The Effect of Multiple Query Representations on Information Retrieval System Performance

N.J. Belkin, C. Cool
School of Communication, Information & Library Studies
Rutgers University, New Brunswick, NJ 08903

W.B. Croft, J.P. Callan
Department of Computer Science
University of Massachusetts, Amherst, MA 01003

Abstract

Five independently generated Boolean query formulations for ten different TREC topics were produced by ten different expert online searchers. These different formulations were grouped, and the groups, and combinations of them, were used as searches against the TREC test collection, using the INQUERY probabilistic inference network retrieval engine. Results show that progressive combination of query formulations leads to progressively improving retrieval performance. Results were compared against the performance of INQUERY natural language based queries, and in combination with them. The issue of recall as a performance measure in large databases was raised, since overlap between the searches conducted in this study, and the TREC-1 searches, was smaller than expected.

1. Introduction

The concept of using multiple representations, of either queries or texts, or of using multiple retrieval techniques, in order to improve information retrieval (IR) system performance, has been suggested by several investigators (reviewed below). However, very few of these suggestions have actually attempted to investigate the effect of multiple representations or retrieval techniques on performance. In this paper, we report on a project designed explicitly to study the effect of combining multiple representations of information problems, on the performance of a single IR technique. In the course of our investigation, we ran into problems in evaluation of performance which have not been previously encountered. These problems may affect our results to some extent, but, more importantly, we believe them to have general significance for evaluation of IR systems in large databases. We therefore discuss in this paper, the problem of recall in large databases, as well as the issue of multiple query representations.

Using multiple representations of a single query or text, or using multiple techniques for a single query, has been suggested fairly often in the IR research literature. There are two basic theoretical rationales for such suggestions. The first derives from the observation that different representations of the same query, or of the documents in the database, or different retrieval techniques for the same query, retrieve different sets of documents (both relevant and non-relevant).

ACM-SIGIR'93-6/93/Pittsburgh, PA, USA
© 1993 ACM 0-89791-605-0/93/0006/0339...$1.50

A possible explanation of this observation is that the processes of representation and retrieval are so complex and uncertain that any one representation or technique captures at best only a part of the complexity. Then, the use of multiple techniques or representations is justified on the grounds that their combination will address more aspects of the situation, and thus retrieve more relevant documents. The other rationale is founded in the general probabilistic IR framework (e.g. Robertson, 1978), which suggests that the more sources of evidence are available about the query, or documents, or the relationships between query and documents, the more accurate the judgment of the probability of relevance of a document to a query will be. Thus, each representation of a query is another source of evidence about that query, which could, in principle, be used to improve prediction of probability of relevance.

McGill, Koll & Norreault (1979), in the course of a study of IR ranking mechanisms, noticed that there was surprisingly little overlap between document sets for the same query topic, when searched for by different intermediaries, or by the same intermediary using controlled versus free-text vocabularies. These different conditions define, in effect, different query representations. Although they commented upon this phenomenon, they were unable to investigate it. A later project at the same institution (Katzer, *et al.*, 1982) considered the effect of different document representations (*e.g.* title, abstract) on retrieval performance, and discovered the same phenomenon: that the various representations had similar performance levels, but that there was little overlap in the documents retrieved. Saracevic and Kantor (1987), in a large study of factors affecting retrieval performance in operational online IR systems, also found this result, but were able to follow up on it to a greater extent. Their study had different intermediaries constructing Boolean search formulations based on the same information problem description. Once again, the retrieved sets had little overlap. However, they found that the odds of a document being judged relevant increased monotonically according to the number of retrieved sets that it appeared in. This seems to be the first study to formulate explicitly the concept of combining different query representations to increase retrieval performance. Turtle and Croft (1991) also suggest the use of different query formulations, but based on theoretical rather than empirical considerations. Their argument is derived from a concept of IR as plausible inference, which suggests the use of multiple sources of evidence concerning the relevance of documents to a query, in order to improve performance. They found that combining different query formulations, in particular, combining 'natural language' and 'Boolean' formulations of the same information problem,

increased performance over that of either formulation alone. Both Saracevic and Kantor (1987) and Turtle and Croft (1991) used query representations that seemed to be generated 'independently', and both suggest that this is an important condition on the effectiveness of their combination Foltz and Dumais (1992) have reported that combining the results of multiple retrieval techniques used for the same query formulation improved retrieval performance in an information filtering environment, which extends the general concept of combination beyond that of multiple representations.

The aim of our study is to investigate explicitly the progressive effect of combining multiple query representations of one type on IR system performance. Turtle and Croft (1991) considered the effect of the combination of only two representations, of two different types. Saracevic and Kantor (1988), on the other hand, considered nine different formulations for each query, all Boolean, but were unable to go beyond the observation of increased odds of relevance according to the number of retrieved sets. Furthermore, Turtle and Croft (1991) worked within standard (small) test collections and were thus able to perform 'classic', although perhaps unrealistic, experiments. Saracevic and Kantor (1988) worked in a large, operational environment, and were unable to perform comparative experiments.

