Combination of Multiple Searches

Edward A. Fox and Joseph A. Shaw
Department of Computer Science
Virginia Tech, Blacksburg, VA 24061-0106

Abstract

The TREC-2 project at Virginia Tech focused on methods for combining the evidence from multiple retrieval runs to improve retrieval performance over any single retrieval method. This paper describes one such method that has been shown to increase performance by combining the similarity values from five different retrieval runs using both vector space and P-norm extended boolean retrieval methods.

1 Overview

The primary focus of our experiments at Virginia Tech involved methods of combining the results from various divergent search schemes and document collections. We performed both routing and ad-hoc retrieval experiments on the provided test collections. The results from both vector and P-norm type queries were considered in determining the probability of relevance for each document in an individual collection. The results for each collection were then merged to create a single final set of documents that would be presented to the user.

2 Index Creation

This section outlines the indexing done with the document collection provided by NIST. Each of the individual collections was indexed separately as document vector files; limitations in disk space prohibited the use of inverted files and the creation of a single combined document vector file.

All processing was performed on a **DECstation 5000/25** with 40 MB of RAM using the 1985 release of the **SMART** Information Retrieval System [2], with enhancements from previous experiments as well as a new modification for our TREC-2 experiments.

The index files were created from the source text via the following process. First, the source document text provided by NIST was passed through a preparser to convert the SGML-like format to the proper format for the 1985 version of **SMART**. The extraneous sections of the documents were filtered out at this point. The TEXT sections of the documents, as well as the various HEADLINE, TITLE, SUMMARY, and ABSTRACT sections of the collections were indexed; all of the other sections were ignored. The subsections of the TEXT fields, where they existed, were considered as part of the TEXT field, with the subsection delimiters removed.

Table 1: **SMART** weighting schemes used for TREC-2.

SMART label	**term_weight =**
ann	$0.5 + 0.5 * \frac{tf}{max_tf}$
bnn	1
mnn	$\frac{tf}{max_tf}$
atn	$0.5 + \frac{tf}{2*max_tf} * log(\frac{num_docs}{coll_freq})$
nnn	tf

The resulting filtered text was tokenized, stop words were deleted using the standard 418 word stop list provided with **SMART**, and the remaining non-noise words were included in the term dictionary along with their occurrence frequencies. Each term in the dictionary has a unique identification number. A document vector file was created during indexing which contains for each document its unique ID, and a vector of term IDs and term weights. The initially recorded weights can be changed based on one of several schemes after the indexing is complete. The various **SMART** weighting schemes referred to within this paper are summarized in Table 1. The dictionary size for each collection was approximately 16 MB, while the document vector files ranged from 31 MB to 124MB (see Table 2).

Table 2: Collection statistics summary. Text, Dictionary and Document Vector sizes in Megabytes.

Collection	Text	Dict.	Doc. Vectors	Total Doc.s
AP-1	266	16.0	120.2	84678
DOE-1	190	15.9	97.9	226087
FR-1	258	15.8	53.8	26207
WSJ-1	295	16.2	124.8	98735
ZIFF-1	251	15.7	88.4	75180
D1	1260	N/A	485.1	510887
AP-2	248	15.9	110.4	79923
FR-2	211	15.6	42.7	20108
WSJ-2	255	16.0	105.5	74520
ZIFF-2	188	15.4	63.6	56920
D2	902	N/A	322.2	231471
D1 & D2	2162	N/A	807.3	742358
AP-3	250	15.9	111.2	78325
PATN-3	254	15.6	31.3	6711
SJM-3	319	16.1	114.4	90257
ZIFF-3	362	16.0	109.8	161021
D3	1185	N/A	366.7	336314
Total	3347	N/A	1174.0	1078672

3 Retrieval

3.1 Queries

All of the queries were created from the topic descriptions provided by NIST. Two types of queries were used, P-norm extended boolean queries and natural language vector queries. A single set of P-norm queries was created, but was interpreted multiple times with different operator weights (P-values). Two different sets of vector queries were created from the topics, one containing information from fewer sections of a topic description. The Title, Description and Concepts sections of the topic descriptions were used in the creation of all three sets of queries, the Definitions section was used also in both sets of vector queries, while the P-norm query set and one of the vector query sets also contained information from the Narrative section of the topic descriptions. The vector query set that included the Narrative section of the topic is referred to as the long vector query set, for obvious reasons, while the other is referred to as the short vector query set.

The P-norm queries were written as complex boolean expressions using AND and OR operators. Phrases were simulated using AND operators since the queries were intended only for soft-boolean evaluation. The query terms were not specifically weighted; uniform operator weights (P-values) of 1.0, 1.5 and 2.0 were used on different evaluations of the query set.

Table 4: Summary of the five individual runs.

Title	Query Type	Similarity Measure
SV	Short vector	Cosine similarity
LV	Long vector	Cosine similarity
Pn1.0	P-norm	P-norm, P = 1.0
Pn1.5	P-norm	P-norm, P = 1.5
Pn2.0	P-norm	P-norm, P = 2.5

3.2 Individual Retrieval Runs

The first step in our TREC-2 experiments involved determination of what weighting schemes would be most effective for P-norm queries. Our TREC-1 experiments with P-norm queries had obtained mixed results, performing poorly based on binary document term weights in our Phase I experiments and performing well for a P-value of 1.0 and very poor with larger P-values in our Phase II experiments using a *tf-idf* weighting scheme [4]. We performed several P-norm retrieval runs on the two AP and two WSJ training collections with topics 51 to 100 to determine the most effective term weighting scheme for P-norm queries with large test collections. The results from these experiments are shown in Table 3 using the standard TREC-2 average non-interpolated precision and the exact R-precision measures. The most effective weighting scheme turned out to be the **SMART** *ann* weighting scheme, which confirmed the result obtained originally by Fox for the much smaller classical document collections [3].

