Information Technology: Research and Development (1982), *2* (261–274)

A STUDY OF THE OVERLAP AMONG DOCUMENT REPRESENTATIONS*

J. KATZER, M. J. McGILL, J. A. TESSIER, W. FRAKES AND P. DASGUPTA

Syracuse University, School of Information Studies, 113 Euclid Avenue, Syracuse, New York 13210, USA

(*Received 4 January 1982, revised 8 June 1982*)

ABSTRACT

In the past 15 years there have been a great many investigations of the comparative performance of different document representations. It is misleading, however, to consider representations equivalent (even though they perform approximately equally in the aggregate) unless they also retrieve a very similar set of documents. This study compared seven document representations in the INSPEC database. Performance measures (recall and precision) were computed in addition to measures of 'overlap' —the proportion of retrieved documents that were the same for each pair of representations. The results indicate slight differences in performance measures (no representation exceeded another in performance by more than 0.18). Low average overlaps were found between all pairs of representations, even between those that should have retrieved similar document sets such as 'abstract' and 'title–abstract'. The findings are interpreted in terms of optimal combinations of document representations and the contribution each representation makes to the combination.

1. INTRODUCTION

For some time now, researchers have been comparing the relative effectiveness of different document representations. The results of numerous investigations of this question (e.g., title versus abstract or free-text terms versus descriptor terms) conducted in the past 15 years or so are equivocal.† First of all, the studies do not

* A preliminary version of this paper was presented by Mr Frakes at the annual meeting of the American Society for Information Science in Washington, DC in October 1981. The complete report of this investigation is in process at the ERIC Clearinghouse, Syracuse University, Syracuse, New York 13210.

† There are too many studies contributing to this set of conclusions to be reviewed here. For illustrative purposes, Cleverdon (1967), Salton (1968), Keen (1973) and McGill *et al.* (1979) report no sizeable differences among the representations they examined. On the other hand, results from the second Cranfield Project, and from studies by Sparck Jones and Jackson (1970), Hersey *et al.* (1971), Salton (1973) and Sparck Jones (1974), indicated differences in average levels of performance.

always agree—even when comparing the same representation. Second, no representation has been found to perform high enough on all performance criteria (e.g., recall and precision) that it can be singled out. Third, even when differences have been found between representations, these differences have, for the most part, been small.

What is known, however, is that combinations of representations perform better than do representations taken singly—suggesting that the redundancy or overlap among the representations is not total. That is, different representations retrieve different subsets of the collection. One of the more recent studies supporting this assertion was conducted by Williams (1977). She computed the overlap among five different document representations in a random sample of 50 documents taken from *Chemical Abstracts*. No queries were obtained from users; rather representations were compared for matching terms. The results gave the degree of uniqueness or lack of overlap among representations. Title, for example, is claimed to be an important representation for retrieval because an average of two title terms per document did not appear in other representations. Smith (1979) provided some indication of the overlap among seven document representations in a portion of the INSPEC database. No users were employed; a random sample of 35 documents was selected and treated as queries. None of the average conditional probabilities (measures of asymmetrical overlap) exceeded 0.5, meaning that different document representations tended to retrieve different documents. A third study (McGill *et al.*, 1979) compared documents retrieved using free text and controlled terms in a portion of the ERIC database. 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 per cent. Other queries were searched by intermediaries using different representations. In this situation, the average overlap dropped to five per cent. Both of these figures are surprisingly low indicating that users retrieve quite different sets of documents when the free and controlled representations are used.

While many other studies of the effectiveness of combined representations have been conducted,* the overall conclusions are somewhat limited in their generality:

1. Typically, these studies compared very few representations—often only two.
2. Many of these studies used exceedingly small databases—occasionally fewer than ten documents or queries.
3. Frequently, these same studies examined the overlap among document representations without any consideration of document relevance—giving the likelihood of retrieving the same document using different representations, but unable to determine if that likelihood is comparable for relevant and non-relevant documents.

