Pergamon

Information Processing & Management, Vol. 30, No. 3, pp. 305–314, 1994

0306-4573/94 $6.00 + .00

0306-4573(93)E0016-H

RELEVANCE ODDS OF RETRIEVAL OVERLAPS FROM SEVEN SEARCH FIELDS

MIRANDA LEE PAO
School of Information and Library Studies, The University of Michigan,
550 East University, Ann Arbor, MI 48109, U.S.A.

(*Received* 9 *December* 1992; *accepted in final form* 22 *July* 1993)

Abstract—Data contained in a 1982 paper were analyzed in terms of relevance odds of common items retrieved by searching any two content-bearing search fields. While the 1982 study compared the relative retrieval performance of 7 search fields, the present study shows that duplicate documents retrieved by the use of terms from any two of the fields would have higher odds of being judged relevant than those retrieved by only one of the fields. Sixty-three relevance odds were computed using the log cross product technique. The highest relevance odds were associated with common items retrieved from assigned descriptors and from truncated free-text terms from either the title or abstract fields; their relevance odds were 19 to 2 in favor of overlaps. Overlap retrieval could be considered a strategy for high precision searching.

INTRODUCTION

The literature refers to retrieval overlap as the degree of agreement between two sets of retrievals derived from two parallel searches conducted for the same query. Retrieved items may not all be relevant. Relevant overlap, however, is the percentage of overlap of the relevant retrieval from the same two parallel searches.

With the exception of an overlap study by Katzer and his colleagues, overlap data have been incidental to the main findings of retrieval experiments (Katzer *et al.*, 1982). Most studies focused on the search outcomes by the use of different search fields, or different searchers (McGill, Koll, & Norealt, 1979; Saracevic & Kantor, 1988a,b). An early example referred to by Katzer's team (1982) is Williams' work in which she examined the unique contributions by the use of title words and terms found in abstracts. In comparing the coverage of interdisciplinary subjects from different databases, researchers from Drexel University also noted that title words added 10% to descriptor searching in MEDLINE, 17% to searching Excerpta Medicus, 7% to Social Science Citation Index®, and 14% to PsychInfo Index (McCain, White, & Griffith, 1986). More recently, with the proliferation of full-text databases, the relative retrieval performance of full-text searching has been the subject of several studies (see Tenopir, 1985; Ro, 1988; McKinin & Sievert, 1989; McKinin *et al.*, 1991).

In all cases, researchers reported that searches conducted for the same query produced complementary sets. In other words, more materials could be expected if several output sets for a given topic were pooled regardless of the choice of variable. Simultaneously, authors also reported small overlaps in comparing the retrieval performance of parallel searches.

For example, the retrievability tests of Medical Behavioral Sciences topics on PsychInfo, Excerpta Medicus, MEDLINE, and BIOSIS produced small overlaps from different databases on identical topics (McCain *et al.*, 1987). Comparing descriptor-based retrieval with cocitation clusters of the same topics, the Drexel team also found small overlaps (White *et al.*, 1984). Comparisons of subject retrievals and citation retrievals from the same seed documents from MEDLINE and SCISearch respectively produced no more than 10% overlaps (White, 1989). Experiments by McCain, and by Pao produced only approximately 10% overlaps from using citations and terms to search the same topics (McCain, 1989; Pao

1986, 1993, Pao & Worthen, 1989). Small overlaps also resulted from searches by different searchers on identical topics (Saracevic & Kantor, 1988b).

Although low overlaps are the norm, Saracevic and Kantor (1988b) stated that one of the most important findings in their study of online searching was that for the same query, as the number of searchers who retrieved a given item increased, so did the odds of its relevance. Indeed, an item retrieved by more than 4 searchers was nearly 6 times more likely to be judged relevant as an item retrieved by only one searcher. Compared with items retrieved by one searcher, there seemed to be a higher likelihood that multiple retrievals would be judged relevant. Pao (1993) also found high precision in the overlap sets in both of her studies. A mean of 9% overlap of papers (relevant and nonrelevant) was found in an average query on nutritional problems (Pao & Worthen, 1989). Yet, all of the overlap items in 20 of the 25 queries with multiple retrievals were judged relevant. Consequently, although the overlap set was small, an average of 92% precision was found. Similar results were found by Pao (1993) in a sample of searches collected in several health sciences libraries. In this study, all retrieval overlaps in 63% of the topics searched were relevant. Eighty percent of the overlap items in both studies comparing citation-based and term-based searching of the same queries were found to be relevant. She also showed that the odds for overlap items to be judged relevant were significantly higher than nonoverlap items.

