Combination of Multiple Searches

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

Abstract

The TREC-3 project at Virginia Tech focused on methods for combining the evidence from multiple retrieval runs and queries to improve retrieval performance over any single retrieval method or query. The largest improvements result from the combination of retrieval paradigms rather than from the use of multiple similar queries.

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. In performing our TREC-3 ad-hoc retrieval experiments on the provided test collections, the results from both vector and P-norm [3] queries were considered in estimating the similarity 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. Our TREC-3 experiments built upon our TREC-2 experiments and focused more on determining where the improvements in combination were derived from rather than on evaluating different combination methods.

2 Index Creation

This section outlines the indexing done with the document collections 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 5000/125** 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 simply removed.

The resulting filtered text was tokenized, stop words were deleted using the standard 418 word stop list provided with **SMART**, plural removal stemming was performed, 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 **SMART** *ann* weighting scheme, defined as $term_weight = 0.5 + 0.5 * \frac{tf}{\max_tf}$ proved to be the most effective in our TREC-2 experiments [5] and was used to evaluate all the queries in our TREC-3 results. The dictionary size for each collection was approximately 16 MB, while the document vector files ranged from 40 MB to 120MB (see Table 1).

3 Retrieval

3.1 Queries

All of the queries were created by the researcher 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), while two different sets of vector queries were created from the topics. The Title, Description and Narrative sections

Table 1: Collection statistics summary. Text, Dictionary and Document Vector sizes in Megabytes.

Collection	Text	Dict.	Doc. Vectors	Total Doc.s
AP-1	266	15.8	116.6	84678
DOE-1	190	15.7	95.3	226087
FR-1	258	15.7	50.9	26207
WSJ-1	295	16.0	120.4	98735
ZIFF-1	251	15.5	83.9	75180
D1	1260	N/A	467.1	510887
AP-2	248	15.7	107.1	79923
FR-2	211	15.5	40.1	20108
WSJ-2	255	15.9	101.7	74520
ZIFF-2	188	15.2	60.5	56920
D2	902	N/A	309.4	231471
Total	2162	N/A	776.5	742358

of the topics were used in the creation of all three sets of queries, while the P-norm query set and one of the vector query sets also contained a limited amount of additional terms added from the general knowledge of the query author to compensate for obvious omissions in the topic descriptions. The vector query set that included the addtional terms 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.

The five queries used for TREC-3 are similar in structure to our TREC-2 ad hoc queries, with the exception that one our TREC-3 vector queries contained terms that were not present in the topic descriptions, while the longer of our two TREC-3 vector queries did.

3.2 Individual Retrieval Runs

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 to simplify the merging of retrieval results across the various collections. The resulting similarity values were not based on collection statistics which would have differed for each collection. 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 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 2, and are equivalent to our TREC-2 runs with the exceptions in query construction noted above.

Table 2: 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 = 2.0

3.3 Combination Retrieval Runs

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. We explored several methods for combining the individual similarity values and found that simply combining the similarty values in a linear fashion– summing the similarity values–worked better than trying to select a given similarity value. This method of combination, called Comb-SUM in our TREC-2 report, [5] was used exclusively in our TREC-3 experiments.

4 TREC-3 Results

The procedure described above was used for our official TREC-3 ad-hoc results. We submitted two sets of results: one run labeled VTc5s which used the Comb-SUM method to combine all five individual runs, as per our official TREC-2 results, and one run labeled VTc2s which combined only the short vector query with the Pnorm query evaluated using a p-value of 1.5. The official results are reported in the last column of Table 3, in the rows labeled for the two runs.

Note that for all the collections the long vector query set containing terms not included in the topic performed better than the short vector query set. On a per-query

Table 3: Average Precision and Exact R-Precision for the five individual runs (Ad-hoc Topics 151-200).

