System Fusion for Improving Performance in Information Retrieval Systems

M. C. McCabe, A. Chowdhury, D. Grossman, O. Frieder
Illinois Institute of Technology
{cmccabe, chowdhury, grossman, frieder}@ir.iit.edu

Abstract

Fusion of various retrieval strategies has long been suggested as a means of improving retrieval effectiveness. To date, testing of fusion was done by combining result sets from widely disparate approaches that include an uncontrolled mixture of retrieval strategies and utilities. To isolate the effect of fusion on individual retrieval models, we have implemented probabilistic, vector space, and weighted Boolean models and tested the effect of fusion on these strategies in a systematic fashion. We also tested the effect of fusion on various query representations and have shown up to a twelve percent improvement in average precision.

1 Introduction

Information Retrieval search strategies use different models and different similarity measures to retrieve relevant documents. It has been shown that disparate systems implementing different retrieval strategies bring back different result sets [Fox94] although the overlap of relevant documents in the top 100 documents is high [Harman99]. Fox showed that combining results from these disparate systems generates a better answer set [Fox94]. Our work last year, isolating the leading similarity measures and fusing the results, showed that the variation of similarity measure alone does not result in improvement. The result sets overlap greatly and are not different enough to warrant combination. The conclusion from that work is that implementation variations such as parsing rules, stop lists, stemmers, phrase rules, etc. have a profound impact on results [McCabe99]. Consider a parser whose stop word lists contains the word *after* and another parser that does not. A query such as "Find reviews of the song *The Morning After*" will result in a very different set of retrieved documents for each system. This paper extends the work of fusing isolated similarity measures. We now examine fusion in the same common environment with varied query representation in addition to similarity measures. Our results show that fusion of different similarity measures in combination with utilities such as relevance feedback is effective. The more varied the query representation, the better the improvements through fusion.

This paper is organized as follows: Section 2.0 describes prior work exploring fusion of retrieval systems. Section 3.0 explains the algorithm we use for fusion. Section 4.0 presents experimental design and results. Finally, conclusions are discussed in section 5.0.

2 Prior Work

Initial work on fusion was done by Fox, Shaw and Thompson as early as TREC-1 [Fox94, Thompson90] Thompson considered each result set an 'expert' and used an existing approach for combining experts. The idea is to weight better experts more than others based on their prior performance. Thompson's merged results were not significantly different from simply using the best of the sources. Fox proposed several combination algorithms:

COMBSUM = sum of the individual measures
COMBMIN = minimum of the individual measures
COMBMAX = maximum of the individual measures
COMBAVG = average of the individual measures
COMBMNZ = COMBSUM* num of runs w/document

He found that combinations of the same types of runs (i.e. long and short queries within vector space) did not achieve improvements over individual runs. However, improvement did occur when merging different models: vector space and p-norm Boolean [Shaw95].

2.1 Characteristics for Fusion

Several research groups have attempted to characterize what makes result sets good candidates for effective fusion [Lee97, Vogt98, Bartell94]. Vogt's work determined the five best characteristics to be:

0-7695-1062-0/01 $10.00 © 2001 IEEE

1) at least one result has high precision/recall,
2) high overlap of relevant documents,
3) low overlap of nonrelevant documents,
4) both distribute scores similarly, and
5) each rank relevant documents differently.

We use these characteristics for examining our result sets for fusion. Each measure used is further explained below.

2.1.1 Overlap of Result Sets

A measure of overlap between results sets is an important indicator of whether fusion will be effective for given sets [Lee97]. Specifically, relevant and nonrelevant overlaps as shown in Equations 1 and 2.

$$R_{overlap} = \frac{R_{common} \times 2}{R_{VSM} + R_{PROB}} \qquad N_{overlap} = \frac{N_{common} \times 2}{N_{VSM} + N_{PROB}} \qquad (1,2)$$

2.1.2 Correlation of Rankings

For the comparison of rankings, we use the Spearman measure of correlation of ranked sets. The Spearman rank correlation coefficient, $\hat{\rho}$, is a measure of similarity between two rankings where x is the rank assigned to the document by the first system and y is the rank assigned to the same doc by the second system. Thus, the Spearman coefficient is stated:

$$\hat{\rho}_s = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \qquad (3)$$

3 Algorithm for Fusion Combinations

Work presented in [Fox94] showed that the COMBMNZ algorithm for fusion was the most effective. One variation of this algorithm – normalization -- is used when retrieval approaches have different ranges of similarity measures. This approach takes the sum of the normalized similarity of each input run and multiplies it by the number of input runs containing the document. Documents found in both sets clearly get promoted. So a high overlap among relevant and low among nonrelevant is the key to this approach.

