Combining Evidence for Information Retrieval

N.J. Belkin, P. Kantor, C. Cool, R. Quatrain
School of Communication, Information & Library Studies
Rutgers University
New Brunswick, NJ 08903 USA
[belkin"kantorp/ccoollquatrain]@cs.rutgers.edu

Abstract

This study investigated the effect on retrieval performance of two methods of combination of multiple representations of TREC topics. Five separate Boolean queries for each of the 50 TREC routing topics and 25 of the TREC ad hoc topics were generated by 75 experienced online searchers. Using the INQUERY retrieval system, these queries were both combined into single queries, and used to produce five separate retrieval results, for each topic. In the former case, results indicate that progressive combination of queries leads to progressively improving retrieval performance, significantly better than that of single queries, and at least as good as the best individual single query formulations. In the latter case, data fusion of the ranked lists also led to performance better than that of any single list.

1. Introduction

The general goal of our project in the TREC-2 program was to investigate the effect of making use of several different formulations of a single information problem, on information retrieval (IR) system performance. The basis for this work lies in both theory and empirical evidence. From the empirical point of view, it has been noted for some time, that different representations of the same information problem retrieve sets (or ranked lists) of documents which contain different relevant, as well as non-relevant documents (see, e.g. McGill, Koll & Norreault, 1979; Saracevic & Kantor, 1988). There is some implication from this evidence (made explicit by Saracevic and Kantor, 1988), that taking account of the different results of the different formulations, could lead to retrieval performance that is better than that of any of the individual query formulations. From the theoretical point of view, IR can be considered as a problem of inference (see, e.g. van Rijsbergen, 1986). That is, IR is concerned with estimating, given available evidence about such things as information problems and documents (or in general, retrievable information objects), the likelihood (or probability, or degree) of relevance of a document to the information problem. From this point of

view, different query formulations constitute different sources of evidence which could be used to infer the probable relevance of a document to an information problem, and it is thus reasonable to consider ways in which to use (i.e. combine) these sources of evidence in the inference process.

These ideas are general to any source of evidence which might be used for IR, such as the evidence of different retrieval techniques, or different document representation techniques, or, in general, different IR systems. One aspect

of our project uses the example of different query formulations as a simulation of the general problem of combination of evidence from different systems.

An additional argument is available for the special case of different query representations. That is, if we consider an information problem to be a complex, and in general difficult-to-specify entity (see, e.g. Taylor, 1968; Belkin, Oddy & Brooks, 1982), then we might conclude that each different representation, derived from some statement by the user, is a different interpretation of the user's underlying information problem, highly unlikely to be like anyone else's (or any other system's) interpretation. Given the empirical evidence, whether any one such interpretation is `better' than another seems mooL However, we might say that each captures some different, yet pertinent aspect of the user's underlying problem; or, that those aspects of the different interpretations which are common to them all (or more than one) reflect some `core' aspect of the problem. Although techniques for making use of the different interpretations might vary according to which of these two views one takes, the general position suggests that it will always be a good idea to take advantage of as many such interpretations as possible. For this case, we therefore consider the issue of combination of different query representations within the `same' IR system.

Our project, thus, considers the problem of inference in IR at two levels of analysis. The first level, as introduced by Turtle & Croft (1991), asks about the effect of evidence obtained when two or more formal query statements are produced for the same information problem. The second level, which is simulated in this study, asks about combination of evidence provided by two or more distinct systems, ranking the same set of documents in response to the

same problem. To distinguish th~se two levels, and in keeping with earlier discussions of the issues involved, we henceforth refer to the combination of query statements as "query combination", and we refer to the combination of evidence from differing systems as "data fusion". Others have also addressed various aspects of this general question. Apart from those akeady cited, we mention in particular the work of Fox and his colleagues (Fox et al., 1993; Fox and Shaw, this volume), and that of Belkin, et al. (1993). These studies in fact address precisely the question of query combination, the Belkin et al. work being a direct precursor to this, and the Fox et al. studies using different query formulation, combination and retrieval techniques, but with very similar results.

Why ought either of these two methods work in the IR situation? The central idea is that either the specific internal score, assigned to a document for a query, or the rank of a document in the list produced for a query, represents information about the relevance of the document to the query. For Boolean retrieval, we may address this question with concepts of signal detection. In this framework, there are two conditional probabilities. The probability that a relevant document is retrieved by system S is d5. The probability that a not relevant document is retrieved is f5. If two Systems (or formulations) are independent, the posterior relevance odds are increased by the product dld21flf2. In actual application (Saracevic and Kantor, 1988), improvements are not this large, suggesting either the existence of an effective base of not-relevant documents, or some effect of interdependence. It can be shown that if several query formulations are drawn from a normal distribution centered at the optimal query formulation, then some fraction of the time, the simple average of these formulations will be closer to the optimum than even the best of them. An even larger fraction of the time, there will be an optimum linear combination which is more nearly optimal than any of the cases from which it is formed ~antor, 1993).

