set of documents retrieved by various strategies, and these strategies can thereby be compared.

However, such relevance judgments are known to vary among judges. Lesk and Salton (1968) found only 30% agreement in relevance judgments between the authors of queries and non-author judges. Others (Gull 1956; Barhydt, 1964; Rees, 1967; Rees & Schultz, 1967; O'Connor, 1969; Kazhdan 1979) cite higher degrees of agreement among assessments of document relevance—as high as 85%—but acknowledge that disagreement among judges is high enough to be of concern for the evaluation of retrieval performance.

Researchers have also suggested that relevance judgments may be sensitive to a number of factors other than the relevance relationship between queries and documents, and may therefore vary under certain conditions extraneous to that relationship. Rorvig (1988) and Saracevic (1970, 1976), in particular, provide reviews of the research in this area.

For example, relevance assessments have been found to vary with the type of document representation—full text, abstract, title—used to judge relevance (Resnick, 1961; Rath, Resnick, & Savage, 1961; Rees & Schultz, 1967; Saracevic, 1969), with stylistic characteristics of documents and queries (Cuadra & Katter, 1967a), with time and stringency pressures (Cuadra & Katter, 1967a), with rating scales used (Cuadra & Katter, 1967a,b; Rees & Schultz, 1967; Eisenberg, 1988), and with the definition of relevance or instructions used in a given experiment (Cuadra & Katter, 1967a,b; Rees & Schultz, 1967). Regazzi (1988) found that the characteristics of judges—whether judges were researchers vs. students, whether judges were senior vs. junior researchers or senior vs. junior students, whether judges had a biomedical vs. a social science specialization—explained much of the variation in relevance assessments. Rees and Schultz (1967) also found that variations in relevance judgments could be linked to the amount of academic training in a subject area as well as the practical experience of the judge. Eisenberg (1988) and Eisenberg and Barry (1986, 1988) found a presentation order effect; significantly different category ratings of relevance were assigned to documents depending on the order in which they were presented. Eisenberg (1988) also found a context effect; the relevance judgments assigned to a particular document appeared to be affected by the other documents in the set being judged.

O'Connor (1969) found that relevance judgments changed when judges discussed discrepancies in judgments.

THE EFFECTS ON EVALUATION

However, while a number of studies have found that relevance judgments vary, few have examined the degree to which the variation of such judgments actually affects the evaluation of performance in retrieval systems that make use of the judgments.

Lesk and Salton (1968) investigated the degree to which recall and precision measures vary with deviations in relevance assessments. A collection of 1268 abstracts in library and information science was used, and eight individuals generated a total of 48 queries. The authors of the queries assessed the relevance of the abstracts to each of their queries; in addition, each individual provided relevance judgments for six queries that he did not author. Lesk and Salton were thus able to compare four sets of relevance judgments: those made by the author of a query; those made by someone other than the author; those formed by documents judged relevant by *either* the author or the non-author judge; and those formed by documents judged relevant by *both* the author and the non-author judge.

In spite of a low level of agreement among the authors and the non-authors (30%), performance differences based on the different sets of relevance judgments were negligible. In addition, regardless of the set of relevance judgments used to evaluate retrieval results, the relative ranking of three different document representations (word form, word stem, thesaurus) remained the same. The authors concluded 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" (Lesk & Salton, 1968, pp. 355–356).

It may be noted here that Lesk and Salton provide both a strong and a weak hypothesis regarding the effects of variations in relevance judgments on the evaluation of retrieval performance. The strong hypothesis states that no non-negligible differences in retrieval performance can be attributed to differences in relevance judgments. The weak hypothesis states that no variations in the relative performance of different processing methods can be attributed to differences in relevance judgments. Lesk and Salton found evidence to support both strong and weak hypotheses.

