A STUDY OF THE OVERLAP AMONG DOCUMENT REPRESENTATIONS [1]

Syracuse University, School of Information Studies,
113 Euclid Avenue, Syracuse, N.Y. 13210, USA

ABSTRACT

Most previous investigations comparing the performance of different representations have used recall and precision as performance measures. However, there is evidence to show that these measures are insensitive to an important difference between representations. To explain, two representations may perform similarly on these measures, while retrieving very different sets of documents. Equivalence of representations should be decided on the basis of similarity in performance and similarity in the documents retrieved. This study compared the performance of four representations in the PsycAbs database. In addition, overlap between retrieved sets was also computed where overlap is the proportion of retrieved documents that are the same for pairs of document representations. Results indicate that for any two representations considered, performance values differed slightly while overlap scores were also low, thus supporting the evidence that recall and precision as performance measures mask differences between the sets of retrieved documents. Results are interpreted to propose an optimal ordering of the representations and to examine the contribution of each representation given this combination.

1 This is a preliminary report of work conducted by Jeffrey Katzer, Judith Tessier, William Frakes and Padmini Das-Gupta at Syracuse University, School of Information Studies (1981-1982). The presentation at SIG/IR was made by Padmini Das-Gupta.

INTRODUCTION

This paper presents some of the second phase data and results of the Document Representation Overlap Study[2] performed at Syracuse University during the period March 1981 to February 1982. The two phases of the study correspond to two data bases used. The two phases also differ in the number of representations and the number of intermediaries used. The second phase investigation was designed to compare four document representations in a portion of the PsycAbs data base using performance and overlap measures.[3] This paper discusses the second phase observations and results pertaining to the overlaps among sets of documents retrieved on different representations.

Past studies show that representations (such as free-text term, or descriptor phrase), when examined using recall and precision performance measures, have not shown consistent results. Further, earlier studies provide evidence that performance measures mask systematic differences among representations. Specifically, different representations result in the retrieval of different items. For example, the study by McGill, et al, 1979 (ref. 3), compared documents retrieved using free-text and controlled terms in a portion of the ERIC data base. Users provided queries which were searched and relevance judgments obtained. Thirty-three of the queries were selected for a study of overlap. When each of the intermediaries searched both document representations, the average overlap was only 14%. Other queries were searched by intermediaries using different representations. In this situation, the average overlap dropped to 5%. Both of these figures are surprisingly low indicating that users retrieve quite different sets of documents when the free and controlled representations are used.

2 Research for this study was supported in part by the National Science Foundation, Division of Information Science & Technology, under Grant 70-21468.

3 The first phase study examined seven representations using seven intermediaries in the INSPEC data base. Ref. 1 and 2.

The phase of the study reported in this paper attempted to duplicate the results in the literature using four representations and the data base PsycAbs.

METHODOLOGY

Variables

The key experimental or independent variable was the representation used in searching the data base. The following four representations were used.

DD - Descriptor terms chosen by an indexer; a controlled vocabulary.

AA - Free-text words from the abstract; trivial words excluded.

TT - Free-text words from the title; trivial words excluded.

II - Free-text phrases chosen by the indexer.

The major dependent or criterion variables were overlap measures. These measures were operationalized as follows:

Assymetric-Overlap: For two representations i and j, this measure is computed by dividing the number of documents retrieved by both representations by the number retrieved by one of the representations. If R_i and R_j are the sets of documents retrieved by representation i and j, then the assymetrical-overlap measure can simply be given as

$$Aij = \frac{n [R_i \cap R_j]}{n [R_i]}$$

where 'n' is the counting operator. Seen this way, assymetric-overlap is the conditional probability of retrieval using representation j given that the data base is restricted to those retrieved by representation i.

Symmetric-Overlap: For two representations i and j, this measure is computed by dividing the number of documents retrieved in common by both representations by the total number of different documents retrieved by either. Or more formally, it is the number of documents in the intersection of the two representations divided by the number retrieved by the union on those representations.

$$Sij = \frac{n [R_i \cap R_j]}{n [R_i \cup R_j]}$$

Union-Overlap: For two representations i and j, this measure is computed by dividing the number of documents retrieved by either of the representations by the number of documents retrieved by all of the representations. In our case r=4.

$$Uij = \frac{n [R_i \cup R_j]}{n [R_i \cup R_j \cup \ldots\ldots \cup R_r]}$$

Thus the union-overlap is more of a recall ratio for a combination of representations. It can be extended to combinations of more than two representations by expanding the numerator.