The existence of such models explains why we might expect combination of evidence, or data fusion, to work for the case of several query formulations, as, for instance, in the INQUERY retrieval system (Turtle & Croft, 1991). But these models do not predict that these techniques must work. The investigation of whether they do work, is the subject of this paper.

Specifically, we investigate whether data fusion meth-

ods will produce better performance than any single method; and, whether combination of query formulations does better than the best individual query formulations, and whether progressive combination of query formulations leads to progressively better IR performance. For each of these questions, we also address the issue of what methods to use in the combination of evidence.

In this paper, we do not discuss the "official" results which we submitted to TREC-2, except in passing. The reason for this is that we are not so much interested in the absolute performance of the techniques which we use, as in their performance relative to one another. what we are most concerned with is what happens to retrieval performance as we combine evidence; if we find that combining evidence in specific ways leads to improvements over our starting point of non-combination, then we can begin to investigate how to optimize starting points, as well as rules for combination.

The general plan of our study was as follows. We collected, from experienced online searchers, five different query formulations for each of the 50 routing topics and for 25 of the ad hoc topics. These query formulations were then put to the INQUERY retrieval system (made available to us by the University of Massachusetts), both as single queries, and as combinations of queries for each topic. The combinations were studied at various levels, with the five-fold combination for each set being reported as "official" TREC-2 results for query combination. The five retrieved lists for the ad hoc topics were merged, and reported as "official" TREC-2 results for data fusion.

2. Methods

2.1 Query Formulation Procedures

The query formulations used in this study were generated by volunteer online searchers, all of whom were experienced users of large bibliographic retrieval systems. In order to obtain the multiple query representations, we asked five different searchers to generate Boolean search statements for each of the TREC topics in our analysis. We asked each of our volunteer searchers to generate a query formulation for five different topics, resulting in five indepenenfly generated query formulations for each topic. Af-

ter formulating each query, searchers were asked to answer four questions about the process: how long it took to formulate the query; how related the topic was to their normal searches; how easy it was for them to formulate the query; and, the extent to which they had enough information to construct the query. A total of 75 searchers participated in our study; 50 for the routing topics, and 25 for the ad hoc topics. In addition to the questionnaire items mentioned above, the ad hoc searchers were also asked how many years of online searching experience they had. Searchers for the routing queries were not asked this question. See the Appendix for a sample response sheeL

Our study is based on analysis of the entire set of 50 routing topics, and a selected sample of 25 ad hoc topics. The sample was stratified according to the domain of the topic, in an effort to represent the distribution of domains in the entire set of ad hoc topics.

In our experiments, we used the INQUERY retrieval engine (version 1.5), developed at the University of Massachusetts (Turtle & Croft, 1991). INQUERY is a probabilistic inference network-based system, which is based upon the idea of combining multiple sources of evidence in order to plausibly infer the relevance of a document to a query. The underlying formalism is that of a Bayesian probabilistic inference network ~earl, 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). See the paper by Croft in this volume for more detail on INQUERY.

2.2 Query Combination Experiments

Each of the Boolean query formulations produced by our searchers was translated into INQUERY syntax. Two methods of query combination were then used in our study,

each specific to the TREC-2 tasks of responding to ad hoc

and routing topics. The first, which we label 1tcomblt1 was applied to the ad hoc topics. In this procedure, we simply combine the five query formulations for each topic directly, into one query, using the INQUBRY `tunweighted sumt1 operator. This query is then used as the search statement in our experiments. In the ad hoc search environment, we cannot expect to have relevance judgments, and so we can do no more than simple combination.

The second combinatorial procedure, called 1'combx", was used for the routing topics. Here, we did a separate search for each separate query formulation for all 50 topics, on the training set supplied from the TRBC- 1 data. From these results, we used the average 11-point precision (in the "official" results reporLed at TRBC-2; precision at 100 documents for the "unofficial results" reported in this paper) of each query formulation as a weight for that formulation in the combination of all five formulations for each topic. For this, we used INQUERY's `tweighted sum" operator. This procedure corresponds to constructing a simple combined query, learning something about how that query's components perform on the current database, and taking account of that evidence to modify the query formulation for searching the next database.

These methods of combining queries give us a very straightforward way to test our hypotheses about the effectiveness of multiple sources of evidence. For our experiments (as opposed to the results which were submitted to TRBC-2, which were just the comb I and combx results as described above), we divided the query formulations for both ad hoc and routing topics, into five different groups. In each group, each topic was represented by one query, and no searcher was represented more than once in any one group. This distribution was meant to control for possible searcher effects. We then did runs for each single group, and for each combination of groups, for both ad hoc and routing topics. With these data, we were able to compare retrieval performance of different levels of query combination, and to compare retrieval performance of combined queries with uricombined.