The two sets of vector queries were evaluated using the standard cosine correlation similarity method as implemented by **SMART**. The same **SMART** *ann* weighting scheme used for the P-norm queries was used on the vector queries for several reasons. First, a weighting scheme that did not use any collection statistics was needed for the routing experiments. Second, the methods used in combining runs described in the next section required a similar range of possible similarity values produced by each run. Finally, the necessity of merging results from each collection into a single set of results was simplified since the resulting similarity values were not based on collection statistics which would have differed for each collection. The P-norm queries were evaluated using three different P-values, again using the **SMART** *ann* weighting scheme based on specific P-norm experiments described below. The five individual runs are summarized in Table 4.

The five individual runs were performed and evaluated for each of the nine training collections on topics 51 to 100. The results for these experiments are given in

Table 3: Average Precision and Exact R-Precision for P-norm experiments on weighting with the AP and WSJ collections (Ad-hoc Topics 51-100).

Coll.	P-value	Average Precision			R-Precision		
		ann	bnn	mnn	ann	bnn	mnn
AP-1	1.0	**0.2810**	0.2419	0.1419	**0.2688**	0.2660	0.1689
	1.5	**0.3122**	0.2581	0.1444	**0.2976**	0.2732	0.1757
	2.0	**0.3027**	0.2510	0.1457	**0.2968**	0.2775	0.1707
AP-2	1.0	**0.3004**	0.2672	0.1826	**0.3165**	0.2864	0.2046
	1.5	**0.3332**	0.2999	0.1831	**0.3412**	0.3118	0.2161
	2.0	**0.3300**	0.2922	0.1847	**0.3339**	0.3057	0.2284
WSJ-1	1.0	**0.2941**	0.2485	0.1742	**0.3221**	0.2830	0.2181
	1.5	**0.3199**	0.2753	0.1774	**0.3443**	0.2994	0.2225
	2.0	**0.3217**	0.2752	0.1776	**0.3470**	0.3013	0.2277
WSJ-2	1.0	**0.2206**	0.1881	0.1356	**0.2367**	0.2094	0.1722
	1.5	**0.2327**	0.2013	0.1174	**0.2511**	0.2234	0.1549
	2.0	**0.2325**	0.1970	0.1098	**0.2442**	0.2158	0.1445

Table 5. In general, the P-norm queries performed better than the vector queries. The most effective P-value however differed between the collections: The AP runs performed better with a P-value of 1.5, while a P-value of 2.0 performed better for the WSJ collections.

3.3 Combination Retrieval Runs

Our experiments in TREC-1 involved combining the results from several different retrieval runs for a given collection either simply taking the top N documents retrieved for each run, or modifying the value of N for each run, based on the eleven point average precision for that run. We felt these efforts suffered from considering only the rank of a retrieved document and not the actual similarity value itself. In TREC-2, our experiments concentrated on methods of combining runs based on the similarity values of a document to each query for each of the runs. Additionally, combining the similarities at retrieval time had the advantage of extra evidence over combining separate results files since the similarity of every document for each run was available instead of just the similarities for the top 1000 documents for each run. While our results for four of the training collections indicated that the P-norm queries performed better than the vector queries, this result was likely specific to the actual queries involved and not necessarily true in general. This lead to a decision to weight each of the separate runs equally and not favor any individual run or method. In general, it may be desirable or necessary to weight a single run more, or less, depending on its overall performance; this could be especially useful in a routing situation.

For any given information retrieval ranking metohd, there are two primary types of errors that can occur:

Table 6: Formulas for combining similarity values.

Name	$Combined\ Similarity =$
CombMAX	$MAX(Individual\ Similarities)$
CombMIN	$MIN(Individual\ Similarities)$
CombSUM	$SUM(Individual\ Similarities)$
CombANZ	$\frac{SUM(Individual\ Similarities)}{Number\ of\ Nonzero\ Similarities}$
CombMNZ	$SUM(Individual\ Similarities) * Number\ of\ Nonzero\ Similarities$
CombMED	$MED(Individual\ Similarities)$

assigning a relatively high rank to a non-relevant document, and assigning a relatively low rank to a relevant document. It has been shown that different retrieval paradigms will perform differently on the same set of data, often will little overlap in the set of retrieved documents. [5] For instance, when one retrieval method assigns a high rank to a non-relevant document, a different retrieval method is likely to assign that document a much lower rank. Similarly, when one retrieval method fails to assign a high rank to a relevant document, a different retrieval method is likely to assign that document a high rank. This characteristic of information retrieval methods indicates that some method for considering both retrieval methods together should help to decease the probability of this happening; of course, it is also possible for both methods to highly rank a non-relevant document or to poorly rank a relevant document.