If indeed overlap retrievals were more likely to be judged relevant to the search topic, multiple searches could be deployed as a strategy to produce high precision sets. Since online searchers routinely make use of different content-bearing search fields for subject retrieval, multiple searches might be a practical quality-filtering strategy. More specifically, the present study seeks to produce empirical evidence that duplicate items retrieved by any two search fields would be more likely judged relevant than those retrieved by only one field.

RESEARCH OBJECTIVE

In a 1982 paper entitled "A Study of the Overlap Among Document Representations," Katzer and colleagues (1982) at Syracuse University compared the retrieval performance of 7 search fields. In this NSF-sponsored project, the investigators used the expression "document representation" to denote a content-bearing search field or two fields used in combination, such as title and abstract. For simplicity, the term "search field" will be used in its place in the following discussion. The Syracuse team used terms from each of the 7 search fields to search 84 topics. For each topic, the team also examined the extent of retrieval overlap between each field and every other field. Additionally, overlaps of relevant retrieval were reported.

Major findings from the 1982 paper were: (1) none of the 7 fields performed significantly better than the others in terms of recall, precision, total items retrieved, and two overlap measures—pairwise asymmetric and union overlap; (2) there were small overlaps between all pairs of fields; and (3) a relation existed between relevance and the size of the overlap set. The third finding stated that the "most relevant documents have higher overlaps than do all relevant documents, and these have higher overlaps than do all retrieved documents" (Katzer *et al.*, 1982, p. 272).

The last findings also suggests that while overlaps between two retrieved sets conducted for the same topic may be small, it would contain a relatively high proportion of relevant documents. For high precision, one might consider identifying duplicate items retrieved from two separate searches using terms from two different subject fields.

The 1982 study analyzed the relative performance of the 7 search fields. The researchers also attempted to identify an optimum ordering of the fields for retrieval purpose. In contrast, the focus of the present study is to show the relationship between two properties of the retrieved items, namely, whether the items were retrieved by two fields or by only one field, and the relevance or nonrelevance of the items retrieved. The purpose of this study is to seek empirical evidence to support the claim that duplicate documents retrieved by the use of any two search fields would have higher odds of being judged relevant than

those retrieved by only one field. Data for this study had been extracted entirely from the 1982 study. No new searches were added.

This study seeks answers to the following question:

> What were the odds that an item was judged relevant when it was retrieved by two search fields as opposed to being retrieved by only one field?

METHODOLOGY

Methods and materials used in the 1982 study

The 1982 study collected a total of 84 queries searched in a 12,000-item database. Each query was searched using each of the 7 fields in turn. Three fields were composites of other fields. The 7 search fields tested by the Syracuse team are as follows:

DD	assigned descriptors
AA	text words from abstracts
TA	text words from titles and abstracts (TT + AA)
DI	descriptors and identifiers (DD + II)
ST	stemmed text words from titles and abstracts
TT	text words from titles
II	text words assigned by indexers as identifiers

Searchers were instructed to construct "high recall" searches using only one field. Seven retrieved sets for each query were merged to eliminate duplicates. Thus, the pooled set had the potential to contain most of the relevant document from the database. These 84 sets were then evaluated by users on a four-point scale: '1' for "definitely relevant," '2' for "probably relevant," '3' for "probably not relevant," and '4' for "not relevant." Two versions of recall were computed. The first was the percentage of relevant items retrieved, that is, items with either a score of 1 or 2 on the four-point scale. This was based on an all-inclusive relevance criterion. In the second version, the recall ratio included only those definitely relevant documents with the score of '1.'