In contrast, the study reported here compared seven representations, formed for each of 12000 documents, in terms of both overlap and performance. The overlap and performance measures are then used together to identify those few representations which jointly will retrieve the most unique documents and the most unique relevant documents.

* Cattley *et al.* (1966), Bottle (1970), Barker *et al.* (1972), Fisher and Elchesen (1972), Maloney (1974), Byrne (1975), Markey and Atherton (1979), Mansur (1980), Chapman and Subramanyam (1981) and Waldstein (1981).

2. METHODOLOGY

A subset of the INSPEC database was used in this study: 12000 documents from the September–December 1979 issues of *Computer and Control Abstracts.* Eighty-four queries posed by 69 users were searched on this database by experienced and trained professional intermediaries. For each query, searchers were assigned a representation and told to construct a 'high recall' search acting as if the documents had that one representation. Queries, representations and intermediaries were balanced so that each query was searched under each representation by one of the seven different intermediaries. DIATOM, a system that simulates most of the features of DIALOG, was used to carry out the actual searches.* The seven retrieved document sets for each query were merged into a single non-redundant list and placed in reverse accession number order. No clue was present which indicated either the intermediary or the representation that retrieved each document. The users then judged each retrieved document for relevance using a four point scale (1—definitely relevant; 2—probably relevant; 3—probably not relevant; 4—definitely not relevant).

The key experimental or independent variable was the representation used in searching the database. Seven representations were chosen:

1. TT—free-text terms from title (trivial words excluded).
2. AA—free-text terms from abstract (trivial words excluded).
3. DD—descriptor terms (controlled vocabulary terms assigned by an indexer).
4. II—identifier terms chosen by indexer from the document (free-text terms).
5. TA—free-text terms from title and abstract (combining TT and AA).
6. DI—indexer selected terms (combining DD and II).
7. ST—stemmed free-text terms from title and abstract.

The major dependent or criterion variables were performance measures (recall and precision) and measures of overlap. The total numbers of documents retrieved were also analyzed. These measures were operationalized as follows.

2.1 Recall

The recall ratios were formed by dividing the number of relevant documents retrieved by each representation by the total number of relevant documents retrieved by all seven representations. Two versions of recall were computed: Recall-1 only counted those documents rated '1' as relevant, Recall-2 used all relevant documents (rated '1' or '2' on the four-point scale).

2.2 Precision

The precision ratio was formed by dividing the number of relevant documents retrieved by each representation by the total number of documents retrieved by that representation. Two versions of precision were computed—depending upon whether 'definitely relevant' or all relevant documents were used in the numerator.

* There are a few major differences between DIATOM as implemented for this study and DIALOG:

1. DIATOM includes a routine for finding word stems.
2. The adjacency operator could not be used with stemming.
3. Adjacency at times ran very slow; the field operator was an available alternative.

2.3 Total retrieved

This measure is simply the number of documents retrieved by each representation; it is the denominator of the precision ratio. It was included because it is an indication of user effort required to read the output from the system.

2.4 Asymmetric-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 representations i and j, then the asymmetrical-overlap measure can simply be given as

$$A_{ij} = \frac{n(R_i \cap R_j)}{n(R_i)}$$

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

Three versions of asymmetrical-overlap were computed. The versions differ depending upon which set of retrieved documents was included:

1. Only those definitely relevant.
2. All relevant.
3. All documents retrieved.

2.5 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 seven representations.

$$U_{ij} = \frac{n(R_i \cup R_j)}{n(R_i \cup R_j \cup R_k \cup \ldots \cup R_0)}$$

Thus, the union-overlap can be viewed as a recall ratio for a combination of representations. It can be extended to combinations of more than two representations by expanding the numerator. As in the case of the asymmetrical-overlap measure, there are three versions of the union-overlap depending upon the document set included.