Six methods of combining the similarity values were tested in our TREC-2 experiments, as summarized in

Table 5: Average Precision and Exact R-Precision for the five individual runs (Ad-hoc Topics 51-100).

Average non-interpolated Precision

Run	Disk 1					Disk 2				Both Disks
	AP	**DOE**	**FR**	**WSJ**	**ZF**	**AP**	**FR**	**WSJ**	**ZF**	
SV	0.2387	0.0605	0.0222	0.2203	0.1026	0.2543	0.0330	0.1503	0.0770	0.1418
LV	0.2435	0.0586	0.0302	0.2414	0.0864	0.2664	0.0324	0.1633	0.0753	0.1555
Pn1.0	0.2605	0.0658	0.0611	0.2941	0.1110	0.3004	**0.0879**	0.2206	0.1003	0.1988
Pn1.5	**0.2939**	0.0771	0.0639	0.3199	**0.1278**	**0.3332**	0.0878	**0.2327**	0.1065	0.2242
Pn2.0	0.2849	**0.0847**	**0.0706**	**0.3217**	**0.1278**	0.3300	0.0865	0.2325	**0.1136**	**0.2250**
CombSUM	0.3493	0.1001	0.0741	0.3605	0.1475	0.3748	0.0842	0.2752	0.1273	0.2620
Chg/Max	18.84%	18.18%	4.95%	12.06%	15.41%	12.48%	-4.20%	18.26%	12.05%	16.44%

Exact R-Precision

Run	Disk 1					Disk 2				Both Disks
	AP	**DOE**	**FR**	**WSJ**	**ZF**	**AP**	**FR**	**WSJ**	**ZF**	
SV	0.2624	0.0564	0.0183	0.2616	0.1180	0.2649	0.0202	0.1744	0.0922	0.2169
LV	0.2672	0.0493	0.0274	0.2800	0.0802	0.2704	0.0176	0.1860	0.0843	0.2311
Pn1.0	0.2688	0.0661	0.0533	0.3221	0.1123	0.3165	0.0971	0.2367	0.0969	0.2708
Pn1.5	**0.2976**	0.0762	0.0572	0.3443	0.1218	**0.3412**	**0.1016**	**0.2511**	0.1068	0.2962
Pn2.0	0.2968	**0.0765**	**0.0654**	**0.3470**	**0.1254**	0.3339	0.0820	0.2442	**0.1158**	**0.3008**
CombSUM	0.3590	0.0950	0.0619	0.3767	0.1357	0.3732	0.0887	0.2851	0.1216	0.3292
Chg/Max	20.63%	24.18%	-5.35%	8.55%	8.21%	9.37%	-12.69%	13.54%	5.00%	9.44%

Table 6. The rational behind the CombMIN combination method was to minimize the probability that a non-relevant document would be highly ranked, while the purpose of the CombMAX combination method was to minimize the number of relevant documents being poorly ranked. There is an inherent flaw with both of these methods; namely, they are specialized to handle specific problems without regard to their effect on the other retrieved documents: for example, the CombMIN combination method will promote the type of error that the CombMAX method is designed to minimize, and vice versa. The CombMED combination method is a simplistic approach to handling this, using the median similarity value to avoid both scenarios. What is clearly needed is some method of considering the documents' relative ranks, or similarity values, instead of simply attempting to select a single similarity value from a set of runs. To this end, we tried three other methods of combining retrieval methods. CombSUM, the summation of the set of similarity values, or, equivalently, the numerical mean of the set of the set of similarity values; CombANZ, the average of the non-zero similarity values, that ignores the effects of a single given run or query failing to retrieve a relevant document; and CombMNZ to provide higher weights to documents retrieved by multiple retrieval methods. Clearly, there are more possibilities to consider; the advantages of those chosen are simplicity, in terms of both execution efficiency and implementation, and generality, in terms of not being specific to a given method or retrieval run.

These six methods were evaluated against the AP and WSJ test collections for topics 51 through 100, combining the similarity values of each of the five individual runs specified above. The results are shown in Table 7 below the results of each of the corresponding individual runs from Table 5. Note that while the CombMAX runs performed well compared with most of the individual runs, they did not do as well as the best of the individual runs in most cases. The CombMIN runs performed similarly for the AP collection, but performed worse than every individual run for the WSJ collection.

The CombANZ runs and the CombMNZ runs both performed better than the best of the individual runs, with the CombMNZ runs performing only slightly better than the combANZ runs for three of the four collections, and performing basically the same for the fourth. The primary reason for the similar performance of the two runs is that the two methods produce the same ranked sequence of for all the documents retrieved by all five individual runs. Thus, the

The CombSUM retrieval run was performed for each of the nine collections on the two training CD-ROMs. The results are shown in Table 5. Breaking this analysis down to a per topic basis in Table 11, it can be

seen that the CombSUM method performs significantly better than the best single individual run, Pn2.0; a two-tailed paired t test on the CombSUM and Pn2.0 average precisions results in a p value of 3.1e-05, which indicates these results are conclusive. However, comparing the CombSUM results with the best individual runs for each query basis, results in a p value of approximately about 0.16, indicating that there is a 16 percent chance that the CombSUM method is no better than the best individual run, Pn2.0, for any specific query. Performing the same calculation on the R-Precision results in similar significance findings.