Our present project extends both of these studies in important ways. First, we were able to use the rather large TREC test collection (Harman, 1993), allowing us to experiment in a close to realistic environment. Secondly, the study was designed explicitly to investigate the effect of progressively cumulating the evidence of a number of independently generated query representations of one type. And, because we used the INQUERY probabilistic-inference network retrieval system, we were able to compare the results of cumulated searches to a very high baseline performance (Turtle and Croft, 1991), and to investigate combination of natural language with Boolean queries.

2. The INQUERY Retrieval System

In our experiments, we used the INQUERY retrieval engine, a probabilistic inference network-based system developed at the University of Massachusetts (Turtle and Croft, 1991). This system is based upon the idea of combining multiple sources of evidence in order to more plausibly infer the relevance of a document to a query. The underlying formalism is that of a Bayesian probabilistic inference network (Pearl, 1988), which provides strict rules for how to combine sources of evidence. Turtle and Croft (1991) give a detailed description of the model and its implementation; a more general description is available in Belkin and Croft (1992). Here, we note a few characteristics of the system which are germane to the project at hand

First, INQUERY provides a natural means for combination of multiple query formulations, as a function of its design. Second, it incorporates a large set of operators which allow, in addition to sophisticated natural language query formulations, complex Boolean formulations. The Boolean operators in INQUERY are not strict, however, which allows ranking of output, and also leads to significantly better performance than strict Boolean retrieval (Turtle and Croft, 1991). Finally, INQUERY has been used as the basis for an overall query analysis, indexing and retrieval system developed for the DARPA-sponsored TIPSTER initiative, whose database and query set was the basis for the TREC test collection. This version of INQUERY, (referred to here as INQC) performs a sophisticated analysis of the TREC topics (see Appendix B for an example topic), including recognition of country names and automatic syntactic phrase generation, which leads to a complex query formulation based on the full text of the TREC topics and highly effective retrieval performance (see Harman, 1993). INQUERY thus gives us a very high baseline against which to compare our Boolean query results, and also a very different kind of query formulation both to compare to, and to combine with.

3. Method

To obtain the multiple query representations in this study, we recruited experienced online searchers to generate search statements for the same search topics. Ten online searchers, all experienced users of large commercial databases, participated in the project. Searchers were given written 'topic descriptions' of users' information needs, and asked to construct search statements in the form of Boolean queries In creating these queries, searchers were instructed to use the standard Boolean operators AND, OR, and NOT in combining terms. They were allowed to use any degree of nesting Adjacency, proximity and order relationships were permitted, as was truncation. The complete set of instructions to the searchers is included in Appendix A.

A total of ten topic descriptions were used in this study Each of the ten volunteer searchers created a query representation for five different topics, resulting in five independently generated queries for each of the ten topic descriptions. The topic description given to each searcher was a complex statement of an information need. These topic descriptions, taken from the TREC/TIPSTER projects, (Harman, 1993) represented real user problems. Each topic description was structured in the same way, containing information about: problem domain; general topic of the problem; a brief description of the contents of desired documents; and a short narrative describing the criteria for evaluating the relevance of documents. Some of the topic descriptions also contained a list of concepts or terms thought to be useful in searching for relevant documents; factors or limiting conditions on relevant documents; and definitions of terms relevant to the topic description. Searchers were asked to use all of the relevant information in each topic description in constructing their query. Appendix B presents an exemplar topic description.

All of the topic descriptions used in this study were taken from the first 50 user topics distributed by TREC organizers. The databases searched in TREC/TIPSTER included three years of full text of the Wall Street Journal, a year of AP newswire articles, a selection of abstracts from the ComputerSelect database, a set of Federal Register documents and a set of abstracts supplied by the U.S. Department of Energy. The complete TREC data set consisted of approximately 2 gigabytes of text documents, distributed in two halves. In this study we ran our queries against the second half only.

Ten topic descriptions were selected from the 50 that were available. Half of the 50 topic descriptions were in the domain of 'International Economics' and the other half in the domain of 'Science and Technology'. We selected five topic descriptions from each domain. We also chose topic descriptions for which we knew some relevance judgements existed at the time that we began our project. Table 1 presents the titles of the ten topic descriptions used in this study.

Five independently generated queries were collected for each of the ten topic descriptions. Five query groups were then constructed, each one containing different representations of all ten topics. These query groupings enabled us to compare the average retrieval performance in two ways - first on a group by group basis, to investigate the effect of different types of query representations on retrieval performance, and then cumulatively to directly assess the effect of multiple query representations. In

this latter approach, query groups were added to create increasingly more complex query representations.