The overall design can be characterized as a 7×7 latin square replicated 12 times (see, for example, King and Bryant, 1971: 301–302). The measures of recall, precision and total retrieved are analyzed using standard analysis of variance (AOV) computations. The design and the analysis control for extraneous variables; separate effects for representations, intermediaries, and if desired, replications can be identified. Approximately 10 per cent (66) of the precision results had to be excluded from the analysis because no documents were retrieved for a given query under a given representation. Fourteen queries had to be excluded from all Recall–1 analyses, and seven from the Recall–2 analyses, because in each situation no relevant documents were retrieved.

The overlap measures may have been adversely affected by the latin square

design. Because each pair of representations for a given query was searched by different intermediaries, there is a possibility that the overlap measures confound representations with intermediaries.

3. RESULTS AND DISCUSSION

Descriptive summary statistics for the five performance measures are presented in Table 1. The means were tested for statistically significant differences (see Tables 2–4 for the AOV summary tables).* Tukey's procedure used in these tables is described in Kirk (1968: 88–90). Representations differed significantly in the Recall–1, Recall–2 and total-retrieved scores. The bottom of Table 1 indicates that descriptors (DD) and titles (TT) perform rather poorly as representations on the recall measures, while identifiers (II) and title-abstracts (either TA or ST) perform much better.

Even though no pairs of representations differed significantly in either precision measure, it is useful to include some consideration of precision into these findings. Considering all five measures, the descriptor (DD) representation performs uniformly poorly on the recall and precision measures while title–abstract (TA) performs reasonably well on them—though not as strongly as DD's negative performance. Interestingly, the free-text words assigned by indexers (II) perform moderately well over all five measures. Stemming (ST) which would tend to increase the total number retrieved performs quite well on the recall measures, but poorly on the precision measures. The title representation (TT) shows the opposite pattern—high on the precision measures and low for recall. The other representations fluctuate quite a bit over the five measures.

The recall and precision means given in Table 1 are the average of individual ratios—each query contributed equally to the final average. Another way (see Salton, 1971: 71, for distinction between micro- and macro-recall and precision) to compute the average performance values is to compute the ratio last. For example, for Recall–1, sum the number of relevant documents retrieved from all 70 queries using a particular representation and divide this total by the number of relevant documents retrieved from all 70 queries using all seven representations. This is a more conservative approach and these values are presented in Table 1. This approach is useful, however, because the unique contribution of single (perhaps atypical) queries is removed. The average values computed in this manner are presented in Table 5. There are several parallels between the patterns in the two tables. Again, the II representation performs well on all four measures. Descriptors (DD) still show an overall poor performance and title–abstract (TA) performs well (though the similarity is weakened in the Precision–2 measure). Titles (TT) have the same pattern here as in Table 3, while stemming (ST) is not quite as good in the recall measures and is just as poor in the precision measures.

Turning to overlaps, the simplest analysis is pairwise, comparing each representation with every other representation. Tables 6 and 7 contain the pairwise overlaps for asymmetrical and union overlap. Each table reports the overlap for relevant documents (only those judged a '1', and those judged a '1' or a '2') and for all documents.

As might be expected, the pairwise overlaps decrease as the number of documents

* The other AOV summary tables are not included here to save space. They are part of the full report of this investigation.