2.3 Data Fusion Experiments

Data fusion was accomplished by a list-merging method which is the natural ex~ension of a 3-out-of-S data

fusion logic in the binary case. The basic data used was the five lists of documents retrieved by the five different query formulations for each topic. Every document has some rank in each of the five lists being joined together. An effective rank is calculated by taking the third highest of the five ranks which the document has. This has the same effect as moving a threshold along the list of effective ranks, and including a document in the output when it has appeared on three of the lists. Since there are five scores all together, this can also be thought of as a median rule.

In practice, to maintain consistency with other parts of our work, we did not calculate the rank of every document, but worked with the lists of the top 1000 documents produced in response to each query formulation. This meant that some documents would appear on all five of the lists, others on just four, or three, or even fewer. Of course, the

whole logic of data fusion suggests that those which appear on more lists are more likely to be relevant. We implemented this, in fact, by forming a combined sort key consisting of (10-degeneracy, 3-rd rank). The degeneracy is the number of lists on which a specific document appears in the top 1000. We used a lexicographic sort, so that all items with degeneracy 5 appeared before any items with degeneracy 4, and so on. Within a given degeneracy, items with lower values for the 3rd rank were ranked first.

3. Results

3,1 Caveats

The results presented in this paper differ in several ways from those submitted as "official" results to TREC-2, which are published at the end of this volume. According to our experimentai design, there are five independently produced query formulations for each of the TREC topics. However, due to uneven return rate among our searchers, we were missing one searcher's set of queries for the ad hoc topics, and three searchers' sets of queries for the routing topics, when we did the "official" runs. Consequently, in the official results, five ad hoc topics and fifteen routing topics are represented by four searches, rather than five. However, we were subsequently able to obtain substitute

searchers, and so for the "unofficial" results presented and discussed in this paper, we have the full complement of 75 searchers and five query formulations per topic.

We were unable to report the data fusion results for routing topics for the official results, because of time constraints. We have subsequently been able to do those runs, and report them here as unofficial results.

We also caution that one query for one of our ad hoc topics is known to have a syntactic error which resulted in very poor performance for that single query, and for all unweighted combinations of queries in which it was present. Therefore, some of our comparative results in the ad hoc case may be slightly incorrect.

3.2 General Results

Because our analyses of ad hoc topics are based on a subset of the total sample, we here consider questions of the sample representativeness. As explained above, the sample was originally chosen to represent topic domains. To see if this had introduced some other bias, we compared the distribution of our 25 topics along the three dimensions of topics proposed by Harman (this volume). These are: broadness, operationally defined as the total number of relevant documents found for that topic; hardness, operationally defined as inverse to the median average precision for that topic; and, restriction, defined according to linguistic characteristics of the topic. The distribution of the 25 topics in our sample did not differ significantly from the total ad hoc topic distribution on any of these dimensions, so we feel reasonably confident that we did not select a markedly biased subset of topics.

Tables 1 and 2 present a descriptive profile of the queries and topics in our study, based upon the query formulations, and the searchers' responses to our questionnaire. Table 1 shows the distribution of numbers of words and operators per query, and also of time required to construct a query. Table 2 shows the distribution of searchers' attitudes to the topics, each indicated on a scale of one to five, from least to most.

[_____	Mean	StdDev.	Mm.	Max.	N
Operators	9.94	5.72	1.00	44.00	375
[Words	19.40	14.63	1.0	145.00	375

LL minutes	11.31	7.48	1.00	40.00	367

Table 1. Characteristics of queries for ad hoc and routing topics.

	Mean	StdDev.	Mm.	Max.	N
Familiarity	1.81	1.15	1.00	5.00	388
Easeof construction	2.82	1.11	1.00	5.00	372
Enough information	3.20	1.11	1.00	5.00	322

Table 2. Characterization of topics by searchers, for routmg and ad hoc topics.

Our ad hoc questionnaire also included a question on how many years of experience each searcher had in online searching. The mean response was 6.8 years. Unfortunately, we do not have these data for the routing searchers.

We wished to consider whether there were any relationships between the various characteristics of queries and topics and the performance of the queries themselves. For this purpose, we constructed a table in which each separate query formulation (75x5=375) is associated with performance measures, the characteristics enumerated in tables 1 and 2, and the three topic categories of broadness, hardness and restriction defined by Harman (this volume). For performance, we considered using one or more of three measures: average of 11-point precision; precision at 100 documents; and R-precision (defined by Harman, this volume). Factor analysis of these three measures showed that a single factor accounts for more than 90% of the variance among them, so that they represent, in effect, a single aspect or factor of performance. The average precision was chosen as representative of this factor, and we have used it both in evaluation of our retrieval results, and in attempting to determine the effect of the other variables we have considered, on retrieval performance. Since this variate does not exhibit a normal distribution, logarithmic and logistic transforms were explored. The logistic leads to a most nearly normal distribution of the transformed score, but we can still not say that the transformed variable follows a normal distribution.