Two different methods were used to form the query groups. The first approach was more or less unsystematic, and resulted in groups in which individual searchers had unequal representation. That is, within each of these original five groups at least four of the query representations came from a single searcher. Our second approach to query grouping attempted to neutralize potential searcher bias within groups by creating an even distribution or equal representation of searchers in each group. In this second method of query grouping, every searcher was represented once in every group.

002: Acquisitions
005: Dumping charges
006: Third World debt relief
016: Marketing of agrochemicals
017: Measures to control agrochemicals
027: Expert systems and neural networks in business or manufacturing
031: Advantages of OS/2
034: Entities involved in building ISDN applications and developing strategies to exploit ISDN
038: The role of minicomputers and mainframes in an environment increasingly dominated by LAN's PC's and workstations
039: Client-server plans and expectations

Table 1. Titles of the TREC topics used in this study

The Boolean queries produced by our volunteer searchers were translated into INQUERY commands and run on that system against half of the database, which consisted of 231,471 documents in total. INQUERY allows a relaxation of Boolean operators in order to produce ranked output. Consistent with TREC, the top 200 documents were retrieved. The combination of groups was done on INQUERY by using the unweighted sum operator.

Retrieval performance was looked at in the following ways. First, individual query groups were compared against each other and then against INQUERY. Next, the cumulative performance of the combined groups was assessed, by looking at query group 1, groups 1+2, groups 1+2+3, and so on. The performance of each increasingly more complex query group was compared against the original single group results, and against INQUERY. Finally, we added INQUERY to our most complex Boolean group, the combination of groups 1-5, and looked at retrieval performance. At this step in the analysis we were particularly concerned with investigating the effectiveness of combining Boolean queries with INQUERY results (Turtle & Croft, 1991).

4. Results

The results of our experiments are presented as level of precision at standard levels of recall. These figures are computed on the basis of the TREC relevance judgements. For each of the TREC topics, the top 200 documents retrieved by each of the participating systems (20+) were judged for relevance by the persons who prepared the topics. The characteristics of these cumulated retrieved document sets, for each of our ten topics, are displayed in Table 2.

Topic No.	Unique ret.	Relevant	Precision
002	1728	380	0.22
005	1224	183	0.15
006	1094	164	0.15
016	850	71	0.08
017	785	206	0.26
027	830	291	0.35
031	890	172	0.19
034	852	371	0.44
038	1200	876	0.73
039	1000	557	0.56

Table 2. Characteristics of retrieved documents for ten TREC topics.

Table 3 presents the retrieval performance for each of our five original query groups (each group having at least four queries by the same searcher). Although there are some differences in performance, each group seems to perform at approximately the same overall level. Table 4 shows the performance as these groups are combined, in a sequential fashion. Here, it is of interest to note that adding a group which performed less well than a previous one nevertheless increased performance; that performance increased with each additional group; and, that the performance of all groups combined was better than that of any single group.

Being concerned with possible searcher effects, we re-grouped the queries, so that each searcher was represented only once in each group. The results for this grouping are presented in Tables 5 and 6. In Table 5, one sees that group 2 performs substantially better than any of the other groups, which are mostly quite similar to one another, with the exception of group 3, which performs somewhat more poorly than the others. The effect of these differences seems to show up when the groups are combined, as shown in Table 6. There is a quite significant increase in performance when group 2 is added to group 1, but adding group 3 to these slightly decreases performance. Thereafter, there is again a steady improvement in performance, with the complete combination being effectively the same as that in the previous groupings (Table 4). In both cases, the combination of all groups leads to performance dramatically better than that of the single weakest one, and at least somewhat better than the best single one.

Although the increase in performance through combination was more regular in our original groupings, for the rest of our analyses we use the groupings in which searchers are evenly distributed. This is primarily because we do not yet understand the effect of individual searchers on performance, especially how it might affect the weight given to a particular source of evidence. We would like, for the present, to be neutral on this issue, and even distribution of searcher allows us to do this.

We now address the question of how well this combination performs relative to some other standard. To do this, we compare it to the performance of the current best version of the INQUERY indexing/representation routines and retrieval techniques, which have been optimized for the TREC collection. The results are shown in Table 7. The INQUERY results (INQC) are substantially better than those of the combined Boolean queries, with the greatest advantage being at the middle recall levels.