Table 1. Means and standard deviations by representations*

Representation	*Recall-1*	*Recall-2*	*Precision-1*	*Precision-2*	*Total retrieved*
DD (descriptor)	0.229	0.200	0.173	0.336	13.238
	(70)	(77)	(62)	(62)	(84)
	0.319	0.257	0.260	0.330	15.824
AA (abstract)	0.365	0.270	0.197	0.352	17.488
	(70)	(77)	(77)	(77)	(84)
	0.314	0.241	0.255	0.315	16.850
TA (title and abstract)	0.404	0.290	0.224	0.352	18.583
	(70)	(77)	(78)	(78)	(84)
	0.317	0.236	0.286	0.318	16.245
DI (descriptor and identifier)	0.330	0.284	0.221	0.361	16.369
	(70)	(77)	(75)	(75)	(84)
	0.328	0.284	0.270	0.300	16.166
ST (stemmed title and abstract)	0.392	0.317	0.188	0.338	19.833
	(70)	(77)	(81)	(81)	(84)
	0.352	0.263	0.231	0.291	15.814
TT (title)	0.273	0.205	0.264	0.422	12.429
	(70)	(77)	(70)	(70)	(84)
	0.292	0.207	0.335	0.370	13.744
II (identifier)	0.339	0.321	0.218	0.403	16.131
	(70)	(77)	(79)	(79)	(84)
	0.323	0.276	0.282	0.334	15.181
Minimum difference between means that are significantly different at 0.05†	0.133	0.106	—	—	5.450
Pairs of representations that differ	DD<TA DD<ST DD<AA	DD<II DD<ST TT<II TT<ST	none	none	DD<ST TT<ST TT<TA

* The three values given in each cell of the table are respectively the mean, the sample size and the standard variation.
† Using Tukey's HSD procedure. See Tables 2–4 for details.

under consideration increases—or more precisely, as the relevance criterion becomes less stringent. That is, the average overlap is highest when only most relevant documents are included; it is lowest when all documents are included.

The major finding in these data is that the overlaps are quite small as indicated by the averages. This is true even between representations that should have retrieved very similar sets such as abstract (AA) and title–abstract (TA) or descriptor (DD) and descriptor–identifier (DI). One possible explanation for the small overlaps is searcher differences. The analysis of variance tables (see the relative sizes of the sums of squares in Tables 2–4) support this contention; they show that searcher differences account for one of the largest portions of the variance. However, the results from the McGill *et al.* (1979) study cast doubt on the contention that searchers are the sole or major cause of the low amount of overlap. In that study, overlaps between different representations searched by the same searcher only

Table 2. AOV summary table: Recall–1

Source	*Sum of squares*	*df*	*Mean square*	*F*
Between squares	2.624	11	0.239	
Queries in squares	10.415	58	0.180	
Searchers	4.072	6	0.679	
Squares × searcher	7.940	66	0.120	
Representations	1.415	6	0.236	3.324*
Square × representation	6.021	66	0.091	1.282†
Residual (by subtraction)	19.714	276	0.071	
Total	52.201	489		

* Region of rejection begins at 2.14 ($\alpha = 0.05$) or 2.89 ($\alpha = 0.01$)
† Region of rejection begins at 1.12 ($\alpha = 0.25$). Since obtained value falls within the region of rejection, the square × representation source of variation is not pooled into the residual.

Note 1: Tukey's HSD region of rejection begins at 4.17.
The standard error = 0.0318.
Note 2: Missing values in the data (14 queries retrieved no highly relevant documents) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed.

Table 3. AOV summary table: Recall–2

Source	*Sum of squares*	*df*	*Mean square*	*F*
Squares	0.963	11	0.088	
Queries in squares	5.678	65	0.087	
Searchers	4.088	6	0.681	
Squares × searchers	4.842	66	0.073	
Representations	1.032	6	0.172	3.44*
Pooled error (by subtraction)	19.038	384	0.050	
Total	35.641	538		

* Region of rejection begins at 2.14 ($\alpha = 0.05$) or 2.89 ($\alpha = 0.01$)

Note 1: Tukey's HSD region of rejection begins at 4.17.
The standard error = 0.0255.
Note 2: Missing values in the data (seven queries retrieved no relevant documents at all) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed.

equalled 14 per cent for retrieved documents—a figure which certainly falls in the range of values reported here.