The results of applying ANOVA to seek a predictor of p are shown in Table 3. No significant relations appear. Because of the range of values assumed by the variables Operators, Words and Time, the relation was sought using

regression analysis. Once again, no significant relations were found, and the scatter plots (not included here) make it clear that there is no trend to be found.. Both hardness and broadness are significantly related to performance. The former is expected, since the hardness is determined by median average precision; the latter is less obvious.

Anal sis of variance for lo I 1-

Independent variable	Significance
Familiaritv	0.149
Easiness	0.169
Information	0.907

Table 3. Significance levels of F-tests using ANOVA to seek dependence of the logistically transformed average precision on the searcher's assessments of their query formulation.

The search for relations between average precision and characteristics of the query formulation, whether provided by the search, or determined from the query text itself, was motivated by the results, discussed below, which show that it is desirable to weight formulations in proportion to their average precision. Thus, if we could find a surrogate for average precision which can be known without evaluating the retrieved documents, it would be possible to approximate the effective combination on the first pass of a retrieval operation. This hope is frustrated at this time.

3.3 Query Combination and Data Fusion

Results: Ad hoc Topics

The official results reported to JkBC-2 were for the overall performance of each of two treatments for the ad hoc topics, and of one treatment for the routing topics. For those results, we refer the reader to the relevant section of this volume. Here we report on our further investigations

on the effect of combination of queries, and of data fusion, on performance.

Our first investigation in query combination was to see if combining query formulations has a regular, beneficial effect, as hypothesized. To do this, we generated the five different search groups for the ad hoc topics, as described in section 2.2, and did experimental runs on all single query groups, all 2-way combinations of queries, all 3-way combinations of queries, all 4-way combinations of queries, and the combination of all 5 query formulations. The results are presented in Table 4, where it is evident that the average performance increases monotonically as more evidence is added. The increase is strict and significant, as shown in Table 4a, where we display the number of times that each combination level performed better than each other level. We note that the data fusion results are not significantly better than any but 1-way combination (that is, average performance for single queries), but also that its performance is not significantly different from unweighted 5-way combination

1-way ~2:w~ 3-wa~ 4-way 5-way fusion
0.1441 0~~.2~016~ 0.2235 0.2361 0.2349 0.2042
0.1571 0~~L82~ 0.2304 0.2225
0.1121 9~.2051 0.2102 0.2292
0.1589 0.1951 0.2043 0.2200
0.1378 0.1763 ~79 0.2166
O~..2L13 0.2171: -
0~.1~727 0.2172
0.1~683 0.1873:
0~.1633 0.2116~~~
_____ 0.1885 0.1934

0.14200.18640.2103 0.22490.23490.2042

For example, column two represents the ten possible ways of choosing two groups of query formulations from the collection of five groups. Each entry is an average over 25 topics.

Table 4. For ad hoc topics, average 11-point precision, by group, for each combination of queries, and mean average precision for all groups at each level of combination.

---

*~	1-way	2-way	3-way	4-way	5-way	fusion
~1-way		1**	3**	3**	3**	5**
~2-way	24**		5**	6**	5**	9
3-way	22**	20**	*****	3.5**		11
4-way	22**	19**	19**		3**	12
5-way	22**	20**	21.5**	22**		15
fusion	20**	16	14	13	10	

**= significant difference at p < .01, sign test
*= significant difference at p < .05, sign test
Read row with respect to column, e.g. 2-way performed better than 1-way 24 out of 25 times, or 1-way performed better than 2-way 1 out of 25 times

Table 4a. Number of times, for average performance of combinations for ad hoc topics, that one treatment performed better than another.

The results presented in Tables 4 and 4a are based on the average performance for the query formulations in any one set. In Tables 5 and 5a, we present data on performance, for ad hoc topics, when only the best query formulation, or best combination of query formulations, for each topic is used. These results are compared with the single 5-way combination (which is the only combination possible at this level with our data), and with the fusion results. It is of some interest to note that the ranring of level of combination is now very much different than that for average performance, with 2-way and 3-way combination being signficanfly better than 1-way, 4-way, 5-way and fusion (see Table Sa).

1-way	2-way I	3-way I	4-way I	5-way I	fusion I
0.2712 I	0.3002 I	0.2959 I	0.2702 I	0.2350 I	0.2042 I

Table 5. For ad hoc topics, mean 1 1-point precision for best-performing combination of queries for each topic.