Kazhdan (1979) examined a test collection of 2600 documents in electrical engineering and 60 queries and compared the results obtained from seven different representations—abstract only, title only, title plus the first sentence from the abstract, and so on. Two sets of relevance judgments were obtained: one from a group of 13 experts and one from a single expert. The author found that variations in relevance judgments did produce some non-negligible differences in the evaluation of retrieval results but that the relative ranking of the seven different document representations remained the same, regardless of the deviations in relevance assessments.

In other words, Kazhdan found evidence to support only the weak hypothesis of Lesk and Salton (1968). The stronger hypothesis, that no non-negligible differences in retrieval performance can be attributed to differences in relevance judgments, was not supported. Kazhdan found several such differences of 10% or more in one of the parameters he used to judge retrieval performance (mean relative noise). Nevertheless, Kazhdan did find evidence to support the weaker hypothesis of Lesk and Salton, that no variations in the relative performance of different processing methods can be attributed to differences in relevance judgments.

The study reported here also investigated the degree to which variations in relevance judgments affect the evaluation of retrieval performance. The study represents an effort to replicate the findings of Lesk and Salton (1968) and Kazhdan (1979) regarding both the strong and weak hypotheses of Lesk and Salton. To some extent, the present study also represents an improvement on the previous studies, as it made use of four sets of relevance judgments (more than either of the previous studies) and six different document representations (more than Lesk and Salton).

Table 1. Relevance categories, definitions, and interpretations used by judges (adapted from Wood, Wood, & Shaw, 1990, p. 15)

Relevance category	Definition	Interpretation
Highly relevant	The paper is a direct response to the query.	I would be angry if the system failed to retrieve this paper.
Marginally relevant	The paper is topically relevant, but not a direct response to the query.	I would not be displeased if the system retrieved this paper or failed to retrieve this paper.
Not relevant	The paper is not relevant to the query.	I would be insulted if the system retrieved this paper.

METHODS

Because test databases generally employ a single set of relevance judgments, few test collections make possible a study of the effects of variations in relevance judgments on the evaluation of retrieval performance. However, the recently developed cystic fibrosis collection (Shaw, Wood, Wood, & Tibbo, 1990; Wood, Wood, & Shaw, 1990) does include multiple sets of relevance judgments, thus allowing this question to be investigated.

The present study was based on a test collection of 1239 papers, published between 1974 and 1979 and indexed with the term *cystic fibrosis* in the National Library of Medicine's Medline file, and 100 queries with three sets of relevance evaluations from subject experts and one set of relevance evaluations from a medical bibliographer (Shaw *et al.*, 1990; Wood et al., 1990).

The existence of these multiple sets of relevance evaluations made it possible to test whether variations in relevance judgments would affect performance measures in an information retrieval system. Set *A* of relevance judgments was contributed by the principal author of the queries, a professor of pediatrics actively involved in cystic fibrosis research. Set *B* represents the evaluations of nine faculty colleagues of the principal author of the queries; each faculty member evaluated a subset of queries relevant to his or her area of expertise in cystic fibrosis research. Set *C* comprises the judgments of four postdoctoral fellows in pediatric pulmonology, each of whom evaluated a subset of the queries. The final set of relevance judgments, set *D*, was contributed by a medical bibliographer with extensive online search experience. The individuals were asked to examine the full text of the documents assigned to them and to code documents as "highly relevant" to the query, "marginally relevant," or "not relevant." The definitions and interpretations of the three categories of relevance are given in Table 1.

The retrieval experiments described in this article examined all four sets of relevance judgments but restricted relevant documents to those judged "highly relevant" to a query. Table 2 shows the number of "highly relevant" document-query pairs for each set of relevance judgments. The fact that the medical bibliographer rated more documents as relevant to queries than did any of the sets of subject experts is consistent with observations by Rees and Schultz (1967) and Saracevic (1970) that the greater the subject knowledge of a judge or set of judges, the fewer the documents that will be judged relevant. The fact that search intermediaries tend to be more liberal in their assessments of relevance is perhaps neither surprising nor undesirable; the greater part of the burden of rejecting documents is thereby placed on end users, and the likelihood that a potentially relevant document is discarded is lessened.