Going beyond pairwise overlaps, the question arises as to the optimum combination of representations, or more precisely, the optimum ordering of representations. That is, if a retrieval environment were limited to a single representation, which one

Table 4. AOV summary table: total retrieved

Source	*Sum of squares*	*df*	*Mean square*	*F*
Between squares	10688.347	11	971.668	
Queries in squares	40273.878	72	559.359	
Searchers	19316.177	6	3219.363	
Squares × searchers	13719.415	66	270.870	
Representations	3654.511	6	609.085	4.24*
Residual	61236.183	426	143.747	
Total	148888.51	587		

* Region of rejection begins at 2.14 ($\alpha = 0.05$) or 2.89 ($\alpha = 0.01$).

Note: Tukey's HSD region of rejection begins at 4.17.
The standard error = 1.308.

Table 5. Mean performance by representation across queries

Representation	*Recall-1*	*Recall-2*	*Precision-1*	*Precision-2*
DD (descriptor)	0.237	0.216	0.173	0.335
AA (abstract)	0.328	0.283	0.181	0.332
TA (title and abstract)	0.369	0.294	0.192	0.324
DI (descriptor and identifier)	0.309	0.268	0.182	0.336
ST (stemmed TA)	0.304	0.281	0.148	0.291
TT (title)	0.285	0.229	0.221	0.378
II (identifier)	0.348	0.306	0.208	0.389

would it be? If a second could be added, which of the remaining six representations contribute the most over and above the effect of the first representation? A third representation could be added over and above the first two, and then a fourth representation, and so on.

The most sensible measure to use in answering this question is the union overlap. Tables 8 and 9 present the results of this analysis. Table 8 uses all seven representations and gives the highly relevant and the total relevant measures across queries. Since three representations (TA, DI, ST) are composed of other representations, the analysis was repeated in Table 9 omitting these 'compound' representations.

Tables 8 and 9 present four different models—different orderings of representations. Such models, if consistent, would allow a searcher to know which combinations 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. Unfortunately, these data suggest that the models are not consistent. What does appear to be highly consistent, however, 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 four models are based on highly interrelated data—data that are subsets of one another. When the cumulative percentages are plotted against the order, the resulting curves appear to be Zipfian in form and when broken down according to Bradford's law of scatter, the obtained proportions are 1:3:7. The theoretical proportions could easily be in the form of 1:3:9, but no attempt was made to verify this analytically.

Table 6. Asymmetric pairwise overlaps*

	AA	*TT*	*TA*	*ST*	*II*	*DI*	*DD*
Most relevant documents							
AA	1.000	0.329	0.401	0.496	0.340	0.368	0.266
TT	0.286	1.000	0.328	0.293	0.348	0.332	0.323
TA	0.451	0.424	1.000	0.520	0.355	0.420	0.344
ST	0.459	0.312	0.428	1.000	0.284	0.332	0.234
II	0.361	0.424	0.334	0.325	1.000	0.508	0.365
DI	0.346	0.359	0.351	0.337	0.450	1.000	0.490
DD	0.192	0.268	0.221	0.183	0.248	0.376	1.000
AVG	0.349	0.353	0.344	0.359	0.338	0.389	0.337
All relevant documents							
AA	1.000	0.276	0.348	0.381	0.275	0.323	0.233
TT	0.223	1.000	0.237	0.212	0.258	0.274	0.268
TA	0.361	0.304	1.000	0.402	0.281	0.310	0.241
ST	0.379	0.261	0.385	1.000	0.233	0.247	0.172
II	0.297	0.344	0.292	0.254	1.000	0.418	0.292
DI	0.305	0.319	0.283	0.235	0.366	1.000	0.458
DD	0.178	0.253	0.178	0.132	0.207	0.370	1.000
AVG	0.291	0.293	0.287	0.269	0.270	0.324	0.277
All documents							
AA	1.000	0.145	0.250	0.229	0.210	0.193	0.103
TT	0.103	1.000	0.113	0.088	0.140	0.131	0.123
TA	0.265	0.169	1.000	0.262	0.188	0.180	0.119
ST	0.259	0.141	0.279	1.000	0.159	0.131	0.080
II	0.193	0.182	0.163	0.129	1.000	0.230	0.131
DI	0.180	0.172	0.158	0.108	0.233	1.000	0.240
DD	0.078	0.131	0.085	0.053	0.108	0.194	1.000
AVG	0.180	0.157	0.175	0.145	0.173	0.177	0.133