In addition to the search-based comparisons using average recall and precision of all individual queries, another set of recall and precision associated with each field across queries was also computed. The latter group of ratios was based on the aggregate number of retrieved items. According to Katzer, these values minimized the bias introduced by the atypical searches. To compute the precision of a search field, the aggregate number of relevant documents from all queries using that field was divided by the aggregate number of documents retrieved from all queries.

The overlap of every other pair of the 7 search fields was examined in terms of: (a) retrieval overlap; (b) overlaps of all relevant items; and (c) overlaps of those judged definitely relevant. With 7 fields, there are 21 possible combinations of pairwise comparisons. With 84 queries and 7 fields, a total 1,764 pairwise comparisons of items were collected. Two types of overlap comparisons were made, namely, the pairwise asymmetric overlap (A_{ij}) and the union overlap (U_{ij}). The computational formula are as follows:

$$\text{Asymmetric overlap: } A_{ij} = \frac{n(R_i \cap R_j)}{n(R_i)}$$

$$\text{Union overlap: } U_{ij} = \frac{n(R_i \cap R_j)}{n(R_i \cup R_j \cup R_k \cup \ldots \cup R_o)}$$

A large amount of overlap and retrieval data on the 7 content-bearing search fields was assembled.

Cross product analysis

Cross product analysis has been shown to be a powerful and useful analytic method, used by Saracevic and Kantor (1988a,b). Basically, it is used to produce the odds that given two values for each of two variables, the high value of the independent variable as opposed to the low value is associated with the high value of the dependent variable as contrasted with the low value. Suppose two parallel searches were conducted on a query using title words and descriptors respectively. Each retrieved item could appear in one of the following: (a) the printout from the use of title words alone; (b) the printout from the use of descriptors alone; or (c) both printouts. Thus, the independent variable is the condition of overlap for the individual item. Since a relevance score was attached to each item, the dependent variable is its relevance judgment.

To compute the odds, two pieces of information are needed for each retrieved item for every pair of search fields: (1) whether the item was retrieved by one or both fields, in other words, whether it was an overlap item or not; and (2) whether the item was judged relevant or not relevant. With this data, each item can be placed in one of the 4 cells of the following 2 × 2 contingency table:

		Independent variable: Condition of overlap		
		Overlap	No overlap	Total
Dependent variable: Relevance judgment	Relevant	p	q	t
	Not rel.	r	s	u
	Total	v	w	n

Each cell needs to contain the aggregate number of items satisfying the two specified conditions. Consequently, an item-level instead of a search-level analysis is called for.

Suppose the odds were computed to be 2. In this case, one may conclude that if an item appeared in both printouts, there is a 2-to-1 chance that it would be judged relevant as compared with items retrieved by only one field alone. The computational procedures with examples are given in detail in the papers by Saracevic and Kantor (1988a,b).

Strong relevance, normal relevance, and weak relevance

Two versions of recall and precision based on two levels of relevance judgment were presented in the 1982 paper. The *all-inclusive* criterion of relevance included all items rated '1' (definitely relevant) or '2' (probably relevant), and the *definitely relevant* criterion was limited only to those rated '1.' The latter imposed a more rigorous requirement in that only 'definitely relevant' items were considered. Investigators such as McKinin (McKinin *et al.*, 1991) and Pao (Pao & Worthen, 1989; Pao, 1993) have since utilized these two levels of relevance criteria. In a similar vein, Saracevic and Kantor (1988b) made the distinction among three types of relevance, namely, strong relevance, normal relevance, and weak relevance. The comparisons were made as follows:

Normal relevance: R + pR vs. NR

Strong relevance: R vs. NR

Weak relevance: R vs. pR + NR

where: R are those items judged as definitely relevant, pR are those items judged as partially relevant, and NR are those items judged as not relevant.

As the design of the 1982 study incorporated relevance judgment using a system of four-point score, the data may be reanalyzed based on the three gradations of relevance

with the "partially relevant" substituted for "probably relevant" category. For the all-inclusive normal relevance, items in both categories of relevance can be included. On the other hand, a more stringent requirement can be imposed by the use of only those rated '1' for strong relevance and discarding those rated as "probably relevant." For weak relevance, those with a score of '1' can be juxtaposed with all other retrieved items. Both versions of analyses will be used in the following sections.