Average non-interpolated Precision

Run	Disk 1					Disk 2				Both
	AP	**DOE**	**FR**	**WSJ**	**ZF**	**AP**	**FR**	**WSJ**	**ZF**	**Disks**
SV	0.2611	0.0320	0.0397	0.1957	0.0189	0.2355	0.0290	0.1811	0.0461	0.1340
LV	0.3139	0.0536	0.0547	0.2544	0.0459	0.2815	0.0357	0.2313	0.0588	0.1960
Pn1.0	0.3276	**0.0852**	0.0956	0.3240	0.0582	0.3038	0.0883	0.2840	**0.1019**	0.2062
Pn1.5	**0.3396**	0.0812	**0.1048**	**0.3435**	0.0599	**0.3201**	**0.0922**	**0.2915**	0.0974	0.2245
Pn2.0	0.3223	0.0758	0.1028	0.3283	**0.0651**	0.3120	0.0911	0.2894	0.0970	**0.2270**
VTc5s	0.3822	0.0855	0.1244	0.3866	0.0734	0.3604	0.1013	0.3322	0.1133	0.2914
Chg/Max	12.5%	0%	18.7%	12.5%	12.7%	12.6%	9.8%	14.0%	11.2%	28.4%
VTc2s	0.3944	0.0853	0.1125	0.3915	0.0732	0.3642	0.1005	0.3411	0.1192	0.3021
Chg/Max	16.1%	5.0%	7.3%	14.0%	22.2%	13.8%	9.0%	17.0%	22.4%	34.6%

Exact R-Precision

Run	Disk 1					Disk 2				Both
	AP	**DOE**	**FR**	**WSJ**	**ZF**	**AP**	**FR**	**WSJ**	**ZF**	**Disks**
SV	0.2996	0.0292	0.0406	0.2263	0.0225	0.2649	0.0217	0.2199	0.0336	0.2058
LV	0.3440	0.0562	0.0504	0.2833	0.0404	0.2907	0.0287	0.2427	0.0415	0.2607
Pn1.0	0.3500	0.0752	0.0903	0.3428	0.0575	0.3087	0.0751	0.2961	0.0799	0.2748
Pn1.5	**0.3528**	**0.0831**	0.0944	**0.3591**	**0.0623**	**0.3261**	0.0753	0.3043	0.0867	0.2855
Pn2.0	0.3453	0.0740	**0.0968**	0.3451	0.0554	0.3236	**0.0760**	**0.3082**	**0.0900**	**0.2895**
VTc5s	0.3947	0.0853	0.1181	0.3840	0.0633	0.3650	0.0824	0.3470	0.0938	0.3404
Chg/Max	11.9%	2.6%	22.0%	6.9%	1.6%	11.9%	8.4%	12.6%	4.2%	17.6%
VTc2s	0.4082	0.0875	0.1090	0.3938	0.0651	0.3728	0.0712	0.3487	0.0980	0.3538
Chg/Max	15.7%	5.3%	15.5%	9.7%	4.5%	14.3%	-5.4%	14.6%	13.0%	23.9%

basis, this held true for 39 of the 50 queries. Furthermore, the pnorm queries built from the long vector queries performed better on average than both sets of vector queries, though the improvement over the long vector query set was slight.

The VTc5s run shows a significant overall improvement over the five individual runs, and on a per query basis performed better than the best of the individual runs for 36 of the 50 topics. This matches the results obtained in our TREC-2 experiments. However, unlike our TREC-2 exerpiments, the VTc2s run combining only two of the five runs also performed significantly better than the best of the individual runs, and in fact performed better than the combination of all five runs overall and for several of the collections. On a per query basis, the VTc2s run performed better than the best of its two component runs for 42 of the 50 topics. However, the difference in overall performance between the two combination runs is not significant.

Further experiments involving all the possible combinations of two individual runs, as reported in Table 4 reveals further interesting trends. Combining two of the same type of runs, either both vector queries or two of the pnorm queries shows little improvement over the individual runs, and performs worse than the best of the two runs in many instances. However, combining one of the two vector queries with one of the pnorm queries always shows an improvement. This indicates that the primary source of improvements seen in the combination runs submitted for TREC-3 derives from the combination of retrieval paradigms and not simply from the use of multiple queries. This may be due to the similarity inherent in the five queries; combining two queries composed of two widely different sets of query terms may well result in significant improvements. But given a single set of query terms, it is still possible to achieve significant improvements by combining different retrieval paradigms.