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

An ancillary question is that of unique contribution of the different representations. That is, 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 models when the representation is the last entered into the model. Tables 10 and 11 report incremental improvement for each representation, assuming the representation entered the model first or last. These are the maximum and minimum incremental improvements for each representation. Again, the identifier phase is distinctively unique, but more so under the full model than under the restricted one. Table 11 shows AA's unique contribution to be equivalent to II when the overlaps with the compound field (of which AA was a part) are not included in the model. These systematic differences in incremental improvement suggest that the patterns of overlap may be representation specific. It should be noted though that the best unique contributor, II, in the full model retrieved only 20 per cent (i.e., 0.091/0.44) of the uniquely found documents and performed at the 0.35 recall level. Table 10 also reports the sum of the unique percentages, 44 per cent for the Relevant–1 measure, 58 per cent for Relevant–2. In other words, only 56 per cent and 42 per cent of the documents were overlapped—another indication of the low probability of overlap observed in this and other studies.

Lastly, it is important to restate the difficulty of clearly interpreting the overlap

Table 7. Union pairwise overlaps

	AA	*TT*	*TA*	*ST*	*II*	*DI*	*DD*	*Average*
Most relevant documents								
AA	0.328	0.520	0.549	0.481	0.558	0.523	0.502	0.495
TT	0.520	0.285	0.533	0.500	0.512	0.491	0.446	0.470
TA	0.549	0.533	0.369	0.525	0.594	0.548	0.525	0.519
ST	0.481	0.500	0.515	0.304	0.553	0.510	0.485	0.478
II	0.558	0.512	0.594	0.553	0.348	0.500	0.499	0.509
DI	0.523	0.491	0.548	0.510	0.500	0.309	0.430	0.473
DD	0.502	0.446	0.525	0.485	0.499	0.430	0.237	0.446
All relevant documents								
AA	0.283	0.449	0.475	0.457	0.505	0.465	0.449	0.441
TT	0.449	0.229	0.453	0.451	0.456	0.424	0.388	0.407
TA	0.475	0.453	0.294	0.462	0.514	0.479	0.458	0.448
ST	0.457	0.451	0.462	0.281	0.516	0.483	0.461	0.445
II	0.505	0.456	0.514	0.516	0.306	0.462	0.459	0.460
DI	0.465	0.424	0.479	0.483	0.462	0.268	0.385	0.424
DD	0.449	0.388	0.458	0.461	0.459	0.385	0.216	0.402
All documents								
AA	0.220	0.353	0.395	0.412	0.380	0.386	0.369	0.359
TT	0.353	0.156	0.363	0.384	0.331	0.335	0.302	0.318
TA	0.395	0.363	0.234	0.418	0.398	0.402	0.380	0.370
ST	0.412	0.384	0.418	0.249	0.420	0.428	0.402	0.388
II	0.380	0.331	0.398	0.420	0.203	0.361	0.347	0.349
DI	0.386	0.335	0.402	0.428	0.361	0.206	0.332	0.350
DD	0.369	0.302	0.380	0.402	0.347	0.332	0.166	0.329

Table 8. Representations ordered by incremental improvement

Most relevant documents							
Order	1st	2nd	3rd	4th	5th	6th	7th
Representation	TA	II	AA	DD	TT	ST	DI
No. of documents	299	444	574	656	722	768	810
Cumulated percentage	0.369	0.548	0.709	0.810	0.891	0.948	1.000
All relevant documents							
Order	1st	2nd	3rd	4th	5th	6th	7th
Representation	II	ST	DI	TA	TT	AA	DD
No. of documents	527	889	1118	1318	1466	1602	1723
Cumulated percentage	0.306	0.516	0.649	0.765	0.850	0.930	1.000