	1-way	2-way	3-way	4-way	5-way	fusion
1-way	_____	6**	7*	13	17	21**
2-way	19**		16.5	20**	22**	24**
3-wa	18*	8.5		20.5*.	22**	22**
4-way	12	5**	4.5*.		22.5**	20**
5-way	8	3**	3**	2.5**	_____	15
fusion	4**	1**	3**	5**	10	

**= significant difference at p < .01, sign test
* = significant difference at p <~ .05, sign test
Read row with respect to column, e.g. 2-way performed better than 1-way 19 out of 25 times, or 1-way performed better than 2-way 6 out of 25 times

Table Sa. Number of times, for performance of best combinations for ad hoc topics, that one treatment performed better than another.

3.4 Adaptive Combination: Ad hoc Topics

Finally, to get an overall idea of how query combination in the ad hoc case worked, and to estimate whether taking account of the evidence of scaech performance could improve subsequent performance, we compared performance of simple combination of all five query formulations (comb 1) with performance when only the best single query formulation for each topic was used (best), with combination of all five query formulations weighted according to the precision at 100 documents retrieved, of each formulation (comby). The results, reported in Tables 6 and 6a, show that there is no significant difference between combi and best, but that comby is significanfly better than comb 1. While formation of comby would not be possible under the conditions of the ad hoc ~fl~EC task, these results are of interest because they simulate the kind of operations that could be implemented in a fully interactive interface to an IR system.

I combi	best	I comby	I fusion I
0.2350	0.2712	0.2819	0.2042

combi = unweighted combination of all queries for each topic
best = best performing query for each topic
comby = weighted (by prec.@ 100 docs) combination of all queries for each topic

Table 6. For ad hoc topics, mean 11-point precision for four treatments.

In reading Tables 6 and 6a, note that the choice referred to as "best" corresponds exacdy to the choice called "1-way" in Table 5. However, it does not correspond to any of the

entries in the first column of Table 4. The entries in Table

4 refer to combinations based upon the fixed groups. But, the combination of groups which performs best for, say, Topic 57, need not be the one which performs best for Topic 72. In Tables 6 and 6a, the best possible combination is chosen for each topic individually. Note also that, in the INQUERY system, the "unweighted sum" corresponds to a symmetrical assignment of each weight to all formulations.

I combi I
best
I combv

	1-way	2-way	3-way	4-way	5-way	fusion
1-way		3**	2**	1**	2**	8**
2-way	47**		5.5**	6**	5.5**	13**
3-way	48**	44.5**		9**	7**	18*
4-way	49**	44**	41**		8**	22.5
5-way	48**	44.5**	43**	42**		28
fusion	42**	37**	32*	27.5	22	

I combi I best I comby I fusion I ** = significant difference at p <- .01, sign test
I 8 4** 115 I * = significant difference at p < .05, sign test
117 I I 9 I 21** I Read row with respect to colunm, e.g. 2-way performed better
I 21** 116 I I 20** I than 1-way 47 out of 50 times, or 1-way performed better than
I ~ I I 2-way3outof5o times

I fusion 110 I 4**

** = significant difference at p < .01, sign test
* - significant difference at p < .05, sign test
Read row with respect to column, e.g. comby performed better than combi 21 times , or combi performed better than comby 4 times.

Table 6a. Number of times that one treatment for ad hoc topics performed better than another.

3.5 Query Combination and Data Fusion

Results: Routing Topics

We ran further experiments on the routing queries, analogous to those we used for the ad hoc queries. Our first set of results shows the progressive effect of unweighted combination of query formulations, by level of combination, when average performance at each level is considered (tables 7 and 7a). Again, as for the ad hoc queries (tables 4 and 4a), there is a progressive, significant effect of level of query combination. For the routing queries, data fusion appears to have a somewhat stronger effect than for ad hoc, being significantly better than 1-, 2- and 3-way combination. It is of some interest to note that the overall level of performance for routing topics is much higher than for the ad hoc topics.

1-way	2-w~ay~	3-way	4-way	5-way	fusion
0.1763	0.2311	0.2599	0.2619	0.2807	
0.1890	0.2202	0.2503	0.2748		
0.1684	0.2258	0.2603	0.2735		
0.2025	0.2229	0.2314	0.2512		
0.1793	0.2436	0.2415	0.2745		
	0.2364	0.2471			
	0.2388	0.2509			
	0.2160	0.2654			
	0.2149	0.2642			
	0.2338	0.2417			

~

Each entry is an average over 50 topics.

Table 7. For routing topics, average 11-point precision, by group, for each combination of queries, and mean average precision for all groups at each level of combination.

Table 7a. Number of times, for average performance of combinations for routing topics, that one treatment performed better than another.