Data extraction

First, consider normal relevance as the basis of judgment of relevance for the dependent variable. The last column in Table 1 in the 1982 paper contains the average number of retrieved papers by the use of a single search field. For example, for TT (text words from titles), it is 12.429. By using TT alone, the aggregate retrieval, whether relevant or not, across all 84 queries, is simply the product of 84 and the average number of retrieved items 12.429. That is, the aggregate retrieval for TT is 1,044 (or 12.429 × 84). Similarly, the aggregated retrieval for DD (assigned descriptors) is computed to be 1,112 (or 13.239 × 84) items. From the last section of Table 6 of the same paper, the asymmetric retrieval (relevant and nonrelevant) overlap ratio between TT and DD is found to be 0.131. In other words, 13.1% of papers retrieved by TT overlapped with DD. Thus the number of common items retrieved by both TT and DD is 137 (or 0.131 × 1044) papers which is the value for (v) in the contingency table. As a result, eliminating one set of common retrieved items, the union retrieval from the use of TT and DD (n) is computed to be 2,019 (or 1044 + 1112 − 137). From Table 5 of the 1982 paper, the aggregate precision ratios are given as the ratios of the aggregate number of relevant items retrieved across all queries to the aggregate retrieval. For TT, the ratio is 0.378, if all "relevant" or "probably relevant" items were included. Given that the aggregate retrieval for TT is 1,044, the number of aggregate *relevant* items retrieved is 395 (or 0.378 × 1,044). By the same token, the aggregate *relevant* retrieval by the use of DD is 373 (or 0.335 × 1,112). From the second section of Table 6 in the same paper, the asymmetric relevant overlap ratio between TT and DD is 0.253, or 25.3% of the total relevant retrieval by TT are overlapped with DD. With the data for the relevant overlap ratio and the number of relevant items retrieved, one is able to compute the number of common relevant items between TT and DD (p) as 100 (or 0.253 × 395). Finally, the number of union relevant retrieval by the use of TT and DD (or t) is computed to be 668 (or 395 + 373 − 100). Hence, the remaining values of the contingency table can be easily filled as follows:

	Overlap	No overlap	Total
Relevant	p = 100	q = 568	t = 668
Not rel.	r = 37	s = 1314	u = 1351
Total	v = 137	w = 1882	n = 2019

Relevance odds

The relevance odds are computed as follows: overlapping items: 100/37 = 2.70; non-overlapping items: 568/1314 = 0.43; odds ratio: 2.70/0.43 = 6.25; Ln ratio: 1.83; standard error: 0.19; t-value: 9.75. In other words, the odds that overlaps from both TT and DD be judged "relevant" or "probably relevant" to nonoverlaps from these two fields are in excess of 6 to 1.

Using the same procedure, the contingency tables for all 21 possible combinations of pairs of search fields were filled. Table 1 shows the 21 cross product relevance odds ratios. Since, in every case, the t-value exceeds 2, all 21 odds ratios were statistically significant. Moreover, as all odds ratios were greater than unity, one may conclude that *the odds that overlap items from the use of any combination of two search fields judged either as definitely relevant or probably relevant are higher than those resulting from the use of any single field.* Specifically, one notes that the odds for overlap items are highest when common