Table 9. Representations ordered by incremental improvement*

Most relevant documents				
Order	1st	2nd	3rd	4th
Representation	II	AA	TT	DD
No. of documents	282	452	554	634
Cumulated percentage	0.348	0.558	0.684	0.783
All relevant documents				
Order	1st	2nd	3rd	4th
Representation	II	AA	DD	TT
No. of documents	527	870	1093	1275
Cumulated percentage	0.306	0.505	0.634	0.740

* Compound representations omitted.

Table 10. Recalls and unique contributions of seven representations

Representation	*Entered 1st* no. of documents*	*Per cent*	*Entered last* no. of documents*	*Per cent*
Most relevant documents				
AA	266	0.328	49	0.060
DD	192	0.237	44	0.054
DI	250	0.309	42	0.052
II	282	0.348	74	0.091
ST	246	0.303	44	0.054
TA	299	0.369	53	0.065
TT	231	0.285	52	0.064
				0.440
All relevant documents				
AA	488	0.283	137	0.080
DD	373	0.216	127	0.074
DI	462	0.268	120	0.070
II	527	0.306	196	0.114
ST	485	0.281	149	0.086
TA	506	0.244	134	0.078
TT	395	0.229	133	0.077
				0.579

* Entered 1st is the equivalent of Recall–1 across queries when no overlap is taken into account. Entered last are the unique documents found only by that representation.

Table 11. Unique contributions of four representations*

Representation	No. of documents	Per cent
Most relevant documents		
AA	125	0.196
DD	85	0.133
II	114	0.178
TT	88	0.138
All relevant documents		
AA	269	0.210
DD	197	0.154
II	271	0.213
TT	182	0.143

* Recalls on 1st entered are same as in Table 10. Compound representations excluded.

measures. As previously mentioned, representations may be confounded with searchers.

4. SUMMARY

From the data, several conclusions seem warranted. First of all, the performance differences among the representations are not remarkable—none is substantially larger than the others on any of the five measures used and none performed well on all of the performance measures (though the identifier terms are notable here). Secondly, the overlap measures between pairs of representations are, on the average, quite low—even among representations which by their definitions ought to have moderate to high overlaps. Only two of the pairwise union overlaps and none of the asymmetrical overlaps exceeded 0.5 on the average. Third, there is a relationship between relevance and overlap. Most relevant documents have higher overlaps than do all relevant documents, and these have higher overlaps than do all retrieved documents. In terms of the incremental contribution of representations to the total number of relevant documents retrieved, approximately 70 per cent of the most relevant documents can be retrieved with only three representations (TA, II and AA or II, AA and TT if the compound representations are excluded). Interestingly, these are all based on free-text terms, though one of them, the identifier terms, was generated by indexers. The II representation performed surprisingly well in a number of analyses. The results here suggest that the identifier terms contribute unique relevant documents over and above those contributed by other representations. Obviously this will not necessarily be true for all databases, but II should be a top candidate if additional representations are being considered for possible inclusion into an existing system.

While interesting patterns appeared among the representations, none was consistent. Evidently, further studies are needed to see if the sources of this inconsistency can be identified. Future research about document representations, 'may not lead to a choice of one form of representation to the exclusion of all others, but

rather an assessment of which form is most appropriate in a particular situation' (Smith, 1981: 103). Several factors need to be considered. Cost is one. It has been frequently noted that there is an implicit trade-off between adding representations to increase the probability of obtaining high-recall retrievals and the additional expense (in terms of dollars, storage space and processing time) required by those additional representations. Another factor is the database. It is likely that term attributes (e.g., specificity, exhaustivity, homonymy) affect the results of any study of representation overlaps. To this end, we are now in the process of replicating this study with a social science database.