As for the ad hoc topics, we then compared the results of the best query formulation combinations for each level of combination, with the unweighted 5-way combination, and fusion results. As for the ad hoc queries, this gave us quite a different ranking of levels of combination, with 3-way and

2-way combinations being significantly better than all others, and 4-way being significanfly better than 5-way and fusion (tables 8 and 8a).

I 1-way I 2-way I 3-way I 4-way I 5-way I fusion I
I 0.29311 0.31731 0.31991 0.30691 0.28071 0.26611

Table 8. For routing topics, mean 11-point precision for best-performing combination of queries for each topic.

	1-way	2-way	3-way	4-way	5-way	fusion
1-way		8.5**	13.5**	22	29	36**
2-way	41.5**		20.5	34*	38**	39**
3-way	36.5**	29.5		37**	42**	45**
4-way	28	16*	13**		44**	40**
5-way	21	12**	8**	6**		28
fusion	14**	11**	5**	10**	22	

** = significant difference at p <~ .01, sign test
* = significant difference at p <.05, sign test
Read row with respect to column, e.g. 2-way performed better than 1-way 41.5 times, or 1-way performed better than 2-way 8.5 times

Table 8a. Number of times, for performanGe of best combinations for routing topics, that one treatment performed better than another.

3.6 Adaptive Combination: Routing Topics

Finally, we wished to investigate the effectiveness of progressively taking account of retrieval performance in

modification of the query formulation. To do this, we compared performance of unweighted 5-way query combination (combi) with performance using the best- performing query formulations in the training database (1)estl), the best performing query formulations in the test database (bes~), the weighted 5-way query combination using weights from the training database (combx), the weighted 5-way query combination using weights from the test database (comby), and 5-way query combination weighted by the mean of the

weights for test and training databases. The weights that we used were the precision at 100 retrieved documents for each query formulation. In the official results, we used average 11-point precision. The reason for the change, is that precision at some cutoff level is a realistic measure for the routing task in general, and especially in an operational environment, whereas the average precision is a measure that we cannot realistically expect to have in an operational environment. When we compared the performance of both weights in the combx formulation, there was no significant difference. The results are presented in tables 9 and 9a, and show that talting account of subsequent evidence has a positive and significant effect on performance. When reading Tables 9 and 9a, note that the entries for combi and fusion have already appeared in Table 7, as 1'5-way11 and "fusion", respectively. Also, "best2" has already appeared in Table 8, as the best `11-way't combination.

	comi	bestl	best2	comx	comy	cxv	fus
comi		~29	21 **	13.5	16* **	14.5	28
besti	21		13**	14** **	14**	12.5	22.5
best2	29	37**		23	20	23	36**
comx	36.5	36** **	27		21.5	18**	40**
comy	34*	36**	30	28.5 **		25.5	36.5
cxy	35.5 **	37.5 **	27	32*	24.5		37**
fus	22	27.5	14** **	10**	13.5	13**	

** = significant difference at p < .01, sign test
* = significant difference at p < .05, sign test
Read row with respect to column, e.g. combx performed better than combi 36.5 times, or combi better than combx 13.5 times.

Table 9a. Number of times that one treatment for routing topics performed better than another.

I combl I	besti I	best2 I	combx I	comby I	comxy I	fusion I
.2807	.2721	.2931	.3012	.3090	.3068	.2661

combi = unweighted combination of all queries for each topic
besti = best performing query (on training set) for each topic

bes~ = best performing query (on test set) for each topic
combx = weighted (by prec.@ 100 docs in training set) combination of all queries for each topic
comby = weighted (by prec.@ 100 docs in test set) combination of all queries for each topic
combxy = weighted (by mean of the sum of prec.@ 100 docs in training and test sets) combination of all queries for each topic

Table 9. For routing topics, mean 11-point precision for seven treatments.

Table 9a encapsulates all of the key concepts of the several approaches to combination that we have explored. We have two approaches which are a priori and symmetric in their treatment of the query formulations (fus and combi). As expected, the fusion system, using the least information, performs worse. comb 1, the symmetric formulations does better, although the difference is not statistically significant. Both of these methods often perform better than the best of the individual formulations, and their relations to other combination schemes are (except for the relation to bes~) quite similar. The query ~hat performs best on the training set (besti) does not perform significantly better than any of the combination schemes. But that formulation which performs best on the test set (best2, also called 1-way in Table 8) is significantly better than besti and the fusion scheme.

Of greater interest are the methods representing adaptive weighting schemes: combx, comby and combxy. Most significantly, combx, the adaptive weighting formulation, is better than the symmetrically weighted combination (comb 1), the fusion rule, and the best single formulation in a substantial fraction (over 70%) of all cases. The weighting based on the test set (comby) stands also in essentially the same relation to those three other schemes. Finally, the weighting scheme combxy simulates a situation which might arise in updating or tuning a combination rule after two batches of documents have been retrieved. This is accomplished by averaging the weights assigned to each formulation in the training run, with those assigned based on the test run. This scheme shows essentially the same pro-

file as combx and comby when compared with the comb 1, fusion, besti and bes~ schemes. It performs significantly better than combx, but not significantly better than comby.