Recall	Precision (% change, with respect to group 1).				
	1	2	3	4	5
0	77.4	59 6 (-23.0)	78.6 (+1.5)	75.0 (-3.1)	68.8 (-11.1)
10	46.1	48.7 (+5 6)	46.4 (+0.6)	47.3 (+2.7)	42.2 (-8 4)
20	39.0	45.5 (+16.7)	37.4 (-4.1)	42.7 (+9.5)	34.4 (-11.8)
30	34.7	41.1 (+18.6)	33 4 (-3 7)	41.1 (+18.5)	32.5 (-6.4)
40	28.0	36.3 (+30.0)	27.5 (-1 6)	37 5 (+34 2)	27.4 (-2.1)
50	23.8	31.2 (+30.8)	22.1 (-7.1)	34.1 (+43.1)	22.9 (-3.7)
60	20.9	27.4 (+31.1)	17.9 (-14.3)	29.6 (+41.5)	18 5 (-11.5)
70	12.3	22.6 (+83.8)	15.6 (+26.6)	24.1 (+95.9)	16.4 (+33.2)
80	10.3	14.3 (+38.3)	9.4 (-9.3)	16.2 (+57.2)	9.6 (-7 1)
90	8.4	10.4 (+23.9)	5.0 (-40.8)	9.7 (+16.4)	5.0 (-39.9)
100	3.7	3.9 (+5.4)	0.6 (-82.9)	3.4 (-6.6)	0.6 (-83.6)
avg	27.7	31.0 (+12.0)	26.7 (-3.5)	32.8 (+18.5)	25 3 (-8 6)

Table 3. Retrieval results for each group of Boolean queries, groups with uneven distribution of searchers.

Recall	Precision (% change with respect to group 1 alone).				
	1	1+2	1+2+3	1+2+3+4	1+2+3+4+5
0	77.4	71.8 (-7.3)	74.3 (-4.0)	75.1 (-3.0)	75.6 (-2.3)
10	46.1	48.7 (+5.7)	48.3 (+4.7)	50.9 (+10.4)	54.6 (+18.6)
20	39.0	43.0 (+10.2)	44.6 (+14.4)	46.4 (+19.1)	49.0 (+25 7)
30	34.7	40.0 (+15.2)	41.9 (+20.7)	43.1 (+24.2)	44.7 (+28.8)
40	28 0	35.6 (+27.2)	37 2 (+33 1)	39.3 (+40.6)	40.3 (+44.3)
50	23.8	31.0 (+30.0)	33.1 (+38.9)	35.5 (+48.9)	35.6 (+49 7)
60	20.9	26.9 (+28.5)	28.6 (+36.6)	31.4 (+50.0)	32.0 (+52.7)
70	12.3	23 1 (+87 8)	24 1 (+96 0)	27.6 (+124 8)	28 0 (+127.8)
80	10.3	14.8 (+43.2)	17.4 (+68.7)	20.1 (+94.8)	21.0 (+104.1)
90	8.4	10.1 (+20.5)	11.3 (+35.1)	12.7 (+51.7)	12.6 (+51.2)
100	3.7	4.0 (+8.1)	3.9 (+6.3)	4.2 (+14.5)	4.1 (+11.0)
avg	27.7	31.7 (+14.5)	33.1 (+19.7)	35.1 (+26.9)	36.2 (+30.6)

Table 4 Retrieval results, combining query groups, groups with uneven distribution of searchers

Recall	Precision (% change, with respect to group 1).				
	1	2	3	4	5
0	69.3	77.3 (+11.5)	61.9 (-10.7)	69.3 (+0.1)	77.9 (+12.4)
10	47.6	51.8 (+8.8)	39.2 (-17.7)	45.5 (-4.4)	41.9 (-12.1)
20	41.9	47.9 (+14.3)	32.7 (-21.9)	38.3 (-8.7)	35.0 (-16.4)
30	37.8	44.0 (+16.2)	29.2 (-22.8)	36.4 (-3.8)	33.5 (-11.4)
40	31.5	38.6 (+22.7)	25.3 (-19.6)	31.9 (+1.3)	27.5 (-12.6)
50	25.9	32.1 (24.1)	22.6 (-12.8)	27.0 (+4.4)	24.8 (-4.4)
60	21.0	27.9 (+33.2)	20.5 (-2.0)	22.0 (+4.8)	21.2 (+1.3)
70	11.7	23.7 (+102.0)	17 5 (+49.1)	17.1 (+45.8)	18.7 (+60 1)
80	8.5	16.4 (+91.7)	12.2 (+42.7)	9.6 (+12.6)	11.9 (+39.0)
90	5.5	11.4 (+1081.2)	8 2 (+50.3)	4.5 (_17.1)	7.8 (+41.7)
100	0.7	3.7 (+428.1)	3.3 (+379.2)	0.4 (-47.2)	3.8 (+440.4)
avg	27.4	34.1 (+24.3)	24.8 (-9.5)	27.5 (+0.2)	27.6 (+0.8)

Table 5 Retrieval results for each group of Boolean queries, groups with even distribution of searchers.