Table 2. Number of "highly relevant" document-query pairs per set of relevance judgments

Set *A*	Set *B*	Set *C*	Set *D*
1104	919	1028	1411

Table 3. Agreement between sets of relevance evaluations $(X \cap Y)/(X \cup Y)$

	Set A	Set B	Set C	Set D
Set A		58.7%	63.6%	42.7%
Set B	748/1275		53.4%	36.0%
Set C	829/1303	678/1269		40.9%
Set D	752/1763	617/1713	708/1731	
Mean	55.0%	49.4%	52.7%	39.9%

Table 3 shows the agreement between sets of relevance evaluations. The measure of agreement is that of Lesk and Salton (1968) and is derived by dividing the total number of common items by the total number of distinct items:

$$\frac{(X \cap Y)}{(X \cup Y)} \tag{1}$$

Of particular interest is the relatively high degree of agreement among the subject experts, represented by sets A through C, and the relatively low degree of agreement between the subject experts and the medical bibliographer, represented by set D. Again, this finding is consistent with those reported by Cuadra and Katter (1967a), Rees and Schultz (1967), and Saracevic (1970) that the agreement among judges will be higher for judges with greater subject knowledge.

The SMART information retrieval system (Buckley, 1985; Salton, 1971) was used to produce single-term indexing representations for the collection based on six different document representations and based on the texts of the queries. The document representations were:

- Titles, abstracts, and MeSH major and minor subject headings; full stemming.
- Titles, abstracts; full stemming.
- Titles, abstracts; no stemming.
- Abstracts only; full stemming.
- Titles only; full stemming.
- MeSH major and minor subject headings only; full stemming.

Weights for the document vector collection were based on augmented normalized term frequencies and the standard inverse document frequency measure, the single-term weighting system found by Salton and Buckley (1988) to produce the best results for most test collections. Here the augmented term frequency of term k in document i, af_{ki}, is equal to

$$0.5 + 0.5 * (tf_{ki}/\max\_tf_i), \tag{2}$$

where tf_{ki} is the raw frequency of term k in document i and $\max\_tf_i$ is the maximum raw term frequency in document i. Thus, all values for af_{ki} are normalized within the range $0.5 < af_{ki} < 1.0$.

The inverse document frequency measured used here is then w_{ki}, the weight of term k in document i, and is equal to

$$af_{ki} * [\log_2 * (d/d_k)], \tag{3}$$

where af_{ki} is the augmented normalized frequency of term k in document i, d is the total number of documents in the collection, and d_k is the number of documents in which term k appears. A high weight in a document denotes a term that occurs relatively infrequently in that document and relatively infrequently in the collection. Such a term, then, is one that is relatively specific to the associated document.

Table 4. Average precision as a function of document representation and relevance evaluation

	Set *A*	Set *B*	Set *C*	Set *D*
Titles, abstracts, headings; full stemming	.3877	.3589	.3582	.3440
Titles, abstracts; full stemming	.3797	.3451	.3490	.3314
Titles, abstracts; no stemming	.3608	.3258	.3265	.3116
Abstracts only; full stemming	.3440	.3137	.3117	.2960
Titles only; full stemming	.2814	.2627	.2628	.2377
Headings only; full stemming	.1926	.1630	.1683	.1599

RESULTS

Retrieval results were obtained by using the SMART information retrieval system to conduct a sequential match of query and document vectors. Differences in performance were based on average precision figures for three intermediate recall levels: .25, .50, and .75. In discussing the magnitude of the differences in performance in these experiments, the definitions of "noticeable" and "material" differences given by Sparck Jones (1974) will be used. Sparck Jones defined a "noticeable" difference between the performance of two retrieval systems as lying in the 5 to 10% range; a "material" difference was defined as more than 10%. (Due to the replicative nature of the study, recall and precision alone were used to measure system performance. The use of other measures, such as expected search length [Cooper, 1968] or the *E* measure [Van Rijsbergen, 1979; Shaw, 1986a], is possible, although difficulties in judging the magnitude of performance differences remain largely unresolved.)