While combining all five runs produced an overall improvement in retrieval effectiveness over each of the runs, the same does not always hold true when combining only two or three runs. Each of the ten combinations of two CombSUM runs was performed for both of the AP test collections, as well as a run combining all three of the P-norm runs. The results of these are given in Table 8. Most of the combinations of two runs performed worse than the better of the two runs while performing better than the poorer of the two runs. One notable exception to this is the combination of the two vector runs, which performed noticeably poorer than either of the two runs.

3.4 Collection Merging

The retrieval results for each of the collections were combined by simply merging the results based solely on the combined similarity values. Since the retrieval runs were based on term weights without collection statistics such as inverse document frequency, the similarity values were directly comparable across collections. The results of merging the CombSUM results by summed similarity value for both disks, is shown in the last column of Table 5.

4 TREC-2 Results

The procedure described above was used for both our official TREC-2 routing and ad-hoc results. The exact queries for ad-hoc topics 51 to 100 used for testing our above method were used for the routing queries against the new collections on disk 3. The results obtained from performing the CombSUM retrieval runs for each of the four collections as well as the merged results are shown in Table 9. The two CombSUM entries in the last column of table are the official TREC-2 results. Since we concentrated on the ad-hoc evaluations, these routing results are included primarily for the benefit of other groups, for purposes of comparison. The ad-hoc queries for topics 101 to 150 were evaluated in the same manner, and are reported in Table 10. Again, the official results are the two CombSUM entries in last column of the table.

As can be seen from Table 12, the CombSUM method performs quite poorly for certain topics while performing very well for others, compared to the best single run's results that that topic. Comparing the CombSUM results to the single best individual run (Pn2.0) shows an improvement for 46 out of the 50 topics, which shows that the CombSUM run performs much better than any single individual run. Performing a two-tailed paired t test on the Pn2.0 and CombSUM precisions results in a p value of about 1.1e-11, which indicates these results are very conclusive. However, comparing the CombSUM results with the best individual runs on a per query basis results in a p value of about 0.2, indicating that there is a 20 percent chance that the CombSUM method is no better than the best individual run for each specific query. Again, performing the same calculation on the R-Precision results in similar values.

4.1 The CEO Model

The **Combination of Expert Opinion (CEO)** model [6, 7] of Thompson can be used to treat the different retrieval methods as experts, and allows combining their weighting probability distributions to improve performance. This could be used in a variety of ways to combine results from a variety of runs and indexing schemes (that could include stemming and/or morphological analysis). For TREC-2, the CEO experiments completed consisted of combining seven individual runs, the three P-norm extended boolean retrieval run types described above, and retrieval runs based on the long vector queries, using both cosine correlation and inner product similarity measures for **SMART** system term weighting schemes of *nnn* and *atn*. Further discussion of this process and the results are described elsewhere in these proceedings.

4.2 Evaluation

Improvements in retrieval effectiveness from combining the evidence from multiple sources of evidence has been performed before in various incarnations, most recently by Belkin *et al.* [1] who evaluated the progressive effect of considering multiple soft boolean representations to improve on a base INQUERY natural language retrieval run. In their experiments, the base INQUERY natural language run performed better than any of the boolean representations, and they report that combining the results from the natural language representation and the combined boolean representations with equal weights performed worse than the best single run. Not until weighting the natural language run four times more

Table 7: Comparison of combination runs and the five individual runs (Ad-hoc Topics 51-100).

Run	Average Precision				R-Precision			
	AP-1	**WSJ-1**	**AP-2**	**WSJ-2**	**AP-1**	**WSJ-1**	**AP-2**	**WSJ-2**
SV	0.2387	0.2203	0.2543	0.1503	0.2624	0.2616	0.2649	0.1744
LV	0.2435	0.2414	0.2664	0.1633	0.2672	0.2800	0.2704	0.1860
Pn1.0	0.2810	0.2941	0.3004	0.2206	0.2688	0.3221	0.3165	0.2367
Pn1.5	**0.3122**	0.3199	**0.3332**	**0.2327**	**0.2976**	0.3443	**0.3412**	**0.2511**
Pn2.0	0.3027	**0.3217**	0.3300	0.2325	0.2968	**0.3470**	0.3339	0.2442
CombMAX	0.2856	0.3205	**0.3337**	**0.2343**	**0.3013**	**0.3484**	**0.3431**	0.2449
CombMIN	0.2863	0.1924	0.3047	0.1308	**0.3036**	0.2214	0.2980	0.1395
CombSUM	**0.3493**	**0.3605**	**0.3748**	**0.2752**	**0.3590**	**0.3767**	**0.3732**	**0.2851**
CombANZ	**0.3493**	**0.3367**	**0.3748**	**0.2465**	**0.3590**	**0.3517**	**0.3732**	**0.2590**
CombMNZ	0.3059	0.3368	**0.3516**	**0.2467**	**0.3175**	**0.3517**	**0.3578**	**0.2590**
CombMED	0.2943	0.3204	**0.3335**	0.2328	**0.2977**	0.3444	**0.3414**	**0.2518**

than the combined boolean schemes did they experience improved retrieval performance when combining different query methods. This differs from our results in several ways. Most importantly, the stage at which we combine the different methods differed: Belkin *et al.* combined the query representations before performing the actual retrieval, while we combined the similarity values produced from retrieval on each method individually. The difference between the two methodologies can best be demonstrated using the standard vector space model: Belkin *et al.* combined by summing the vector representations of each query, while our method is analogous to summing the cosines of the angles between each vector and a document. It is easily shown that the cosine of the angle between a document vector and a combined query vector, that is the sum of two query vectors as in the Belkin *et al.* approach, is not equal to the sum of the cosines between a document vector and the two separate query vectors. Other differences between the two methodologies include the fact that our P-norm queries performed better on average than our natural language vector queries, with exceptions on a per query basis. We used only one P-norm query and modified the operator weights while Belkin *et al.* used five different boolean queries. Finally, combining with five runs with equal weights actually improved performance over each individual run. However, one common trend emerges from both experiments: the more query representations considered, the better the results.