5 Acknowledgements

This research was supported by the Virginia Tech Department of Computer Science and Computing Center. We also thank Russell Modlin, M. Prabhakar Koushik and Durgesh Rao for their collaboration during TREC-1.

Table 4: Average Precision for CombSUM runs combining two or three individual runs compared with combining all five individual runs. (Ad-hoc Topics 151-200).

Average non-interpolated Precision

	Disk 1					Disk 2				Both
Run	**AP**	**DOE**	**FR**	**WSJ**	**ZF**	**AP**	**FR**	**WSJ**	**ZF**	**Disks**
SV	0.2611	0.0320	0.0397	0.1957	0.0189	0.2355	0.0290	0.1811	0.0461	0.1340
LV	0.3139	0.0536	0.0547	0.2544	0.0459	0.2815	0.0357	0.2313	0.0588	0.1960
Pn1.0	0.3276	**0.0852**	0.0956	0.3240	0.0582	0.3038	0.0883	0.2840	**0.1019**	0.2062
Pn1.5	**0.3396**	0.0812	**0.1048**	**0.3435**	0.0599	0.3201	**0.0922**	**0.2915**	0.0974	0.2245
Pn2.0	0.3223	0.0758	0.1028	0.3283	**0.0651**	**0.3120**	0.0911	0.2894	0.0970	**0.2270**
SV-LV	0.3170	0.0473	0.0540	0.2568	0.0389	0.2849	0.0351	0.2310	0.0571	0.1865
SV-Pn1.0	0.3849	0.0894	0.1136	0.3734	0.0720	0.3479	**0.1079**	0.3264	0.1179	0.2826
VTc2s	0.3944	0.0853	0.1125	0.3915	0.0732	0.3642	0.1005	0.3411	0.1192	0.3021
SV-Pn2.0	0.3845	0.0831	0.1158	0.3871	0.0736	0.3608	0.0929	0.3351	0.1192	0.3004
LV-Pn1.0	0.3795	**0.0903**	0.1161	0.3766	0.0723	0.3516	0.0923	0.3296	0.1177	0.2941
LV-Pn1.5	0.3885	0.0844	0.1181	0.3966	**0.0775**	0.3654	0.0989	0.3429	0.1188	0.3104
LV-Pn2.0	0.3816	0.0823	0.1194	0.3890	0.0767	0.3634	0.0940	0.3395	0.1188	0.3100
Pn1.0-Pn1.5	0.3393	0.0846	0.0998	0.3405	0.0595	0.3191	0.0929	0.2940	0.0989	0.2183
Pn1.0-Pn2.0	0.3415	0.0823	0.1032	0.3449	0.0578	0.3216	0.0940	0.2909	0.0980	0.2236
Pn1.5-Pn2.0	0.3331	0.0797	0.1025	0.3353	0.0605	0.3194	0.0906	0.2916	0.0969	0.2267
SV-LV-Pn1.0	0.3953	0.0883	0.1155	0.3782	0.0751	0.3600	0.0961	0.3379	**0.1207**	0.3083
SV-LV-Pn1.5	**0.4056**	0.0841	**0.1253**	0.3947	0.0767	0.3704	0.0994	0.3488	0.1079	0.3204
SV-LV-Pn2.0	0.4029	0.0819	0.1241	**0.3987**	0.0706	**0.3722**	0.0969	**0.3490**	0.1090	**0.3219**
VTc5s	0.3822	0.0855	0.1244	0.3866	0.0734	0.3604	0.1013	0.3322	0.1133	0.2914

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 (TREC-1)*, D.K. Harmon (Ed.), National Institute of Standards and Technology Special Publication 500-207, Gaithersburg, MD, 319–328.

[5] Fox, E.A., Shaw, J.A. (1994). Combination of Multiple Searches. In *The Second Text REtrieval Conference (TREC-2)*, D.K. Harmon (Ed.), National Institute of Standards and Technology Special Publication 500-215, Gaithersburg, MD, 243-252.

[6] 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.