Table 4 shows average precision as a function of the four different sets of relevance judgment and the six different document representations. In no case was there a noticeable or material difference in retrieval performance due to variations in relevance judgments; by contrast, noticeable and material differences in performance due to variations in document representation were found. Thus, the stronger hypothesis of Lesk and Salton (1968) that variations in relevance judgments do not produce substantial variations in retrieval performance, as represented by average precision, was supported for the four sets of relevance judgments investigated here.

Lesk and Salton (1968) suggested four reasons why variations in relevance judgments may not affect recall and precision results. Three of these reasons have been examined in further detail in this study.

First, Lesk and Salton (1968, p. 346) argued that "recall and precision data are normally given as averages over many search requests; these averages may not be sensitive to small variations in the results for individual queries."

Evidence supporting this argument may be seen in Table 5, where average precision figures for each set of relevance judgments are listed for each query. While there were no noticeable differences in the overall average precision figures for the four sets of relevance judgments (see Table 4), there were many noticeable and material differences for individual queries. For example, in the 96 queries for which sets *A* and *B* could be compared, there were 16 noticeable differences and 31 material differences. The number of noticeable differences between sets of judgments ranged from 11 (between sets *B* and *C*) to 18 (between sets *A* and *C* and also between sets *C* and *D*). The number of material differences ranged from 28 (between sets *A* and *C*) to 39 (between sets *B* and *C*).

Lesk and Salton (1968, p. 346) also speculated that

> disagreements among relevance judges may affect mostly the borderline cases, while preserving a general consensus for a large set of items definitely 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.

This speculation was examined by testing two sets of relevance judgments: one in which relevant documents were defined as those that were judged "highly relevant" by all

Table 5. Average precision as a function of relevance evaluation (titles, abstracts, headings; full stemming)