Different versions of these dependent variables were computed; they differed in terms of the stringency of the relevance criterion, where relevance was determined by the requestor. A four point relevance continuum was used from 1 (definitely relevant to 4 (definitely not relevant).

The overall design can be described as a factorial design containing sixteen cells (four searchers by four representations). Each of the 52 queries was searched under all sixteen combinations. This design, required that each intermediary use all representations when searching a query.

Procedure

Permission was granted by the PsychInfo Use Service to use a subset of their 1980 data base whose printed counterpart is Psychological Abstracts. Approximately 12,000 documents were included in the subset. Each document consisted of a series of bibliographic citation fields, the abstract and the indexing information.

Altogether, 45 individuals served as users and submitted a total of 52 search requests written in natural language to the study. Users were from Syracuse University and other institutions, with information needs related to the contents of the data base.

Four experienced search intermediaries were employed for this study. A three hour training session was held. Each intermediary was required to submit two practice searches to the system. Searches were conducted using DIATOM, the Boolean online retrieval system designed to simulate most of the features of DIALOG.

The search procedure was started with some preliminary screening of the search requests obtained for appropriateness to the data base and on-line searching. Each intermediary was given a photocopy of the search request with instructions to conduct four searches for each query; one for each representation. Computer programs within the DIATOM system imposed a random ordering of the representations used. Intermediaries were instructed to carry out high recall searches.

After a query was completely searched: sixteen times, once for each searcher-representation combination, the retrieved document set was merged into a single listing and placed in reverse chronological order. This listing consisted of citations and abstracts of the retrieved documents (if more than 200 documents were retrieved, a random sample of 200 was used). No clue was present to indicate which intermediary or representation retrieved the document.

Two copies of this listing were produced. Both copies were sent to the user for relevance judgements. A four point scale was used from '1' indicating definitely relevant to '4' indicating definitely not relevant. All documents in the listing were judged by the user for relevance.

RESULTS AND DISCUSSION

Our initial concern was to determine if the results from this study repeated the pattern noted earlier: relatively little difference in performance among the representations coupled with relatively little overlap. It is apparent that these results do repeat the pattern observed in other studies. The major conclusion made from the performance analysis is that though some performance measures are significantly different, none of the differences exceed .12 -- which is clearly within the range of values reported in the literature. Consult the NSF report (ref. 2) for a detailed examination of the performance data. The overlaps range from a low of about 23% to a high of about 27%: these correspond to the earlier results.

The remainder of this section will describe in detail the analysis of the overlap data.

Analysis of Overlaps

The simplest analysis of overlaps is pairwise, comparing each representation with every other representation. Tables 1-3 report the overlaps for the data. Each table contains three overlap analysis: (1) most relevant documents, (2) all relevant documents, and (3) all documents retrieved. In these tables, a high value indicates greater overlap and therefore less of the 'second' representation.

It can be seen that the pairwise overlaps decrease as the number of documents under consideration increase. That is, the average overlap is highest when only most relevant documents are considered, it is lowest when all retrieved documents are included. A second general finding is that the overlap figures are lowest when overlap is defined symmetrically, they are the highest for union-overlap. This, of course, is a function of the definition of the three measures of overlap.

The major finding in these data is that the overlaps are quite small as indicated by the averages. For example, the highest symmetric overlap among the relevant documents is only about one-third -- .363 between AA and II.

A possible explanation for the size of overlaps is searcher difference. Analysis of variance computations (see the final report ref. 2) showed that searcher effects occasionally account for significant portions of the variance. However, the ranking study, conducted by McGill et al (ref. 3), casts doubt on the contention that searchers are the sole or major cause of the low amount of overlap. In the ranking study, overlaps between different representations searched by the same searcher only equalled 14% for retrieved documents. That figure certainly falls in the range of values reported here. Furthermore, the design required that each intermediary search each query under all representations: the overlap results were, at best, moderate.

It can be seen from the symmetric measures, in Table 1 that the maximum difference in overlaps does not exceed 0.10. Also, the free-index phrases (II) show a tendency to share more relevant documents with titles and abstract fields than with the descriptor field -- although the size of this overlap is still quite small.