In order to emphasize one input run over another, a scalar multiplier may be used. Several groups have investigated ways to optimize scalar assignment, but none have found that scalars make a big difference in fusion results [Bartell94, Mounir98]. We experiment with linear scalars after the runs have been normalized to determine if scalars have a consistent impact within a common environment for fusion.

4 Fusion Results

We isolated three search features in our experimentation – the similarity measures, the query expansion utilities such as relevance feedback, and the query representation such as title or description. We examine the factors for effective fusion discussed above. The similarity measures selected for fusion include the vector space model (VSM) using Singhal's pivoted normalization, the probabilistic model (PROB) using Robertson's formula and finally a weighted Boolean approach (BOOL). We use the entire 2GB TREC6-8 collection for our experimentation, along with the 150 queries from TREC-6, 7 and 8. The sections below discuss experimentation with each search feature.

4.1 Combining Baseline Similarity Measures

We combined the baseline runs that consist of the retrieval strategy with no utilities. We held the query representation constant and vary the retrieval strategy. For instance, combining the vector space retrieval using the Title query with the probabilistic retrieval using the Title query. The TREC queries have three components – Title, Description and Narrative. We restricted our experimentation to title and description because long, natural language queries such as narrative are seldom the norm in today's search systems. These baseline runs use stemmed terms and two-term phrases from the title-only and the title plus description versions of the TREC topics. We used no query expansion for our baseline runs. The results for these runs are shown in Table 1 and Table 2. The table entries are ordered by the overall effectiveness of the fusion.

The overlap values and the ranking similarity values for the baseline runs are very high. Thus, the impact of fusion is minimal. This result is somewhat surprising given the general belief that probabilistic, Boolean and vector space are very different. However, it is consistent with the finding that there are very few unique relevant documents among TREC participants – that is relevant documents brought back by only one team [Voorhees96]. When the approaches are pared down to the basic retrieval model, and *idf* is a component of the term weighting, the result sets are highly similar, with high overlap in nonrelevant and relevant documents and very similar rankings.

4.1.1 Combinations with Expansiom Utilities

We then added the query expansion utilities commonly used for each similarity measure and combined those results. The Vector Space Method, uses Rocchio relevance feedback. Probabilistic uses probabilistic feedback. For weighted Boolean, we used a simple local context feedback technique consisting of N**nidf* where N is the number of top documents containing the candidate term. Our relevance feedback run added the best terms from the top 20 documents. For all retrieval runs, we used identical original query terms all retrievals.

Table 1 Fusion of Different Retrieval Strategies - Title Queries

INPUT					FUSION				
Test Set	Method-1 TITLE Avg Precision		Method-2 TITLE Avg Precision		Avg Precision	N-Overlap	R-Overlap	Spearman	Change
Trec6	VSM	0.2003	PROB	0.2107	0.2047	86%	95%	0.87	-2.80%
Trec8	VSM	0.2301	PROB	0.2431	0.2367	84%	96%	0.85	-2.60%
Trec6	VSM	0.2003	BOOL	0.2123	0.2069	85%	95%	0.86	-2.50%
Trec8	VSM	0.2301	BOOL	0.2438	0.2395	84%	95%	0.85	-1.80%
Trec7	VSM	0.1476	PROB	0.1529	0.1503	86%	93%	0.87	-1.70%
Trec7	VSM	0.1476	BOOL	0.1529	0.1512	86%	96%	0.87	-1.10%
Trec7	PROB	0.1529	BOOL	0.1529	0.1535	82%	93%	0.83	0.40%
Trec6	PROB	0.2107	BOOL	0.2123	0.2146	84%	95%	0.85	1.10%
Trec8	PROB	0.2431	BOOL	0.2438	0.2468	82%	95%	0.83	1.20%

Table 2: Fusion of Different Retrieval Strategies - Title and Description Queries

INPUT					FUSION				
Test Set	Method-1 TITLE+DESC Avg Precision		Method-2 TITLE+DESC Avg Precision		Avg Precision	N-Overlap	R-Overlap	Spearman	Change
Trec8	VSM	0.2520	BOOL	0.2673	0.2634	79%	95%	0.79	-1.50%
Trec7	VSM	0.1834	PROB	0.1958	0.1938	86%	95%	0.87	-1.00%
Trec8	VSM	0.2520	PROB	0.2657	0.2634	85%	96%	86%	-0.90%
Trec7	VSM	0.1834	BOOL	0.1927	0.1912	80%	92%	0.81	-0.80%
Trec6	VSM	0.2267	BOOL	0.2363	0.2346	79%	93%	0.78	-0.70%
Trec6	VSM	0.2267	PROB	0.2354	0.2344	86%	95%	0.87	-0.40%
Trec7	PROB	0.1958	BOOL	0.1927	0.1963	79%	92%	0.79	0.30%
Trec6	PROB	0.2354	BOOL	0.2363	0.2377	76%	91%	0.75	0.60%
Trec8	PROB	0.2657	BOOL	0.2673	0.2703	78%	94%	0.78	1.10%