Query	Set *A*	Set *B*	Set *C*	Set *D*	Query	Set *A*	Set *B*	Set *C*	Set *D*
1	0.3753	0.3908	0.2667	0.4213	51	0.2423	0.2446	0.2100	0.2434
2	–	–	–	0.3333	52	1.0000	1.0000	1.0000	1.0000
3	0.4530	0.3939	0.0842	0.3976	53	0.1673	0.2533	0.2319	0.0299
4	0.1875	0.1429	0.1875	0.1875	54	0.7721	0.5575	0.8013	0.6656
5	0.2366	0.0970	0.3105	0.1008	55	0.0345	0.0655	0.0572	0.0347
6	0.2500	0.2407	0.1738	0.1523	56	0.0556	0.0556	0.0556	0.0564
7	0.5575	–	0.1607	0.4000	57	0.1593	0.3484	0.3093	0.1619
8	0.1414	0.0808	0.0527	0.0570	58	0.2284	0.1454	0.2302	0.2116
9	0.1143	0.1429	0.1255	0.1346	59	0.0393	0.0213	0.0292	0.1166
10	0.5631	0.2924	0.7222	0.6959	60	0.0685	0.0801	0.0685	0.1455
11	0.9048	0.9333	0.8095	0.7262	61	0.9028	0.7667	0.9111	0.6320
12	0.2536	0.3704	0.3623	0.0556	62	0.0451	0.0401	0.0456	0.0496
13	0.0185	0.0150	0.0150	0.0449	63	0.1447	0.0802	0.1029	0.1728
14	0.1325	0.1344	0.1614	0.1555	64	0.1107	0.1358	0.1107	0.0744
15	0.2994	0.1334	0.3072	0.2327	65	0.2514	0.1738	0.2049	0.2322
16	0.1844	0.0626	0.2042	0.1533	66	0.0915	0.1068	0.0976	0.0895
17	0.0816	0.0282	0.0935	0.0532	67	0.2799	0.2230	0.2485	0.1047
18	0.5119	–	0.7222	0.3371	68	0.2335	0.3287	0.1113	0.6270
19	0.0670	0.0611	0.0670	0.0616	69	0.7500	0.9167	0.8056	0.7222
20	0.6660	0.7042	0.5677	0.4843	70	1.0000	1.0000	1.0000	0.7692
21	1.0000	0.8889	0.7556	0.6992	71	0.0250	0.0250	0.0250	0.0250
22	0.0504	0.0490	0.0784	0.1677	72	0.7556	0.1383	0.7556	0.5000
23	0.0156	0.0188	0.0099	0.0264	73	0.7778	0.8750	0.8611	0.8444
24	0.0193	0.0236	0.0345	0.0069	74	0.6894	0.6818	0.6894	0.6681
25	0.1289	0.0190	0.1318	0.0882	75	0.1814	0.2500	0.2500	0.2750
26	0.4345	0.3083	0.3020	0.6255	76	0.3778	0.7333	1.0000	0.8333
27	1.0000	1.0000	1.0000	0.3398	77	0.0609	0.0608	0.0814	0.0973
28	0.5556	0.0080	1.0000	0.0442	78	0.0292	0.0367	0.0729	0.1197
29	0.0199	0.0182	0.0111	0.0416	79	0.1362	0.0993	0.0883	0.1000
30	0.2000	0.2179	0.0667	0.2429	80	0.2460	0.6667	0.2460	0.2683
31	0.4508	0.3570	0.3902	0.4257	81	0.7315	0.7778	0.3333	0.7083
32	0.5000	0.5000	0.2436	0.7778	82	0.5484	0.4201	0.5238	0.7897
33	0.2252	0.2490	0.2581	0.2243	83	0.4000	0.5576	0.3381	0.4339
34	0.6325	0.6325	0.4526	0.2342	84	1.0000	1.0000	1.0000	0.0833
35	0.1922	0.1345	0.1050	0.1345	85	0.0302	0.0270	0.0208	0.0248
36	0.0601	0.0109	0.0601	0.0038	86	0.1590	0.1448	0.0616	0.0711
37	0.4903	0.4442	0.5257	0.5563	87	0.2442	0.3198	0.2442	0.1241
38	0.1778	0.3278	0.4667	0.3686	88	0.8889	0.6082	0.8889	0.1278
39	0.1413	0.0273	0.1310	0.2116	89	0.6814	0.3978	0.3431	0.7436
40	0.0579	0.0147	0.0819	0.0738	90	0.2863	0.5897	0.3732	0.6154
41	0.2500	0.2500	0.3819	0.3819	91	0.0637	0.0540	0.0572	0.0794
42	0.8750	1.0000	0.4927	1.0000	92	0.3897	0.2902	0.3213	0.2464
43	0.4531	0.1756	0.1402	0.2182	93	1.0000	1.0000	0.4460	0.3417
44	0.4546	0.4698	0.3826	0.4914	94	0.3556	0.1364	0.3250	0.3891
45	0.9444	0.9583	0.2704	0.7778	95	0.5917	–	0.2500	0.6695
46	0.5375	0.9167	0.6993	0.5919	96	1.0000	1.0000	1.0000	1.0000
47	0.2346	0.1988	0.1437	0.2026	97	0.3111	0.3111	0.2000	0.2000
48	0.1623	0.2705	0.1623	0.1649	98	0.1097	0.0220	0.4386	0.0751
49	0.7548	0.8571	0.8119	0.7821	99	1.0000	0.7500	1.0000	1.0000
50	0.7176	0.6789	0.6152	0.7176	100	1.0000	0.6885	1.0000	1.0000

four judges; and one in which relevant documents were defined as those that were judged "highly relevant" by one and only one judge. The results support the above suggestion. (Table 6 shows retrieval results for the set of documents on which all four judges agreed and for the set of documents that were judged "highly relevant" by one and only one judge.) The average precision for the set of documents on which all four judges agreed was materially higher than the average precision for the other set (.3309 vs .0831). As Lesk and Salton (1968, p. 354) expressed it: "The query-document pairs which are most closely and inarguably related are exactly the pairs on which the retrieval performance is best."