Recall	Precision (% change with respect to group 1 alone).				
	1	1+2	1+2+3	1+2+3+4	1+2+3+4+5
0	69.3	72.5 (+4.6)	72.0 (+3.9)	74.5 (+7.5)	74.5 (+8.0)
10	47.6	54.1 (+13.7)	48.5 (+ 2.0)	51.1 (+7.4)	54.9 (+15.4)
20	41.9	50.0 (+19.4)	45.2 (+8.0)	47.9 (+14.4)	49.5 (+18.3)
30	37.8	45.3 (+19.8)	41.8 (+10.4)	43.5 (+14.9)	45.0 (+19.0)
40	31.5	39.9 (+26.5)	37.7 (+16.5)	37.9 (+20.3)	39.1 (+24.1)
50	25.9	34.7 (+34.1)	32.2 (+24.2)	33.7 (+30.3)	34.5 (+33.3)
60	21.0	29.9 (+42.5)	28.1 (+33.9)	30.3 (+44.8)	30.4 (+45.1)
70	11.7	25.4 (+117.1)	24.1 (+105.5)	26.8 (+128.6)	26.8 (+128.6)
80	8.5	17.5 (+104.8)	17.5 (+104.8)	19.6 (+129.2)	19.8 (+131.9)
90	5.5	11.0 (+101.6)	10.9 (+98.9)	11.8 (+114.9)	11.9 (+116.9)
100	0.7	4.0 (+475.3)	4.0 (+475.3)	4.3 (+520.4)	4.1 (+495.7)
avg	27.4	34.9 (+27.5)	32.8 (+19.8)	34.7 (+26.5)	35.5 (+29.7)

Table 6. Retrieval results, combining query groups, groups with even distribution of searchers

Recall	Precision (% change with respect to INQC).	
	INQC	1+2+3+4+5
0	86.4	74.8 (-13.4)
10	72.4	54.9 (-24.1)
20	67.5	49.5 (-26.6)
30	62.0	45.0 (-27.4)
40	54.6	39.1 (-28.4)
50	46.4	34.5 (-25.6)
60	38.5	30.4 (-20.9)
70	32.7	26.8 (-17.9)
80	21.7	19.8 (-8.6)
90	13.9	11.9 (-14.5)
100	1.7	4.1 (+144.4)
avg	45.2	35.5 (-21.4)

Table 7. Comparison of retrieval performance of INQUERY C with combined Boolean queries

Finally, we consider the issue of combining INQC and the combined Boolean queries, as two different sources of evidence, in the spirit of Turtle & Croft (1991), who combined natural language and Boolean queries. The results are presented in Table 8. We began by simply summing these two sources, with equal weights. These are the results in Table 8 in the column headed 12345+c. Since this had the effect of reducing overall performance, we then weighted the evidence of the combined Boolean queries at several levels: twice, one-half and one-quarter. Table 8 shows that the more weight is given to the Boolean query evidence, the worse the overall retrieval performance, but that performance of the combination is improved at fractional weightings.

A potential problem with these results is that some of the documents retrieved by our queries may not have been retrieved by any of the other systems participating in TREC. This would mean that some retrieved documents would be neither relevant nor non-relevant, and thus would not be incorporated in the performance results. We therefore checked the top 200 retrieved items for each topic, in the completely combined condition. The results, displayed in Table 9, show what seem to be rather high numbers of non-judged documents for several topics.

Recall	Precision (% change with respect to INQC)				
	INQC	2.0*12345+c	12345+c	0.5*12345+c	0.25*12345+c
0	86.4	85.2 (-1.4)	83.7 (-3.1)	83.9 (-2.9)	82.9 (-4.0)
10	72.4	62.4 (-13.9)	67.4 (-6.9)	71.6 (-1.2)	72.7 (+0.4)
20	67.5	57.2 (-15.2)	62.9 (-6.9)	66.9 (-0.9)	68.3 (+1.2)
30	62.0	53.2 (-14.1)	57.8 (-6.7)	60.0 (-3.1)	62.2 (+0.3)
40	54.6	47.4 (-13.1)	51.1 (-6.3)	55.5 (+1.7)	56.5 (+3.5)
50	46.4	40.3 (-13.1)	43.3 (-6.7)	47.9 (+3.2)	50.5 (+8.8)
60	38.5	34.6 (-10.0)	37.0 (-3.7)	40.8 (+6.2)	41.7 (+8.3)
70	32.7	29.9 (-8.5)	31.3 (-4.1)	32.8 (+0.3)	33.9 (+3.7)
80	21.7	22.1 (+1.9)	22.8 (+5.4)	23.9 (+9.7)	23.8 (+9.6)
90	13.9	13.6 (-2.5)	14.5 (+4.5)	15.1 (+8.7)	15.6 (+12.2)
100	1.7	4.5 (+162.8)	4.6 (+169.0)	4.5 (+163.5)	3.6 (+110.7)
avg	45.2	40.9 (-9.5)	43.3 (-4.2)	45.7 (+1.0)	46.5 (+2.8)

Table 8. Performance of INQC in combination with combined Boolean queries, at various weights of the Boolean query evidence .