Table 1. Odds ratios based on normal relevance

	Odds	ln	SE	t-Value	Significant?
DD AA	7.53	2.02	0.21	9.83	Y
DD TA	5.24	1.66	0.19	8.58	Y
DD DI	5.28	1.66	0.14	11.67	Y
DD ST	6.56	1.88	0.24	7.96	Y
DD TT	6.25	1.83	0.19	9.75	Y
DD II	6.45	1.86	0.18	10.33	Y
AA TA	2.06	0.72	0.12	6.09	Y
AA DI	3.27	1.19	0.14	8.46	Y
AA ST	2.77	1.02	0.12	8.34	Y
AA TT	6.03	1.80	0.18	10.03	Y
AA II	2.20	0.79	0.13	5.95	Y
TA DI	3.55	1.27	0.15	8.75	Y
TA ST	2.36	0.86	0.12	7.44	Y
TA TT	5.16	1.64	0.16	9.79	Y
TA II	3.13	1.14	0.14	8.17	Y
DI ST	4.71	1.55	0.17	9.15	Y
DI TT	5.61	1.72	0.17	10.40	Y
DI II	3.89	1.36	0.13	10.58	Y
ST TT	6.05	1.80	0.18	9.74	Y
ST II	3.19	1.16	0.15	7.59	Y
TT II	5.30	1.67	0.16	10.53	Y

items are retrieved from the use of assigned descriptors (DD) and from text words found in abstracts (AA).

In addition to the 21 cross product analyses presented in Table 1, the same procedure was used to extract the data from the 1982 paper to compute relevance odds using strong and weak relevance criteria. An additional 42 cross product analyses were conducted. The relevance odds are summarized in Tables 2 and 3.

RESULTS

The most important finding is that all 63 analyses produced statistically significant results, and the odds that an overlap item from the combination of any two search fields to be judged relevant are higher than any nonoverlap item in every pairwise overlap pos-

Table 2. Odds ratios based on strong relevance

	Odds	ln	SE	t-Value	Significant?
DD AA	8.51	2.14	0.29	7.38	Y
DD TA	7.46	2.01	0.25	8.08	Y
DD DI	5.63	1.73	0.20	8.73	Y
DD ST	9.67	2.27	0.32	7.05	Y
DD TT	7.36	2.00	0.25	7.87	Y
DD II	8.46	2.14	0.25	8.73	Y
AA TA	2.77	1.02	0.15	6.82	Y
AA DI	3.97	1.38	0.18	7.46	Y
AA ST	4.10	1.41	0.16	8.70	Y
AA TT	8.11	2.09	0.23	8.98	Y
AA II	2.97	1.09	0.17	6.44	Y
TA DI	5.23	1.65	0.18	8.99	Y
TA ST	3.21	1.17	0.15	7.81	Y
TA TT	8.32	2.12	0.21	10.27	Y
TA II	4.08	1.41	0.18	7.90	Y
DI ST	7.48	2.01	0.23	8.92	Y
DI TT	7.09	1.96	0.22	8.98	Y
DI II	5.55	1.71	0.17	10.08	Y
ST TT	8.56	2.15	0.25	8.73	Y
ST II	4.30	1.46	0.20	7.22	Y
TT II	7.84	2.06	0.20	10.17	Y

Table 3. Odds ratios based on weak relevance

	Odds	ln	SE	*t*-Value	Significant?
DD AA	4.47	1.50	0.26	5.82	Y
DD TA	5.69	1.74	0.24	7.27	Y
DD DI	3.63	1.29	0.18	7.19	Y
DD ST	6.62	1.89	0.31	6.19	Y
DD TT	4.38	1.48	0.23	6.39	Y
DD II	5.08	1.62	0.23	7.21	Y
AA TA	2.62	0.96	0.15	6.59	Y
AA DI	3.18	1.16	0.18	6.60	Y
AA ST	3.70	1.31	0.16	8.26	Y
AA TT	5.48	1.70	0.22	7.74	Y
AA II	2.73	1.00	0.16	6.09	Y
TA DI	4.57	1.52	0.18	8.46	Y
TA ST	2.96	1.08	0.15	7.43	Y
TA TT	7.19	1.97	0.20	9.65	Y
TA II	3.42	1.23	0.17	7.19	Y
DI ST	6.10	1.81	0.22	8.25	Y
DI TT	4.74	1.56	0.20	7.65	Y
DI II	4.40	1.48	0.16	9.17	Y
ST TT	6.01	1.79	0.23	7.70	Y
ST II	3.58	1.28	0.19	6.58	Y
TT II	5.72	1.74	0.19	8.98	Y

sible. It is noteworthy that if an item appeared in both printouts using any two fields, its odds of being judged relevant are at least twice as high than the use of any one field. Tables 1–3 provide strong evidence supporting the claim that items retrieved by the use of any two search fields are more likely to be judged relevant.