Finally, Lesk and Salton (1968, p. 346) suggested that

> 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.

Table 6. Retrieval results average precision at standard recall levels (titles, abstracts, headings; full stemming)

Recall	Documents judged "highly relevant" by all four judges	Documents judged "highly relevant" by one and only one judge
0.00	0.4531	0.2383
0.05	0.4457	0.2197
0.10	0.4227	0.1814
0.15	0.4085	0.1556
0.20	0.3980	0.1422
0.25	0.3873	0.1313
0.30	0.3715	0.1274
0.35	0.3515	0.0979
0.40	0.3492	0.0945
0.45	0.3372	0.0787
0.50	0.3365	0.0769
0.55	0.3106	0.0625
0.60	0.3041	0.0593
0.65	0.2959	0.0559
0.70	0.2730	0.0449
0.75	0.2691	0.0410
0.80	0.2589	0.0389
0.85	0.2526	0.0341
0.90	0.2436	0.0319
0.95	0.2369	0.0306
1.00	0.2347	0.0304
Average precision for 3 intermediate points	0.3309	0.0831
Number of queries	86	87

Support for this argument can be seen in Table 4. For each set of relevance judgments, the relative performance of the six different document representations was the same. The combination of titles, abstracts, and MeSH major and minor subject headings with full stemming performed best, regardless of which set of relevance judgments was being used. Titles and abstracts with full stemming always provided the second best performance; titles and abstracts with no stemming were consistently third best. Abstracts only with full stemming, titles only with full stemming, and MeSH major and minor subject headings only with full stemming ranked fourth, fifth, and sixth, respectively for all sets of relevance judgments. Thus, the weaker hypothesis of Lesk and Salton (1968), that no variations in the relative performance of different processing methods can be attributed to differences in relevance judgments, also appears to be corroborated for the four sets of relevance judgments investigated here.

DISCUSSION

In their discussion of the relevance evaluations provided with the cystic fibrosis test collection, Shaw *et al.* (1990, p. 9) noted that "it remains to be shown how the numerous discrepancies between the relevance judgments of subject experts and those of the medical bibliographer affect performance measures." The study reported here found no evidence that the differences in relevance judgments in the cystic fibrosis test collection affected the evaluation of retrieval performance. Instead, the study found evidence to support the strong and weak hypotheses of Lesk and Salton (1968) and found evidence to support their suggestions as to why the hypotheses might hold.

Obviously, further investigation into this topic is warranted. In his review of research on factors affecting relevance judgments, Saracevic (1970, p. 134) suggested that while the study by Lesk and Salton (1968) represented "a step in the right direction," it would be worthwhile to conduct additional experiments "on different systems under different conditions with differently obtained judgments." The study reported here provides one such experiment—a different test collection with four sets of relevance judgments. However, like the investigation by Lesk and Salton and that reported by Kazhdan (1979), the study re-

ported here examined variations in relevance judgments due to different sets of judges only. The relevance assessments were obtained under otherwise similar conditions; there were no controlled variations in the type of document representation used to judge relevance (all judges were instructed to utilize full text), in definitions of relevance, in rating scales, or in any of the other variables known to affect judgments of relevance. It would be helpful to study sets of relevance judgments obtained under such variations to determine whether and to what extent the evaluation of retrieval system performance would be affected.

It is interesting, for example, to note that the study reported here corroborated the findings of Lesk and Salton (1968) even though the definitions and interpretations of relevance used by the judges in the test collection employed in this investigation (see Table 1) were quite different—and perhaps less objective—than that of Lesk and Salton, who instructed their judges to consider a document relevant " 'if it is directly stated in the abstract as printed, or can be directly deducted from the printed abstract, that the document contains information on the topic asked for in the query' " (Lesk & Salton, 1968, p. 347). It would be useful to examine the degree to which differences in the definition of relevance might affect retrieval performance for the same test collection.