The assymetric measures indicate the proportion of documents that would have been retrieved 'anyway' -- that is, by the other representations. For example, Table 2 reports an assymetric overlap of .378 between DD and II for the most relevant documents. This can be interpreted as follows: of all the documents retrieved by the descriptor representation, approximately 38% of them can also be retrieved by the free-index phrases. Table 2 provides both row and column average figures (the other tables are symmetrical and a single set of averages suffices). A useful interpretation of the difference between row and column averages for a single representation can be given in terms of the sequence the representations are used in searching. The averages of the columns of numbers (presented along the bottom of the table) can be interpreted in terms of being used 'first' in the search process. Given a single representation (indicated by the column heading), the average at the bottom indicates the proportion of documents retrieved by this representation that could also be retrieved by other representations. The averages presented in the right column are understandable in terms of being used 'last' in the search process. Given retrieved documents from other representations, the row average for a given representation indicates its effect if searching were resumed using it alone -- the lower the average, the more the new representation will contribute.

Given this distinction between implementing or using a representation 'first' or 'last', the assymetric overlaps (in Table 2) identify the descriptors as the best choice for 'first' and for 'last' representations.

Union-overlaps give an estimate of the combined effect of two or more representations. Because the numerator of these pairwise union-overlaps includes all distinct documents (in the appropriate version) retrieved by the two representations, the union-overlaps will have higher values than comparable figures for the symmetrical and assymmetrical overlaps.

The union-overlap figures in Table 3 show that most pairs of representations achieve higher than 50% recall levels. The combination of descriptions and abstracts given over 80% of the most relevant documents and over 75% of all documents retrieved.

Union-overlaps are one way to explore the 'marginal utility' or the 'value added' of additional representations. Table 2 provides only pairwise overlaps. The extension to more than two representations is necessary in order to get overall

conclusions. The next section of this paper takes this approach.

Descriptive Models of Overlap

Going beyond pairwise overlaps, the question of concern here focuses on the relationship among all of the representations: what is the optimum ordering of representations? That is, if a retrieval environment were limited to a single representation, which one would it be? If a second one could be added, which of the remaining representations contributes the most over and above the effect of the first representation? A third representation could be added over and above the first two, and so on.

The most sensible measure to use in answering this question is based on the union-overlap. The result of this analysis is a model depicting an ordering of the representations. Table 4 presents this model. Specifically, given the four representations, the model identifies descriptors as the best 'first' representation and abstracts as the best 'second' representation when descriptors is the first representation. Titles rank the best 'third' representation by contributing most over and above descriptors and abstracts, while identifiers is selected as the last representation given the first three.

Such a model, if consistent, would allow a searcher to know which combination of fields would be most likely to retrieve relevant documents. Such models would also point to obvious economies in the design and operation of retrieval systems.

An interesting result that shows up from Table 4 is that the order does not change as a function of relevance stringency. What appears to be highly consistent is the cumulative increase in the percentage of relevant documents accounted for as each additional representation is included. This similarity may simply be due to the fact that the models are based on highly interrelated data.

The overlap among document representations can also be viewed from the perspective of a representation's 'unique' contribution. For a given representation, what documents does it contribute to the relevant retrieved that were not retrieved under any other representation? The question is equivalent to the observed improvements in the model when the representation is the last one entered. Table 5 reports the effect of each representation, assuming the representation entered the model first or last. These are the maximum and minimum incremental improvements for each representation. The 'unique' effect of each representation is reported as the minimun contribution.

The lack of overlap among representations is again evident in the unique percentages. Given the data base and the four representations, the fourth representation can contribute a sizeable number of additional documents -- approximately 25% for the DD representation.

One final indicator of the lack of overlap among document representations is the sum of the unique contributions (Table 5). This total is about 58%. Thus the amount of overlapping documents is about 42%.

To conclude, the four representations examined differed little in performance scores. Further, the overlap measures between pairs of representations were low, therefore supporting the earlier literature where performance measures were observed to mask differences among the sets of documents retrieved by different representations. Low overlaps were also indicated by the unique contribution figures for each representation.

Within the constraints of this study, a closer examination of the overlap data revealed an optimal ordering of the representations in terms of relevant documents retrieved. This study provides evidence for the importance of overlap measures in the comparison of representations. In addition, it identifies the descriptor field as an effective representation (within the PsychAbs data base) with respect to the representations examined.

REFERENCES

1. Katzer, Jeffrey, et al. A Study of the Overlap Among Document Representations. Information Technology, 1982. 1, pp. 261-273.

2. Katzer, Jeffrey. A Study of the Impact of Representations in Information Retrieval Systems. Final report for Grant NSF-IST-79-21468 to the National Science Foundation, July 1982.