4.3 Future Exploration

Planned future work includes studying the following:

- Individually weighting various methods' similarity values when performing combination runs.
- Normalization methods to allow combination of runs made with different weighting schemes.
- Extending the analysis to all combinations of three and four retrieval runs.
- Considering more/different query types.

5 Acknowledgements

This research was supported in part by DARPA and by PRC Inc. We also thank Russell Modlin, M. Prabhakar Koushik and Durgesh Rao for their collaboration during TREC-1.

References

[1] Belkin, N.J., Cool, C., Croft, W.B., Callan, J.P. (1993, June). The Effect of Multiple Query Representations on Information Retrieval Performance. *Proc. 16th Int'l Conf. on R&D in IR (SIGIR '93)*, Pittsburgh, 339–346.

[2] Buckley, C. (1985, May) Implementation of the SMART information retrieval system. Technical Report 85-686, Cornell University, Department of Computer Science.

[3] Fox, E.A. (1983, August). Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. Cornell University Department of Computer Science dissertation.

[4] Fox, E.A., Koushik, M.P., Shaw, J., Modlin, R., Rao, D. (1993). Combining Evidence from Multiple Searches. In *The First Text REtrieval Conference*

Table 8: Average Precision and Exact R-Precision for CombSUM runs combining two or three individual runs (Ad-hoc Topics 51-100).

	Average Precision				R-Precision			
Run	**AP-1**	**WSJ-1**	**AP-2**	**WSJ-2**	**AP-1**	**WSJ-1**	**AP-2**	**WSJ-2**
SV	0.2387	0.2203	0.2543	0.1503	0.2624	0.2616	0.2649	0.1744
LV	0.2435	0.2414	0.2664	0.1633	0.2672	0.2800	0.2704	0.1860
Pn1.0	0.2810	0.2941	0.3004	0.2206	0.2688	0.3221	0.3165	0.2367
Pn1.5	**0.3122**	0.3199	**0.3332**	**0.2327**	**0.2976**	0.3443	**0.3412**	**0.2511**
Pn2.0	0.3027	**0.3217**	0.3300	0.2325	0.2968	**0.3470**	0.3339	0.2442
SV and LV	0.1457	0.1657	0.1611	0.1100	0.1492	0.1887	0.1524	0.1124
SV and Pn1.0	0.2774	0.3111	0.3257	**0.2362**	0.2874	0.3332	0.3360	**0.2554**
SV and Pn1.5	0.3117	**0.3389**	**0.3614**	**0.2463**	**0.3198**	**0.3575**	**0.3675**	**0.2637**
SV and Pn2.0	0.3012	**0.3395**	**0.3584**	**0.2467**	**0.3153**	**0.3580**	**0.3636**	**0.2551**
LV and Pn1.0	0.2744	0.3136	0.3197	**0.2353**	0.2867	0.3357	0.3269	**0.2518**
LV and Pn1.5	0.3057	**0.3413**	**0.3536**	**0.2442**	**0.3181**	**0.3596**	**0.3624**	**0.2568**
LV and Pn2.0	0.2950	**0.3408**	**0.3518**	**0.2458**	**0.3141**	**0.3608**	**0.3596**	**0.2615**
Pn1.0 and Pn1.5	0.2817	0.3109	0.3243	0.2324	0.2898	0.3412	0.3395	0.2476
Pn1.0 and Pn2.0	0.2935	0.3191	0.3330	**0.2367**	0.2944	0.3426	0.3397	0.2507
Pn1.5 and Pn2.0	0.2928	**0.3233**	**0.3336**	**0.2328**	0.2935	**0.3478**	0.3351	0.2489
Pn1.0, Pn1.5 and Pn2.0	0.2943	0.3192	**0.3345**	**0.2339**	0.2953	0.3421	0.3386	0.2497

(TREC-1), D.K. Harmon (Ed.), National Institute of Standards and Technology Special Publication 500-207, Gaithersburg, MD.

[5] Katzer, J., McGill, M.J., Tessier, J.A., Frakes, W., Dasgupta, P. (1982). A Study of the Overlap among Document Representations. *Information Technology: Reseach and Development*, 1(2):261-274.

[6] Thompson, P. (1990) A Combination of Expert Opinion approach to probabilistic information retrieval, Part 1: The conceptual model. *Information Processing & Management*, 26(3):371–382, 1990.

[7] Thompson, P. (1990). A Combination of Expert Opinion approach to probabilistic information retrieval, Part 2: Mathematical treatment of CEO Model 3. *Information Processing & Management*, 26(3):383–394.

Table 9: Average Precision and Exact R-Precision for the five individual runs compared with the combined CombSUM runs (Routing Topics 51-100).