Among the 63 comparisons, the most impressive results were those found in the cases in which only those items judged 'definitely relevant' were considered as relevant.

Comparing Tables 1, 2, and 3, the odds that overlap items would be judged "definitely relevant" as opposed to "not relevant" were consistently higher for all 63 cases. The best performer belonged to the overlaps retrieved from both assigned descriptors (DD) and stemmed text words from titles and abstracts (ST). The odds for an overlap item to be judged definitely relevant increased by a factor of 9.67. In other words, the odds are nearly 10 to 1 for overlap items to be judged definitely relevant as opposed to not relevant when compared to those retrieved from either descriptors or text words from titles and abstracts alone. This is an astounding result in that these overlaps are 865% more likely to be judged definitely relevant. It is also worth noting again that the best performer involves the combined usage of both controlled vocabulary and free-text searching.

In an attempt to detect consistent patterns among the different pairwise overlaps, averages were computed for each field. The relevance odds data for the retrieval overlaps are arranged for each search field in Table 4a–c. For example, in Table 4a, when normal relevance is considered, the average odds for using descriptors assigned from the controlled vocabulary (DD) with any other search field is found to be 6.22. The average is considerably higher than 3.98 which is the average for free-text from abstracts (AA). Consistently, overlaps from the use of assigned descriptors (DD) and stemmed free-text words (ST) produced one of the top two highest relevance odds (Table 4a–c). Furthermore, the columns of data under DD in Table 4a–c all show that one can expect the odds to be at least 3.63.

For those interested in items of definite relevance, data from the column under (TT) in Table 4b shows that the odds for overlap items judged as definitely relevant resulting from the use of free-text terms from the title (TT) with any other field were at least 7 to 1 as compared with nonoverlap items. Similarly, the relevance odds associated with the use of assigned descriptor (DD) with any other field were also high (see under DD in Table 4b). Since DD is a subset of DI or a combination of assigned descriptors and identifiers, and if DI is excluded from consideration, the odds for overlaps from the use of assigned descriptors to be judged definitely relevant were also at least 7 to 1. Although the odds from all 63 pairwise overlaps were more than 2 to 1, using free-text from abstracts (AA) with

Table 4. Odds ratios for (a) normal, (b) strong, and (c) weak relevance

	DD	AA	TA	DI	ST	TT	II
(a) *Normal relevance*							
DD		7.53*	5.24	5.28	6.56	6.25	6.45
AA	7.53*		2.06†	3.27	2.77	6.03	2.20
TA	5.24	2.06†		3.55	2.36	5.16	3.13
DI	5.28	3.27	3.55		4.71	5.61	3.89
ST	6.56	2.77	2.36	4.71		6.05	3.19
TT	6.25	6.03	5.16	5.61	6.05		5.30
II	6.45	2.20	3.13	3.89	3.19	5.30	
Average	6.22	3.98	3.58	4.39	4.27	5.73	4.03
(b) *Strong relevance*							
DD		8.51	7.46	5.63	9.67*	7.36	8.46
AA	8.51		2.77†	3.97	4.10	8.11	2.97
TA	7.46	2.77†		5.23	3.21	8.32	4.08
DI	5.63	3.97	5.23		7.48	7.09	5.55
ST	9.67*	4.10	3.21	7.48		8.56	4.30
TT	7.36	8.11	8.32	7.09	8.56		7.84
II	8.46	2.97	4.08	5.55	4.30	7.84	
Average	7.85	5.07	5.18	5.83	6.22	7.88	5.53
(c) *Weak relevance*							
DD		4.47	5.69	3.63	6.62	4.38	5.08
AA	4.47		2.62†	3.18	3.70	5.48	2.73
TA	5.69	2.62†		4.57	2.96	7.19*	3.42
DI	3.63	3.18	4.57		6.10	4.74	4.40
ST	6.62	3.70	2.96	6.10		6.01	3.58
TT	4.38	5.48	7.19*	4.74	6.01		5.72
II	5.08	2.73	3.42	4.40	3.58	5.72	
Average	4.98	3.70	4.41	4.44	4.83	5.59	4.16

*Indicates the strongest relevance odds.
†Indicates the weakest relevance odds.

any other field for strong and weak relevance resulted in the lowest odds. Free-text from both abstracts and titles (TA) which contains AA, had the lowest odds for normal relevance.