Topic	Unjudged	Relevant	Non-rel.	Precision
002	134	17	49	0.257
005	61	66	73	0.47
006	66	13	121	0.09
016	151	14	35	0.28
017	87	96	17	0.85
027	12	106	82	0.56
031	49	66	85	0.32
034	11	111	78	0.58
038	33	167	0	1.0
039	20	180	0	1.0

Table 9. Characteristics of document sets retrieved by combined Boolean query

Since we were not in a position to obtain relevance judgments for the unjudged documents, we decided to run our tests over again, excluding the two topics, 002 and 016, which had the most non-judged documents, from the tests. The results of these new tests are presented in Tables 10, 11 and 12. Considering these tables, and comparing them to Tables 5, 6, and 7, we note that: the same pattern of performance and performance improvement holds for both cases; and, performance for all groups and combinations, and for INQC, improves substantially. In the next section, we discuss some implications of the data in Tables 2 and 9. But the results in Tables 10, 11 and 12 seem to confirm our overall finding, that combination of query formulations improves IR system performance in a fairly regular way.

Recall	Precision (% change, with respect to group 1).				
	1	2	3	4	5
0	79.8	77.8 (-2.5)	64.6 (-19.1)	76.4 (-4.3)	96.1 (+20.4)
10	58.9	62.6 (+6.4)	45.2 (-23.3)	53.8 (-8.7)	52.0 (-11.8)
20	52.0	58.7 (+12.9)	40.5 (-22.1)	46.1 (-11.4)	43.6 (-16.2)
30	47.1	53.9 (+14.4)	36.3 (-22.8)	44.4 (-5.7)	41.7 (-11.4)
40	39.2	47.6 (+21.4)	31.5 (-19.8)	39.1 (-0.2)	34.3 (-12.7)
50	32.3	39.6 (+22.8)	28.0 (-13.1)	33.3 (+3.2)	30.8 (-4.6)
60	26.1	34.3 (+31.6)	25.5 (-2.4)	27.1 (+3.6)	26.4 (+1.1)
70	14.6	29.1 (+99.4)	21.7 (+48.6)	21.0 (+44.0)	23.4 (+60.0)
80	10.6	20.1 (+88.7)	15.1 (+42.3)	11.8 (+10.8)	14.8 (+38.9)
90	6.8	14.0 (+106.1)	10.2 (50.0)	5.5 (-19.2)	9.6 (+41.5)
100	0.8	4.6 (+443.6)	4.1 (+392.0)	0.4 (-51.6)	4.7 (+457.3)
avg	33.5	40.2 (+20.1)	29.3 (-12.4)	32.6 (-2.6)	34.3 (+2.4)

Table 10. Retrieval results for each group of Boolean queries, 8 queries only per group.

Recall	Precision (% change with respect to group 1 alone)				
	1	1+2	1+2+3	1+2+3+4	1+2+3+4+5
0	79.8	77.7 (-2.7)	76.2 (-4.5)	79.3 (-0.7)	79.6 (-0.3)
10	58.9	66.6 (+13.2)	59.0 (+0.2)	60.4 (+2.6)	66.1 (+12.2)
20	52.0	61.5 (+18.3)	55.7 (+7.1)	56.9 (+9.5)	60.2 (+15.8)
30	47.1	57.9 (+22.9)	52.2 (+10.8)	54.8 (+16.3)	56.1 (+19.1)
40	39.2	51.4 (+30.9)	46.4 (+18.2)	49.0 (+24.8)	50.2 (+27.9)
50	32.3	44.6 (+38.0)	42.1 (+30.5)	44.2 (+37.1)	44.5 (+37.8)
60	26.1	38.6 (+47.8)	36.2 (+38.5)	39.2 (+50.2)	39.8 (+52.1)
70	14.6	32.9 (+125.5)	30.9 (+111.5)	34.8 (+138.5)	35.0 (+139.5)
80	10.6	23.0 (+115.8)	22.0 (+107.1)	25.3 (+137.6)	25.8 (+142.2)
90	6.8	14.3 (+109.7)	13.8 (+102.5)	15.7 (+129.9)	15.8 (+131.7)
100	0.8	4.7 (+456.9)	4.8 (+469.1)	5.3 (+528.9)	5.1 (+502.6)
avg	33.5	43.0 (+28.4)	39.9 (+19.3)	42.3 (+26.2)	43.5 (+29.8)

Table 11. Retrieval results, combining query groups. 8 queries only per group.

Recall	Precision(% change with respect to INQC).		
	INQC	1+2+3+4+5	0.25*12345+ c
0	92.6	79.6 (-14.1)	86.1 (-7.0)
10	76.4	66.1 (-13.5)	75.6 (-1.0)
20	71.8	60.2 (-16.2)	72.2 (+0.5)
30	67.7	56.1 (-17.1)	67.6 (-0.1)
40	60.8	50.2 (-17.5)	61.5 (+1.1)
50	54.1	44.5 (-17.8)	56.2 (+3.9)
60	46.6	39.8 (-14.8)	48.4 (+3.7)
70	39.7	35.0 (-11.9)	41.6 (+4.9)
80	26.6	25.8 (-3.0)	29.2 (+9.8)
90	17.2	15.8 (-8.1)	18.8 (+9.3)
100	2.1	5.1 (+142.1)	4.0 (+91.1)
avg	50.5	43.5 (-14.0)	51.0 (+1.0)

Table 12. Comparison of performance of INQC with combined Boolean query formulation, and with INQC combined with Boolean, 8 queries only .