Average non-interpolated Precision

Run	**AP-3**	**PATN-3**	**SJM-3**	**ZF-3**	**Disk 3**
SV	0.1347	0.0189	0.1139	0.0593	0.0589
LV	0.1189	0.0156	0.1056	0.0587	0.0494
Pn1.0	0.2519	**0.0257**	0.2128	0.1141	0.2039
Pn1.5	**0.2869**	0.0239	**0.2411**	0.1189	**0.2279**
Pn2.0	0.2852	0.0221	0.2390	**0.1303**	0.2225
CombSUM	0.3196	0.0260	0.2696	0.1304	0.2681
Chg/Max	11.4%	1.2%	11.8%	0.07%	17.6%

Exact R-Precision

Run	**AP-3**	**PATN-3**	**SJM-3**	**ZF-3**	**Disk 3**
SV	0.1703	0.0171	0.1337	0.0595	0.0595
LV	0.1444	0.0156	0.1098	0.0547	0.1002
Pn1.0	0.2790	**0.0325**	0.2185	0.1224	0.2594
Pn1.5	**0.3082**	0.0322	**0.2579**	0.1248	0.2786
Pn2.0	0.3062	0.0310	0.2531	**0.1462**	**0.2809**
CombSUM	0.3319	0.0319	0.2900	0.1260	0.3143
Chg/Max	7.7%	-1.8%	12.4%	-13.8%	11.9%

Table 10: Average Precision and Exact R-Precision for the five individual runs (Ad-hoc Topics 101-150).

Average non-interpolated Precision

	Disk 1					**Disk 2**				**Both**
Run	**AP**	**DOE**	**FR**	**WSJ**	**ZF**	**AP**	**FR**	**WSJ**	**ZF**	**Disks**
SV	0.3237	0.0949	0.0630	0.2740	0.0936	0.3068	0.0650	0.2259	0.1166	0.2035
LV	0.3326	0.0697	0.1018	0.2848	0.0997	0.2981	0.0602	0.2483	0.1045	0.2159
Pn1.0	0.3340	**0.0831**	0.1777	0.3153	0.1292	0.3133	0.1927	0.2838	0.1722	0.2205
Pn1.5	**0.3682**	0.0814	**0.1874**	**0.3332**	**0.1430**	**0.3438**	0.1982	**0.2941**	0.1964	0.2543
Pn2.0	0.3647	0.0750	0.1761	0.3290	0.1307	0.3419	**0.1995**	0.2828	**0.2018**	**0.2573**
CombSUM	0.4153	0.1038	0.2133	0.3778	0.1657	0.3959	0.2000	0.3561	0.2200	0.3206
Chg/Max	12.8%	9.4%	13.8%	13.4%	15.9%	15.1%	0.2%	21.0%	9%	24.6%

Exact R-Precision

	Disk 1					**Disk 2**				**Both**
Run	**AP**	**DOE**	**FR**	**WSJ**	**ZF**	**AP**	**FR**	**WSJ**	**ZF**	**Disks**
SV	0.3351	0.0947	0.0567	0.2881	0.1086	0.3052	0.0718	0.2502	0.1130	0.2649
LV	0.3385	0.0714	0.1006	0.3022	0.1055	0.3019	0.0540	0.2712	0.1036	0.2661
Pn1.0	0.3495	**0.0989**	0.1434	0.3322	0.1190	0.3139	0.1653	0.2934	0.1577	0.2867
Pn1.5	0.3721	0.0899	**0.1582**	**0.3440**	**0.1277**	**0.3465**	**0.1716**	**0.3044**	0.1796	0.3203
Pn2.0	**0.3735**	0.0925	0.1579	0.3389	0.1154	0.3401	0.1640	0.2953	**0.1979**	**0.3233**
CombSUM	0.4137	0.1156	0.1938	0.3757	0.1389	0.3712	0.1741	0.3385	0.1909	0.3711
Chg/Max	10.8%	16.9%	22.5%	9.2%	8.8%	7.1%	1.4%	11.2%	-3.5%	14.8%

Table 11: Per-Topic comparison of Average Precisions (Ad-hoc Topics 51-100).