Likewise, it would be worthwhile to consider the effects of various rating scales on system performance. The dichotomous scale used in this study and in Lesk and Salton (1968), whereby documents are judged to be either 100% relevant or not relevant to a given query, has been criticized as overly simplistic (Cuadra & Katter, 1967a; Rees & Schultz, 1967; Eisenberg & Hu, 1987), and it would be useful to examine retrieval performance based on both category rating scales and magnitude estimations, such as those explored by Eisenberg and Barry (1986, 1988).

It would also be worthwhile to study the degree to which other retrieval techniques are affected by variations in relevance judgments. Burgin (1991) has shown that the vector space model used in the study reported here is more robust against variations in indexing exhaustivity than is the single-link clustering model studied by Shaw (1990a, b, 1986b), and it would be helpful to examine whether evaluation of performance for the single-link clustering model is more or less robust against variations in assessments of relevance. Because the single-link clustering model takes interdocument similarities into account prior to any consideration of the relationship between queries and documents, differences between the two models might be expected.

Finally, the effect of variations in relevance judgments on the evaluation of relevance feedback techniques in retrieval systems is of considerable interest because such techniques are implemented to increase the sensitivity of retrieval systems to the varying needs and varying determinations of relevance of different users. Eisenberg and Barry (1988) have suggested that order effects might affect systems that employ relevance feedback, and it would be worthwhile to determine the impact of these and other factors that result in deviations in relevance assessments.

The extent to which variations in relevance judgments affect the evaluation of retrieval system performance is, of course, critical to the standard methods of evaluating such systems as well as to the evaluation of online public access catalogues (O'Brien, 1990), classification schemes (Carpenter, Jones, & Oppenheim, 1978), and generalized text filtering techniques in computer-mediated communications systems (Losee, 1989) and to estimates of expected performance in retrieval systems (Losee, 1991). Some researchers who have found variations in relevance judgments have suggested that these variations invalidate the standard methods of evaluating retrieval system performance, especially when comparisons are made among different systems (Cuadra & Katter, 1967a; Rees & Schultz, 1967). A fuller awareness of the degree to which deviations in relevance assessments affect the evaluation of retrieval system performance should enhance our understanding of the extent to which such intersystem comparisons can be made and the conditions under which they can be made.

REFERENCES

Barhydt, G.C. (1964). A comparison of relevance assessments by three types of evaluator. *Proceedings of the American Documentation Institute, October 5-8, 1964* (pp. 383-385). Washington, DC: American Documentation Institute.

Buckley, C. (1985). *Implementation of the SMART information retrieval system*. Technical Report No. 85-686. Ithaca, NY: Cornell University.

Burgin, R. (1991). The effect of indexing exhaustivity on retrieval performance. *Information Processing & Management, 27*, 623-628.

Carpenter, A.M., Jones, M., & Oppenheim, C. (1978). Retrieval tests on 5 classification schemes—studies on patent classification systems II (1). *International Classification, 5*, 73-80.

Cooper, W.S. (1968). Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation, 19*, 30-41.

Cuadra, C.A., & Katter, R.V. (1967a). *Experimental studies of relevance judgments*. Santa Monica, CA: Systems Development Corporation.

Cuadra, C.A., & Katter, R.V. (1967b). Opening the black box of 'relevance.' *Journal of Documentation, 23*, 291-303.

Eisenberg, M.B. (1988). Measuring relevance judgments. *Information Processing & Management, 24*, 373-389.

Eisenberg, M.B., & Barry, C. (1986). Order effects: a preliminary study of the possible influence of presentation order on user judgments of document relevance. *Proceedings of the 49th ASIS Annual Meeting, Volume 23. Chicago, Illinois, September 28-October 2, 1986* (pp. 80-86). Medford, NJ: Learned Information, Inc.

Eisenberg, M.B., & Barry, C. (1988). Order effects: a study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science, 39*, 293-300.