From the extracted overlap data, percentages of overlaps for every other pair of search fields were computed. Table 5 shows the average percentages of retrieval overlaps across all 21 possible pairwise comparisons using the 7 search fields. Of the items retrieved for all 84 queries, an aggregate of 8.85% were overlap items. Even fewer were relevant overlaps: 5.3% were relevant overlaps, and 3.55% were overlaps of definite relevance. However, 63% of those overlap retrieval were judged relevant, and 42% were of definite relevance. Among those items with any degree of relevance, 67.38% were found to be definitely relevant. These findings confirm the observation by the Syracuse researchers that the average overlap is highest when only the most relevant are included, and is lowest when all documents, whether relevant or not, are considered.

Table 5. Mean percentages of overlaps across 21 pairwise comparisons of 7 document representations

	Definitely relevant retrieved	All relevant retrieved	Overlap retrieved
Based on retrieved set	3.55%	5.30%	8.85%
Based on overlap set	42.47%	63.00%	–
Based on relevant overlap set	67.38%	–	–

A caveat was noted in the 1982 paper: although there are unique strengths for each field, the overlap measures of these fields may be confounded by the use of different searchers. From the results of this analysis, another note of caution may be added. The judgment of "probable relevance" may be suspect. In every pairwise overlap category, when strong relevance is used as the criterion, the odds are higher than those computed for either normal or weak relevance. The criterion of strong relevance necessitates the elimination of the category of items judged to be of probable relevance. On the other hand, in comparing the same category using normal and weak relevance as criteria, inconsistent and mixed results emerged. The odds in 13 out of the total 21 categories were lower when normal relevance was used. Tables 4a and 4c show that including "probably relevant" items in either the all-inclusive "relevant" when normal relevance is used, or all-inclusive "not relevant" for weak relevance may be less reliable than using only those rated as "definitely relevant" as relevant as in the case of strong relevance. Therefore, a much more confident conclusion is that the odds that overlaps derived from two parallel searches using different fields be of *definite relevance* are considerably higher than those retrieved items from any one field alone. The score of '2,' a stand-in for "probably relevant," seems to be a much less reliable indicator of relevance.

CONCLUSION

The major finding of this study is that the odds that an overlap item retrieved from both printouts of any two search fields will be judged "relevant" or "probably relevant" as opposed to not relevant is at least twice as much as if it were retrieved from only one field. Similar results hold even if a more stringent or a more relaxed criterion of relevance were used. Regardless of the search fields chosen, the relevance odds were higher in overlap items than in the nonoverlap items. Statistically significant results were found in all 63 possible pairwise overlap comparisons.

Overlaps resulting from the use of assigned terms from the controlled vocabulary and truncated words from the titles and abstracts have odds of 19 to 2, when compared with retrieved items from only one of the two fields. Certainly, these results give credence to the general practice of supplementing descriptor search terms with free-text used by searchers. In particular, if one is interested in obtaining items of definite relevance, assigned descriptors should be used with one other field. One could expect the odds to be no less than 5 to 1 in favor of overlaps.

Katzer and colleagues observed that the size of overlap is directly related to the level of relevance of the data sets compared (Katzer *et al.*, 1982). Table 6 shows that all overlaps are low. The overlap is higher if only the most relevant items are included. If only definitely relevant items are considered, the mean overlap is 35% without any pair of fields

Table 6. Average asymmetric pairwise overlaps among 7 document representations

Definitely relevant retrieved	All relevant retrieved	All retrieved
AA 0.35	0.29	0.18
TT 0.35	0.29	0.16
TA 0.34	0.29	0.18
ST 0.36	0.27	0.15
II 0.34	0.27	0.17
DI 0.39	0.32	0.18
DD 0.34	0.28	0.13
Mean 0.35	0.25	0.14

*Data taken from Table 6 from "A Study of the Overlap Among Document Representations," by Katzer, J. *et al.* (1982). *Information Technology: Research and Development*, *2*, 261–274.

exceeding 40%. Whereas, if all relevant items are included, the mean is 25% with none exceeding 33%. In the three graded levels of relevance, the relevance odds for the same pair of search fields are higher for strong relevance in all 21 possible overlaps.