Topic	#Rel	SV	LV	Pn1.0	Pn1.5	Pn2.0	CombSUM	Chg/Max	Chg/Pn2.0
74	273	0.0000	0.0099	0.0001	0.0002	0.0002	0.0002	-97.98%	0.00%
90	206	0.1248	0.1134	0.0082	0.0036	0.0015	0.0089	-92.87%	493.33%
77	138	0.1100	0.2241	0.0093	0.0196	0.0246	0.0411	-81.66%	67.07%
91	40	0.1648	0.1039	0.0012	0.0038	0.0068	0.0308	-81.31%	352.94%
72	119	0.1099	0.1173	0.0076	0.0116	0.0138	0.0323	-72.46%	134.06%
73	183	0.0231	0.0124	0.0017	0.0052	0.0120	0.0098	-57.58%	-18.33%
94	310	0.0176	0.0185	0.0191	0.1094	0.2100	0.1043	-50.33%	-50.33%
85	894	0.1587	0.1779	0.0239	0.0433	0.0518	0.1018	-42.78%	96.53%
67	92	0.0011	0.0012	0.0435	0.0487	0.0502	0.0290	-42.23%	-42.23%
55	810	0.2362	0.4535	0.1153	0.1266	0.1256	0.2874	-36.63%	128.82%
64	375	0.1493	0.1641	0.0740	0.0673	0.0706	0.1097	-33.15%	55.38%
57	461	0.0282	0.0605	0.5464	0.4518	0.3447	0.3786	-30.71%	9.83%
80	374	0.0218	0.0266	0.1007	0.1125	0.1025	0.0899	-20.09%	-12.29%
82	602	0.2544	0.3139	0.2603	0.1746	0.1278	0.2521	-19.69%	97.26%
97	352	0.0513	0.0513	0.1133	0.1376	0.1405	0.1190	-15.30%	-15.30%
76	294	0.1481	0.1677	0.1221	0.0966	0.0755	0.1440	-14.13%	90.73%
89	174	0.0365	0.0762	0.0481	0.1159	0.1515	0.1306	-13.80%	-13.80%
98	666	0.0822	0.0677	0.0421	0.0579	0.0703	0.0711	-13.50%	1.14%
84	396	0.1659	0.0985	0.0787	0.1251	0.1559	0.1471	-11.33%	-5.64%
99	288	0.4834	0.4796	0.3529	0.3551	0.3224	0.4480	-7.32%	38.96%
58	159	0.0971	0.0620	0.3459	0.4198	0.4072	0.3937	-6.22%	-3.32%
92	88	0.0253	0.0405	0.0130	0.0154	0.0176	0.0381	-5.93%	116.48%
71	380	0.0595	0.0613	0.1396	0.1978	0.2210	0.2156	-2.44%	-2.44%
88	165	0.1813	0.2341	0.3629	0.3635	0.3542	0.3550	-2.34%	0.23%
59	579	0.2368	0.2272	0.1546	0.3827	0.4370	0.4282	-2.01%	-2.01%
93	171	0.0344	0.0385	0.5233	0.4831	0.3779	0.5141	-1.76%	36.04%
95	263	0.0149	0.0228	0.1450	0.2072	0.2392	0.2375	-0.71%	-0.71%
52	535	0.4918	0.4462	0.4896	0.6623	0.6882	0.6864	-0.26%	-0.26%
78	162	0.2202	0.2632	0.7696	0.7484	0.7250	0.7705	0.12%	6.28%
61	206	0.4106	0.5130	0.3646	0.3786	0.3621	0.5167	0.72%	42.70%
51	138	0.2799	0.3957	0.4751	0.5282	0.5428	0.5476	0.88%	0.88%
70	55	0.1156	0.1198	0.7700	0.7919	0.7982	0.8080	1.23%	1.23%
54	171	0.3063	0.3045	0.3950	0.3207	0.2673	0.4048	2.48%	51.44%
62	298	0.1674	0.1907	0.1519	0.2087	0.2121	0.2198	3.63%	3.63%
81	62	0.1759	0.1410	0.2287	0.2370	0.2321	0.2467	4.09%	6.29%
68	195	0.0960	0.0920	0.1040	0.2098	0.2540	0.2651	4.37%	4.37%
69	52	0.0956	0.1629	0.5382	0.5873	0.5833	0.6227	6.03%	6.75%
53	571	0.1821	0.1874	0.1543	0.2928	0.3241	0.3461	6.79%	6.79%
83	633	0.2673	0.2931	0.1753	0.2412	0.2666	0.3317	13.17%	24.42%
56	878	0.3011	0.3277	0.3391	0.3089	0.2691	0.3955	16.63%	46.97%
100	317	0.1904	0.1972	0.1423	0.1815	0.2094	0.2516	20.15%	20.15%
65	386	0.0100	0.0078	0.1111	0.1190	0.1236	0.1529	23.71%	23.71%
86	213	0.5146	0.4242	0.5216	0.5234	0.4891	0.6624	26.56%	35.43%
60	60	0.0547	0.0887	0.0866	0.0960	0.0992	0.1259	26.92%	26.92%
79	232	0.1376	0.1220	0.2057	0.2784	0.2719	0.3690	32.54%	35.71%
75	365	0.0016	0.0037	0.0443	0.0499	0.0481	0.0684	37.07%	42.20%
63	208	0.0032	0.0028	0.0631	0.0597	0.0688	0.0972	41.28%	41.28%
66	197	0.0001	0.0000	0.0320	0.0388	0.0478	0.0695	45.40%	45.40%
87	188	0.0436	0.0559	0.0410	0.0914	0.1256	0.1867	48.65%	48.65%
96	693	0.0095	0.0084	0.0833	0.1177	0.1302	0.2356	80.95%	80.95%
Avg	15667	0.1418	0.1555	0.1988	0.2242	0.225	0.262	16.44%	16.44%

Table 12: Per-Topic comparison of Average Precisions (Ad-hoc Topics 101-150).