ACKNOWLEDGEMENT

This research was supported in part by the National Science Foundation, Division of Information Science and Technology (Grant IST 79-21468).

REFERENCES

Barker, F. H., Veal, D. C. and Wyatt, B. K. (1972) Comparative efficiency of searching titles, abstracts and index terms in a free-text data base. *Journal of Documentation 28*, 22–36.

Bottle, R. T. (1970) Title indexes as alerting services in the chemical and life sciences. *Journal of the American Society for Information Science 21*, 16–21.

Byrne, J. R. (1975) Relative effectiveness of titles, abstracts and subject headings for machine retrieval from the COMPENDEX services. *Journal of the American Society for Information Science 26*, 223–229.

Cattley, J. M., Moore, J. E., Banks, D. G. and O'Leary, P. T. (1966) Inter-index patent searching by computer. *Journal of Chemical Documentation 6*, 15–26.

Chapman, J. and Subramanyam, K. (1981) Cocitation search strategy. *Proceedings of the 2nd National Online Meeting.* pp. 97–102. Medford, N.J.: Learned Information.

Cleverdon, C. (1967) The Cranfield tests on index language devices. *Aslib Proceedings 19*, 173–194.

Fisher, H. L. and Elchesen, D. R. (1972) Effectiveness of combining title words and index terms in machine retrieval searches. *Nature 238*, 109–110.

Hersey, D. F., Foster, W. R., Stalder, E. W. and Carlson, W. T. (1971) Free text word retrieval and scientist indexing. Performance profiles and costs. *Journal of Documentation 27*, 167–183.

Keen, E. M. (1973) The Aberystwyth index languages test. *Journal of Documentation 29*, 1–35.

King, D. W. and Bryant, E. C. (1971) *The Evaluation of Information Services and Products.* Washington, DC: Information Resources Press.

Kirk, R. E. (1968) *Experimental Design: Procedures for the Behavioural Sciences.* Belmont, Ca.: Brooks/Cole.

Maloney, R. K. (1974) Titles vs. title/abstract text searching in SDI systems. *Journal of the American Society for Information Science 25*, 370–373.

Mansur, O. (1980) On selection and combining of relevance indicators. *Information Processing and Management 16*, 139–153.

Markey, K. and Atherton, P. (1979) Online searching tests. *Online Searching of ERIC: Impact of Free Text or Controlled Vocabulary Searching on the Design of the ERIC Data Base*, Part III. Syracuse University: ERIC Clearinghouse on Information Resources.

McGill, M. J., Koll, M. and Noreault, T. (1979) *An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems.* Final Report for Grant NSF-IST-78-10454 to the National Science Foundation.

Salton, G. (1968) The evaluation of computer-based retrieval systems. In *Automatic Information Organization and Retrieval.* pp. 316–349. New York: McGraw-Hill.

Salton, G. (1971) *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, N.J.: Prentice-Hall.

Salton, G. (1973) A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science 23*, 75–84.

Smith, L. C. (1979) *Selected Artificial Intelligence Techniques in Information Retrieval Systems Research.* Unpublished doctoral dissertation. Syracuse University, School of Information Studies.

Smith, L. C. (1981) Representation issues in information retrieval system design. *ACM SIGIR Forum 16*, 100–105.

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

Sparck Jones, K. and Jackson, D. M. (1970) The use of automatically-obtained keyword classification for information retrieval. *Information Storage and Retrieval 5*, 175–201.

Waldstein, R. K. (1981) *The Role of Noun Phrases as Content Indicators.* Unpublished doctoral dissertation, Syracuse University, School of Information Studies.

Williams, M. E. (1977) Analysis of terminology in various CAS data files as access points for retrieval. *Journal of Chemical Information and Computer Sciences 17*, 16–20.