5. *Discussion*

5.1 Combination of query formulations

The basic message of our results is that combining different, independently generated Boolean query formulations has, in general, a positive effect upon retrieval performance; and, that in general, the more such formulations are used, the better the performance. This seems to indicate that a good rule of thumb on combining query formulations is: The more, the better. Furthermore, even though the best Boolean combination performed less well than INQC, it was still possible to increase performance of INQC by an appropriate combination of the Boolean evidence. This again supports the principle of combination of query formulations, this time with different types of queries. Thus, our results appear to both support and extend those of Turtle and Croft (1991) and Saracevic and Kantor (1988).

There are, however, some problems and open questions in the interpretation of these results. One problem has to do with the dip in performance that occurred when query group 3 was added to groups 1 and 2 in the evenly distributed grouping (Table 6). This appears to have been caused by group 2's exceptionally high level of performance, and group 3's relatively low level. Indeed, the performance of group 2 in this condition is almost as good as that of the combination of all of the query groups. This has at least two possible implications. One is that, given that there is no way to tell, *a priori,* how well any query formulation, or set of formulations will perform, an optimal strategy is to use as many as are available. Another, however, is that these results appear to suggest that some query formulations may be better than others. This in turn suggests that either: there may be some optimum order of combination; or, that there may be some optimum weighting of sources of evidence; or, that there may be some optimum single source, or single query formulation, in any given case. All three of these suggestions require some means of estimating the performance of a source or formulation in advance of its use, an inherently difficult problem which we have been unable to address in our current study.

There has been some research addressing the choice of an optimal representation or retrieval technique for any given information problem, for example Croft (1981); Croft and Thompson (1984); Belkin and Kwasnik (1986). Unfortunately, those studies which tried to predict, on the characteristics of the information problem, which retrieval technique or representation scheme would be best, have mostly failed to be able to do so. The alternative method of prediction, based on a feedback iteration, as suggested by Frants, Shapiro and Voiskunskii (1993) is both practically difficult, and possibly not productive, given that combination might be just as effective.

A final difficulty with our data is that they are based on only ten information problems, or queries. Although the dataset itself is very large, this is clearly a limitation on the validity of this study.

All of these questions are now being addressed in further research on optimal combination of sources of evidence, being undertaken at Rutgers University in the context of the TREC-2 program. In particular, the number of queries will be substantially expanded, different methods of combination will be tested, and relative weighting of different formulations will be investigated, as will automatic generation of multiple query representations, both Boolean and with other structures.

5.2 Recall and performance evaluation in large databases

The results on overlap between our retrieved documents and those retrieved by all of the TREC-1 participants suggest to us some real problems in the use of recall as a performance measure in large databases. For both Topics 002 and 016, there was what seems to be surprisingly little overlap, given that at least 20 different systems had already searched the same databases. This is especially the case for 002, which had a very large number of retrieved items already. This result may be related to shome characteristics of the queries and the databases which were most relevant to them, since we had much greater overlap with some queries than with others.

But we are concerned that our results indicate a larger problem. That is, in large databases, we can, in general, expect each alternative system to retrieve quite different documents, and also at least some unique relevant documents. This is consistent with the experience of TREC-1, in which overlap was in general rather small (see Table 2, for instance). In small databases, like the typical test collections, this has not been a significant issue. But in large databases, given our experience, it may be quite significant. For instance, without going back to the TREC relevance evaluators, we have no precise way to judge the performance of our method, with reasonable confidence, for more than half the topics we searched. And there appears to be no reason, in principle, why this should not be the case for the next method which comes along. This suggests that using recall to evaluate performance in such circumstances is problematic, and that other measures could be considered. This issue is also of concern to us in our further research.

6. *Acknowledgement*

This research was supported in part by the NSF Center for Intelligent Information Retrieval at the University of Massachusetts at Amherst.

7. *References*

BELKIN, N.J. & CROFT, W.B. (1992) Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35,12: 29-38.

BELKIN, N.J. & KWASNIK, B.H. (1986) Using structural representations of anomalous states of knowledge for choosing document retrieval stategies. In: F. Rabitti, ed. *Proceedings* of the 1986 ACM Conference on Research and Development in Information Retrieval. Pisa, ACM-CNR: 11-22.

CROFT, W.B. (1981) Incorporating different search models into one document retrieval system. *SIGIR Forum*, 16,1: 40-45.