Topic	#Rel	SV	LV	Pn1.0	Pn1.5	Pn2.0	CombSUM	Chg/Max	Chg/Pn2.0
103	94	0.2400	0.2379	0.0130	0.0225	0.0309	0.0532	-77.83%	72.17%
122	114	0.1467	0.1059	0.0547	0.0490	0.0544	0.0898	-38.79%	65.07%
101	57	0.2232	0.1932	0.0900	0.1088	0.1175	0.1482	-33.60%	26.13%
140	25	0.0150	0.0520	0.0111	0.0207	0.0315	0.0356	-31.54%	13.02%
135	400	0.4072	0.4449	0.1103	0.1121	0.0835	0.3122	-29.83%	273.89%
104	75	0.1867	0.2214	0.0715	0.0642	0.0828	0.1671	-24.53%	101.81%
121	55	0.0017	0.0028	0.0283	0.0542	0.0628	0.0475	-24.36%	-24.36%
127	223	0.2411	0.2476	0.0543	0.0942	0.1028	0.1911	-22.82%	85.89%
143	397	0.4069	0.3978	0.1628	0.1968	0.2022	0.3257	-19.96%	61.08%
124	173	0.0201	0.0165	0.1711	0.1360	0.1068	0.1501	-12.27%	40.54%
146	358	0.6795	0.7017	0.4394	0.4775	0.4907	0.6255	-10.86%	27.47%
114	138	0.0518	0.0756	0.2381	0.2559	0.2615	0.2525	-3.44%	-3.44%
112	291	0.0041	0.0188	0.3876	0.3515	0.3194	0.3779	-2.50%	18.32%
130	286	0.2301	0.2937	0.4184	0.5477	0.5749	0.5632	-2.04%	-2.04%
102	64	0.2127	0.2308	0.1119	0.1373	0.1505	0.2287	-0.91%	51.96%
128	381	0.0915	0.0751	0.2583	0.3710	0.4190	0.4156	-0.81%	-0.81%
150	458	0.4618	0.4994	0.3362	0.4432	0.4566	0.4985	-0.18%	9.18%
115	165	0.4053	0.4383	0.3262	0.3633	0.3720	0.4407	0.55%	18.47%
113	206	0.0531	0.0768	0.2700	0.3304	0.3030	0.3367	1.91%	11.12%
119	326	0.0644	0.0843	0.2442	0.2252	0.2018	0.2497	2.25%	23.74%
107	98	0.1101	0.1580	0.3156	0.4062	0.4451	0.4592	3.17%	3.17%
116	49	0.2635	0.1886	0.3008	0.2803	0.2487	0.3125	3.89%	25.65%
118	273	0.0259	0.0164	0.1551	0.1736	0.1725	0.1806	4.03%	4.70%
145	162	0.3218	0.2983	0.2169	0.2180	0.1961	0.3359	4.38%	71.29%
138	52	0.0895	0.0992	0.1169	0.1321	0.1322	0.1386	4.84%	4.84%
148	250	0.7256	0.7300	0.7316	0.7904	0.8010	0.8416	5.07%	5.07%
108	294	0.0994	0.1266	0.3089	0.2586	0.1820	0.3264	5.67%	79.34%
136	206	0.1713	0.1976	0.3943	0.5899	0.5894	0.6319	7.12%	7.21%
134	188	0.4033	0.4132	0.5091	0.4897	0.4675	0.5500	8.03%	17.65%
132	201	0.1979	0.3246	0.5671	0.5759	0.5625	0.6222	8.04%	10.61%
133	80	0.2396	0.2793	0.1541	0.2322	0.2425	0.3034	8.63%	25.11%
147	315	0.1371	0.1266	0.2312	0.2820	0.3001	0.3317	10.53%	10.53%
106	201	0.0354	0.0458	0.2784	0.3463	0.3653	0.4039	10.57%	10.57%
126	240	0.2766	0.2990	0.1997	0.3132	0.3548	0.3971	11.92%	11.92%
125	169	0.1632	0.1719	0.2580	0.2224	0.1842	0.2920	13.18%	58.52%
137	158	0.3365	0.4135	0.2339	0.3128	0.3555	0.4715	14.03%	32.63%
131	28	0.0659	0.0648	0.0671	0.0984	0.1061	0.1255	18.28%	18.28%
110	496	0.4909	0.4968	0.4845	0.4768	0.4336	0.6001	20.79%	38.40%
142	660	0.4269	0.4479	0.4060	0.4694	0.4629	0.5721	21.88%	23.59%
117	275	0.1165	0.1030	0.1296	0.1596	0.1637	0.2007	22.60%	22.60%
149	133	0.0309	0.0836	0.0490	0.0732	0.0769	0.1028	22.97%	33.68%
129	207	0.3391	0.2338	0.2402	0.3348	0.3477	0.4369	25.65%	25.65%
144	49	0.2490	0.2275	0.0832	0.1327	0.1958	0.3130	25.70%	59.86%
105	54	0.0656	0.1744	0.1304	0.1439	0.1441	0.2262	29.70%	56.97%
139	55	0.0911	0.1139	0.0592	0.0705	0.0776	0.1481	30.03%	90.85%
111	285	0.3795	0.3540	0.1991	0.3025	0.3756	0.5025	32.41%	33.79%
123	435	0.0697	0.0800	0.2254	0.2252	0.2043	0.3014	33.72%	47.53%
120	83	0.0165	0.0136	0.0490	0.0556	0.0538	0.0746	34.17%	38.66%
109	742	0.0875	0.0879	0.0811	0.1344	0.1500	0.2290	52.67%	52.67%
141	36	0.0084	0.0125	0.0538	0.0517	0.0514	0.0873	62.27%	69.84%
Avg	10760	0.2035	0.2159	0.2205	0.2543	0.2573	0.3206	24.60%	24.60%