4. Discussion

4.1 General Results

As is customary, we begin this section with a general disclaimer. In this case, we need to point out that all of our results were obtained with a very specific kind of query formulation technique and very special kinds of queries, and, that all of our results were obtained within a very special retrieval context, the INQUERY system. It is certainly p05-sible that these circumstances strongly affected our results, so that we cannot make widely general claims for them. On the other hand, the results reported by Fox and Shaw (this volume), using queries generated in quite different ways, and using a quite different IR system and retrieval technique, are quite similar in general form and trend to

ours, although their specific figures are different. So we are willing to believe that the influence of our experimental situation is probably not enough to invalidate our results at some level of generality.

There are several aspects of our general results which are of some interest, apart from the issues of query combination and data fusion. One has to do with the lack of any significant relationship between number of words in a query, and the performance of a query. It has been at least informally suggested in the IR community, that the retrieval performance of queries increases with the number of words in the query. There is no support in our data for this hypothesis. Indeed, in our data, there is at least one one-word query, which performed better than all of the other, multi-word queries for that topic.

It is also of interest that none of the query/searcher characteristics was related to performance. This may be a characteristic of our particular data set, but it also suggests that it will be rather difficult to identify characteristics of people or topics at this level, which will be predictive of performance of the query.

Mthough the level of familiarity by the searchers on the topics was in general rather low, our searchers nevertheless found it not too difficult to formulate queries (mean of 2.82 on a scale of 1 to 5), and felt that they had sufficient information to construct a reasonable query, on the basis of the topic (mean of 3.2 on a scale 1 of 5). This makes us

think that the queries are likely to be reasonable formulations of the search topics, at least as far as the searchers are concerned. But the range and variability of the numbers of words and numbers of operators per topic seems to indicate that the query formulations themselves are rather different (we have not yet compared them for overlap in specific words, but work on this issue is in progress). These two results seem to us to confirm our initial i6ea that each query formulation is indeed a "different" interpretation of the information problem, and thus to substantiate our general approach.

4.2 Query Combination Results

Our results, for both ad hoc and routing topics, seem clearly to show that, in general, the more evidence one has, and uses, in the form of different query formulations, the better the IR performance is going to be. In particular, tables 4, 6, 7 and 9 support this conclusion, in various respects. From the results of tables 6 and 9, we can see that, taking advantage of what one learns about query performance from one iteration doesn't help a lot, after the first iteration, but on the other hand, it doesn't hurt, either. This suggests to us that continual modification and reweighting of the multiple query formulations in a combined query, is likely to be useful in the general routing environment. But even doing it once, given the initial evidence, seems to help. This also suggests that continuing to add new query formulations to a combined query will likely help performance on subsequent runs.

Having said all this, it is worth considering the results

of tables 5 and 8, which showed that picking the best 2-way or 3-way combination of query formulations was significantly better than using 4-way or 5-way combinations. On the face of it, this runs counter to the general result of "the more, the better". However, it is possible that this result is an artifact of our data. For both 2;way and 3-way combinations, it was possible to choose the best from ten different combinations. Because we had only five different query formulations for each topic, we had smaller pools from which to choose, for both single query formulations, and

for the 4-way and 5-way combinations. This issue needs further investigation.

4.3 Data Fusion Results

There are several points to be made with regard to the median-fusion scheme as implemented here. First, as may be expected from general arguments, it sometimes performs better than the best of the lists which are joined by the fusion process. Second, it does not perform as well as even the symmetric (unweighted) combinations made using the internal scores generated by the INQUERY system. This is expected, since those scores contain more information than the rankings alone. One can imagine special cases in which the distribution of scores assigned to a document by several queries is such that the internal combination rule of unweighted sum does not perform as well, but this has apparently not occurred in the cases studied here.

Third, in the application to the routing problem we have, in fact, operated in a batch mode. For a true routing situation, it would be necessary to estimate cutoff scores for the several query formulations, corresponding to the cutoff rank on the fused list. For large data sets this can be done easily. Without this step, it is not possible to make an immediate decision about a newly presented document.

Fourth, the stability induced by using this system was manifested in the case of the one query for which we discovered, too late to make the change, that one of the query formulations was in error. For this case, all of the "average combination of evidence formulations" performed more poorly than the fusion rule. This is because one, or even two disastrously bad query formulations will have little effect on the results of the 3-of-S fusion rule. Of course, expect for the case of combining all five query formulations, the best of one, two, three or four query formulations can do well because the one bad formulation will be missing from the combinations that are best.