3. McGill, Michael J. et al. An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems. Final Report For Grant NSF-IST-78-10454 to the National Science Foundation, October 1979.

TABLE 1

Symmetric Pairwise Overlaps

	II	DD	AA	TT	AVG *
Version - Most Relevant					
II	1.000	0.289	0.363	0.351	0.334
DD	0.289	1.000	0.273	0.264	0.275
AA	0.363	0.273	1.000	0.277	0.304
TT	0.351	0.264	0.277	1.000	0.297
Version - All Relevant					
II	1.000	0.269	0.319	0.328	0.305
DD	0.269	1.000	0.233	0.234	0.245
AA	0.319	0.233	1.000	0.256	0.269
TT	0.328	0.234	0.256	1.000	0.273
Version - All Documents					
II	1.000	0.199	0.182	0.215	0.199
DD	0.199	1.000	0.150	0.159	0.169
AA	0.182	0.150	1.000	0.127	0.153
TT	0.215	0.159	0.127	1.000	0.167

*Averages were computed with the diagonal element omitted.

TABLE 2

Assymetric Pairwise Overlaps

	II	DD	AA	TT	AVG *
Version - Most Relevant					
II	1.000	0.378	0.469	0.551	0.466
DD	0.552	1.000	0.452	0.551	0.518
AA	0.616	0.407	1.000	0.536	0.520
TT	0.491	0.336	0.364	1.000	0.397
AVG*	0.553	0.374	0.428	0.546	
Version - All Relevant					
II	1.000	0.357	0.437	0.523	0.439
DD	0.524	1.000	0.413	0.500	0.479
AA	0.54	0.348	1.000	0.485	0.458
TT	0.468	0.305	0.351	1.000	0.375
AVG*	0.511	0.337	0.401	0.503	
Version - All Documents					
II	1.000	0.289	0.264	0.394	0.316
DD	0.39	1.000	0.256	0.364	0.337
AA	0.371	0.267	1.000	0.307	0.315
TT	0.321	0.220	0.178	1.000	0.240
AVG*	0.361	0.259	0.233	0.355	

* Averages were computed with the diagonal element omitted.

** The representations in the columns form the denominator of the overlap measure.

TABLE 3

Union Pairwise Overlaps

	II	DD	AA	TT	AVG *
Version - Most Relevant					
II	0.377	0.719	0.640	0.528	0.629
DD	0.719	0.550	0.821	0.701	0.747
AA	0.64	0.821	0.495	0.651	0.704
TT	0.528	0.701	0.651	0.336	0.627
Version - All Relevant					
II	0.368	0.715	0.624	0.525	0.621
DD	0.715	0.539	0.806	0.704	0.742
AA	0.624	0.806	0.454	0.624	0.685
TT	0.525	0.704	0.624	0.329	0.618
Version - All Documents					
II	0.314	0.616	0.640	0.469	0.575
DD	0.616	0.424	0.753	0.587	0.652
AA	0.640	0.753	0.442	0.619	0.671
TT	0.469	0.587	0.619	0.256	0.558

* Averages were computed with the diagonal element omitted.

TABLE 4

Representations Ordered
by Incremental Improvement

Order	1st	2nd	3rd	4th
Most Relevant				
Representation	DD	AA	TT	II
Cum. No. Docs.	339	506	573	616
Cum. Percentage	.550	.821	.930	1.000
All Relevant				
Representation	DD	AA	TT	II
Cum. No. Docs.	871	1302	1489	1615
Cum. Percentage	.539	.806	.922	1.000

TABLE 5

Maximum and Minimum Contributions of Four Representations

Repr.	Maximum Contribution* No.Docs.	Percent**	Minimum Contribution* No.Docs.	Percent**
Most Relevant				
AA	310	.475	112	.172
DD	339	.520	158	.242
II	229	.351	42	.064
TT	210	.322	50	.077
				.555
All Relevant				
AA	728	.440	286	.173
DD	870	.526	429	.259
II	579	.350	120	.072
TT	518	.313	131	.079
				.583

*Maximum contribution is the effect of that representation alone--either it is the sole representation in the data base or it was used (entered) first, before the others are used. Maximum contribution is therefore equivalent to micro-recall.

Minimum contribution is the "unique" effect of that representation after all documents retrieved by the other three representations have been removed; thus, it can be considered to have entered the search process last.

**Percentages are based on all documents retrieved by all representations in each category. Here the numbers are 652 for most relevant and 1653 for all relevant.