In a recent article, Kantor (1992) attempted an explanation of the complementary nature of different searches conducted by different searchers on identical topics. The underlying principle of the low degree of overlap is still not well understood and more research is needed. Yet, it is worth noting that if overlap sets are more likely to contain relevant items, it could be regarded as a practical quality filter technique for online searching.

REFERENCES

Kantor, P.B. (1992). Two heads are better than one: The potential of data fusion concepts for improvement of online searching. In *Proceedings of the 13th National Online Meeting*, New York (pp. 147–151).

Katzer, J., McGill, M.J., Tessier, J.A., Frakes, W., & Das Gupta, P. (1982). A Study of the overlap among document representations. *Information Technology: Research and Development, 2*, 261–274.

McCain, K.W., White, H.D., & Griffith, B.C. (1986). Test retrieval as a measure of system performance: MEDLINE and the medical behavioral sciences. *Proceedings of the American Society for Information Science, 23*, 199–203.

McCain, K.W. (1989). Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval overlaps and novelty distribution. *Journal of the American Society for Information Science, 40*, 110–114.

McCain, K.W., White, H.D., & Griffith, B.C. (1987). Comparing retrieval performance in online data-bases. *Information Processing and Management, 23*(6), 539–553.

McGill, M., Koll, M., & Norealt, T. (1979). *An evaluation of factors affecting document ranking by information retrieval systems.* (Final Report for Grant NSF-IST-78-10454). Washington, DC: National Science Foundation, pp. 1–110.

McKinin, E.J., & Sievert, M.E. (1989). A comparison of full-text and abstracts for information retrieval in clinical medicine. In *Proceedings of the 19th National Online Meeting*, New York (pp. 295–301). Medford, NJ: Learned Information.

McKinin, E.J., Sievert, M.E., Johnson, E.D., & Mitchell J.A. (1991). The MEDLINE/full text research project. *Journal of the American Society for Information Science, 42*(4), 297–307.

Pao, M.L. (1986). Comparing retrievals by keywords and by citations. In *Proceedings of the 7th National Online Meeting*, New York (pp. 341–346). Medford, NJ: Learned Information.

Pao, M.L. (1993). Term and citation searching: A field study. *Information processing and Management, 29*(1), 95–112.

Pao, M.L., & Worthen, D.B. (1989). Retrieval effectiveness by semantic and pragmatic relevance. *Journal of American Society for Information Science, 40*(4), 226–235.

Ro, J.S. (1988). An Evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. I: On the effectiveness of full-text retrieval. *Journal of American Society for Information Science, 39*(1), 73–78.

Saracevic, T., Kantor, P., Chamis, A.Y., & Trivision, D. (1988a). A study of information seeking and retrieving: I. Background, and methodology. *Journal of the American Society for Information Science, 39*(3), 161–176.

Saracevic, T., & Kantor, P. (1988b). A study of information seeking and retrieving: III. Searchers, searches, and overlaps. *Journal of the American Society for Information Science, 39*(3), 197–216.

Tenopir, C. (1985). Full-text database retrieval performance. *Online Review, 9*, 149–164.

White, H.D., Griffith, B.C., Cowen, J.A., Selinger, N.E., & Steere, D.T. (1984, January). *Evaluation of the National Library of Medicine's programs in the medical behavioral sciences. Quality of indexing: The development and testing of a behavioral science literature.* (Report to the NLM). Philadelphia: Drexel University.

White, H.D. (1989, December). Toward automated search strategies. In *Proceedings of the 13th International Online Meeting, London* (pp. 33–47). Oxford, UK: Learned Information.