CROFT, W.B. & THOMPSON, R.H. (1984) The use of adaptive mechnaisms for selection of search strategies in document retrieval systems. In: C. J. van Rijsbergen, ed. *Proceedings* of the ACM/BCS International Conference on Research and Development in Information Retrieval. Cambridge, Cambridge University Press: 95-110.

FOLTZ, P.W. & DUMAIS, S.T. (1992) Personalized information delivery: An analysis of information-filtering methods. *Communications of the ACM*, 35,12: 51-60.

FRANTS, V.I., SHAPIRO, J. & VOISKUNSKII, V.G. (1993). Multiversion information retrieval systems and feedback with mechanism of selection. *Journal of the ASIS*, 44,1: 19-27.

HARMAN, D. (ed). (1993) *The First Text REtrieval Conference (TREC1).* National Institute of Standards and Technology Special Publication 200-207. Gaithersburg, MD.

KATZER, J., MCGILL, M.J., TESSIER, J.A., FRAKES, W & DASGUPTA, P. (1982) A study of the overlap among document representations. *Information Technology: Research and Development,* 1: 261-274.

MCGILL, M., KOLL, M. & NORREAULT, T. (1979) An evaluation of factors affecting document ranking by information retrieval systems. Syracuse, Syracuse University School of Information Studies.

ROBERTSON, S.E. (1977) The probability ranking principle in IR. *Journal of Documentation*, 33,4: 294-304.

PEARL, J. (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Palo Alto, CA, Morgan Kaufmann.

SARACEVIC, T. & KANTOR, P. (1988) A study of information seeking and retrieving. III. Searchers, searches, overlap. *Journal of the ASIS*, 39,3: 197-216.

TURTLE, H. & CROFT, W.B. (1991) Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9,3: 187-222.

APPENDIX A: Instructions to Searchers:

We would like you to construct a general search statement for each enclosed Topic Description, using all of the information in the Topic Description which you feel is relevant for doing this. The form of this statement should be a general Boolean query, using precisely and all of the terms which you think would be needed to do a good search for this topic. The search will be against the free text, not indexing terms. In structuring your search statement, please do not use the symbols of a specific query language (e.g. DIALOG). Rather, the terms should be combined using the standard Boolean operators: AND, OR, NOT, using any degree of nesting of statements that you feel necessary. You may indicate the following relationships of terms: adjacency, proximity and order, by using the following operators, respectively: ADJ; n; ORD; n. The difference between proximity and order is that in the former, the words may appear in any order; whereas in the latter, they must appear in the specified order. The ADJ, proximity, or ORD operators should be limited to relations at the word level, rather than more complex expressions, such as phrases. The system to which the queries will be put does automatic stemming, so you need not indicate this feature unless you feel it necessary; if so, please use the symbol '$'. This system does not allow prefix stemming.

APPENDIX B: Example Topic Description:

Number: 002
Domain: International Economics
Topic: Acquisitions
Description: Document discusses a currently proposed acquisition involving a U.S. company and a foreign company.
Narrative: To be relevant, a document must discuss a currently proposed acquisition (which may or may not be identified by type, e.g., merger, buyout, leveraged buyout, hostile takeover, friendly acquisition). The suitor and target must be identified by man; the nationality of one of the companies must be identified as U.S. and the nationality of the other company must be identified as NOT U.S.
Concepts:
1. acquisition, takeover
2 suitor, target
3. merger, buyout, leveraged buyout (LBO)
4. arb, arbitrage, arbitrager, leverage, offer, bid, tender, purchase
5. anti-takeover, poison pill, white knight, restructure, sale of assets, recapitalization
Factors:
Nationality: U.S.
Nationality: Not U.S.
Time: Current
Definitions:
Acquisition- The taking over by one company of a controlling interest in another, also called a takeover. The action may be friendly or unfriendly. The company initiating the takeover is the suitor. The company which is taken over is the target.
Arbitrage- A form of speculation in which the purchase of an asset in one market is accompanied by a simultaneous sale of the same or similar asset in a different market, to take advantage of a difference in price. Arbitragers or arbs buy a company's shares at today's price expecting them to be bid for tomorrow at a higher price in a takeover bid. When they do this knowing that a bid is coming, they are indulging in insider trading.
Hostile Takeover- An acquisition in which the suitor company plans to replace management or liquidate assets, etc. of the target company. The target company may institute some countermeasure, called a poison pill. An investor or group which tries to assist the target company is a white knight.
Leveraged Buyout (LBO)- Takeover of a company using borrowed funds, with the target company's assets serving as security for the loans taken out by the acquiring firm. The acquiring firm repays the loans out of the cash flow of the acquired company or from the sale of the assets of the acquired company.
Merger- The acquisition by one corporation of the stock of another. The acquiring corporation then retires the other's stock and dissolves that corporation. Therefore, only one corporation retains its identity in a merger.
Tender Offer- An offer to buy shares of a corporation, usually at a premium above the shares' market price, for cash, securities, or both, often with the objective of taking control of the target company.