Eisenberg, M.B., & Hu, Xiulan. (1987). Dichotomous relevance judgments and the evaluation of information systems. *Proceedings of the 50th ASIS Annual Meeting, Volume 24. Boston, Massachusetts, October 4-8, 1987* (pp. 66-69). Medford, NJ: Learned Information, Inc.

Gull, C.D. (1956). Seven years of work on the organization of materials in the special library. *American Documentation, 7*, 320-329.

Kazhdan, T.V. (1979). Effects of subjective expert evaluation of relevance on the performance parameters of a document-based information-retrieval system. *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 13*, 21-24.

Lesk, M.E., & Salton, G. (1968). Relevance assessments and retrieval system evaluation. *Information Storage Retrieval, 4*, 343-359.

Losee, R.M. (1989). Minimizing information overload: the ranking of electronic messages. *Journal of Information Science, 15*, 179-189.

Losee, R.M. (1991). An analytic measure predicting information retrieval system performance. *Information Processing & Management, 27*, 1-13.

O'Brien, A. (1990). Relevance as an aid to evaluation in OPACs. *Journal of Information Science, 16*, 265-271.

O'Connor, J. (1969). Some independent agreements and resolved disagreements about answer-providing documents. *American Documentation, 20*, 311-319.

Rath, G.J., Resnick, A., & Savage, T.R. (1961). Comparisons of four types of lexical indicators of content. *American Documentation, 12*, 126-130.

Rees, A.M. (1967). Evaluation of information systems and services. *Annual Review of Information Science and Technology, 2*, 63-86.

Rees, A.M., & Schultz, D.G. (1967). *A field experimental approach to the study of relevance assessments in relation to document searching, final report*. Cleveland, OH: Center for Documentation and Communication Research, School of Library Science, Case Western University.

Regazzi, J.J. (1988). Performance measures for information retrieval systems—an experimental approach. *Journal of the American Society for Information Science, 39*, 235-251.

Resnick, A. (1961). Relative effectiveness of document titles and abstracts for determining the relevance of documents. *Science, 134* (3484), 1004-1006.

Rorvig, M.E. (1988). Psychometric measurement and information retrieval. *Annual Review of Information Science and Technology, 23*, 157-189.

Salton, G. (Ed.). (1971). *The SMART retrieval system—experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*, 513-523.

Saracevic, T. (1969). Comparative effects of titles, abstracts and full texts on relevance judgments. *Proceedings of the American Society for Information Science, Volume 6* (pp. 293-299). Westport, CT: Greenwood Publishing Corporation.

Saracevic, T. (1970). The concept of relevance in information science: a historical review. In T. Saracevic (Ed.), *Introduction to information science* (pp. 111-151). New York: R.R. Bowker.

Saracevic, T. (1976). Relevance: a review of the literature and a framework for thinking on the notion in information science. *Advances in Librarianship, 6*, 79-138.

Shaw, W.M., Jr. (1986a). On the foundation of evaluation. *Journal of the American Society for Information Science, 37*, 346-348.

Shaw, W.M., Jr. (1986b). An investigation of document partitions. *Information Processing and Management, 22*, 19-28.

Shaw, W.M., Jr. (1990a). An investigation of document structures. *Information Processing & Management, 26*, 339-348.

Shaw, W.M., Jr. (1990b). Subject indexing and citation indexing—part II: an evaluation and comparison. *Information Processing & Management, 26*, 705-718.

Shaw, W.M., Jr., Wood, J.B., Wood, R.E., & Tibbo, H.R. (1990). *The cystic fibrosis data base: content and research opportunities*. Unpublished paper. Chapel Hill, NC: University of North Carolina at Chapel Hill, School of Information and Library Science.

Sparck Jones, K. (1974). Automatic indexing. *Journal of Documentation, 30*, 393-432.

Van Rijsbergen, C.J. (1979). *Information retrieval*. London: Butterworths.

Wood, J.B., Wood, R.E., & Shaw, W.M., Jr. (1990). *The cystic fibrosis data base*. Technical Report No. 8902. Chapel Hill, NC: University of North Carolina at Chapel Hill, School of Information and Library Science.