Finally, the application of data fusion here, at the so-called decision level (that is to say, after the documents have been ranked according to several rules) is a simulation for the case to which it should be applied. Since the specific system that we used permits internal manipulation of scores, there is no need to delay combination until after the output lists have been formed. But in realistic settings, several distinct systems will have internal operations which are not compatible, so that, even if it were possible to extract the internal scores, it would not be apparent how to

combine them.

5. Conclusions ARPA.

In general, we conclude that our initial research questions with respect to query combination have been positively answered. That is, if one has available several different representations of a single information problem, then it makes sense to use all of them, in combination, in order to improve retrieval performance, rather than to try to identify and use only the best one. In addition, it is reasonably clear that progressive and continuous combination of query formulations leads to continuing and progressive improvement of performance. This may extend to progressive modification of query formulations in the routing situation, for instance, on the basis of each iteration of retrieval. Nevertheless, some of our results appear anomalous, and in particular we need to address more carefully the issue of how best to combine query formulations.

As far as our data fusion questions are concerned, we have clearly demonstrated that doing data fusion is better than using only one query formulation. Although performance improvement in these experiments was rather low, for operational settings in which there are multiple systems with incompatible scores, a data fusion method that works with the ranked outputs, rather than the scores is the precise method that is needed. In the present study we have shown how that method can be extended from the case of binary (set) retrieval to the case of ranked lists. We have shown that the results are, on the average, better than the results of the individual formulations. In some cases, they are better than the best of the component formulations. This lends support to a program of seeking optimal tunings for fusion of any number of given systems, to achieve results better than any of them alone could provide.

Overall, we find strong support for adaptive weighting in query combination. This is applicable to both routing, as shown direcdy here, and to relevance feedback, which we have simulated in our application to the ad hoc topics. We also find strong support for enlarging the set of query representations. This success raises many interesting possibilities. For example, one might systematically explore the k-way combinations to see how they compare to the adaptive weighting scheme. Or, one might apply the notion of adaptive weighting to the best of the k-way combinations. The possibilities for combining these two concepts ex-

plodes (of course) combinatorially. We feel that the present experiments point a way into the forest of possibilities.

6. Acknowledgments

We wish to thank the 75 searchers who so generously donated their time and effort to this projecL Without them this research could not have been done. We also wish to thank Audrey Gorman and Kathy Mrowka, who provided invaluable assistance on the project in planning, data gathering and input, and Dong Li, who helped with the data analysis. We owe special thanks to Bruce Croft and Jamie Callan, not only for permission to use the INQUERY system for this investigation, but also for the unstinting support they gave us in using it. This research was performed with partial funding from a TREC support grant from the

7. References

BELKIN, N.J., COOL, C., CROFT, W.B. & CALLAN, J.P. (1993). The effect of multiple query representations on information retrieval performance. In: Proceedings of the 16th International Conference on Research and Development in Information Retrieval (SIGIR `93), Pittsburgh, 1993. New York, ACM: 339-346.

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, ODDY, R.N. & BROOKS, H.M. (1982) ASK for information retrieval. Journal ofDocumentation, 38, 2&3: 61-71, 145-164.

FOX, E.A. et al. (1993) Combining evidence from multiple searches. In: D.K Ilarman, ed., The First Text REtrieval Conference (TREC-1). GPO, Washington, D.C.: 319-328.

KANTOR, P. (1993) Vector space models of data combination in information retrieval. Technical Report APlabfrR-93-3. New Brunswick, NJ., Rutgers University, School of Communication, Information & Library Studies.

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.

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

TAYLOR, R.S. (1968) Question negotiation and information seeking in libraries. College and Research Libraries, 29:178-194.

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

VAN RIJSBERGEN, CJ. (1986) A new theoretical framework of information retrieval. In: Proceedings of the 1986 International Conference on Research and Development in Information Retrieval (SIGIR `86), Pisa, 1986. New York, ACM: 194-200.

APPENDIX: SEARCHER RESPONSE SHEET

QUERY FORMULATION

TOPIC NUMBER:_____ PACKET NUMBER:_____

Please formulate a search query for one of the five topics you have been sent, in the space below. Don't
forget to indicate the topic number in the space above. Please read the entire topic description before you
begin your query forumlation. (Use the back of this sheet, if there isn't enough room below)

Please answer the following questions, as they relate to this specific query formulation.

1. About how many minutes did it take (including reading the topic description)?
_____ minutes

2. Is this topic related to things you normally search on (please circle one number)?

1 ----------2-------3----------4---------5
Not at all Somewhat Very much

3. How easy was it to formulate this query?

1----------2-------3----------4---------5
Not easy Somewhat Very easy

4. Do you feel you had enough information to construct an effective query?

1----------2-------3----------4---------5
Too little Adequate Plenty

5. About how many years have you been doing online searching? years.