Table 3: Fusion of Similarity Measures with Query Expansion

INPUT					FUSION				
Test Set	Method-1 Avg Precision		Method-2 Avg Precision		Avg Precision	N-Overlap	R-Overlap	Spearman	Change
Trec8	VSM	0.2700	PROB	0.2880	0.2822	75%	95%	0.75	**-2.0%**
Trec6	VSM	0.2360	BOOL	0.2450	0.2441	74%	90%	0.73	**-0.5%**
Trec8	VSM	0.2700	BOOL	0.2810	0.2809	75%	94%	0.75	**-0.2%**
Trec7	VSM	0.2190	BOOL	0.2320	0.2316	75%	93%	0.75	**-0.2%**
Trec8	PROB	0.2880	BOOL	0.2810	0.2888	71%	94%	0.71	**0.3%**
Trec6	PROB	0.2250	BOOL	0.2450	0.2511	72%	90%	0.70	**2.3%**
Trec6	VSM	0.2360	PROB	0.2250	0.2417	80%	92%	0.80	**2.6%**
Trec7	VSM	0.2190	PROB	0.2260	0.2331	78%	94%	0.77	**3.0%**
Trec7	PROB	0.2260	BOOL	0.2320	0.2449	72%	92%	0.72	**5.6%**

4.1.2 Combinations of Query Representations

We combine Title versions of the query with longer Description versions. These are simply two different expressions of the same information need. It is interesting to note the poor performance of the Description component by itself. This is attributed to the presence of more general terms that serve to dilute the query. However, this poor individual performance does not indicate that the component should not be used in fusion. In fact, the fusion of the description component with the title results in improvement across the board. These results are shown in Table 4.

Table 4: Combining Query Representations

INPUT					FUSION				
Test Set	Method-1 TITLE Alpha = 1.0 Avg Precision		Method-2 DESCRIPTION Beta = 0.5 Avg Precision		Avg Precision	N-Overlap	R-Overlap	Spearman	Change
Trec8	PROB	0.2431	PROB	0.2115	0.2559	36%	83%	0.24	**5.3%**
Trec8	VSM	0.2301	VSM	0.1993	0.2401	37%	83%	0.25	**4.3%**
Trec8	BOOL	0.2438	BOOL	0.1956	0.2572	32%	80%	0.18	**5.5%**
Trec7	PROB	0.1529	PROB	0.1796	0.1806	44%	74%	0.34	**0.6%**
Trec7	VSM	0.1476	VSM	0.1677	0.1707	45%	74%	0.34	**1.8%**
Trec7	BOOL	0.1529	BOOL	0.1695	0.1793	41%	73%	0.29	**5.8%**
Trec6	PROB	0.2107	PROB	0.1489	0.2162	29%	66%	0.12	**2.6%**
Trec6	VSM	0.2003	VSM	0.1426	0.2069	28%	66%	0.12	**3.3%**
Trec6	BOOL	0.2123	BOOL	0.1429	0.2159	25%	63%	0.08	**1.7%**

Table 5: Fusion of Similarity Measures and Query Representations

INPUT					FUSION				
Test Set	Method-1 TITLE Alpha = 1 Avg Precision		Method-2 DESCRIPTION Beta = .5 Avg Precision		Avg Precision	N-Overlap	R-Overlap	Spearman	Change
Trec6	VSM	0.2003	PROB	0.1489	0.2064	28%	66%	0.12	3%
Trec6	VSM	0.2003	BOOL	0.1429	0.2113	24%	63%	0.06	5%
Trec6	PROB	0.2107	BOOL	0.1429	0.2192	23%	62%	0.05	4%
Trec6	PROB	0.2107	VSM	0.1426	0.2153	27%	64%	0.10	2%
Trec6	BOOL	0.2123	VSM	0.1426	0.2108	28%	64%	0.11	-1%
Trec6	BOOL	0.2123	PROB	0.1489	0.2151	28%	65%	0.12	1%
Trec7	VSM	0.1476	PROB	0.1796	0.1748	42%	72%	0.30	-3%
Trec7	VSM	0.1476	BOOL	0.1695	0.1745	37%	71%	0.24	3%
Trec7	PROB	0.1529	BOOL	0.1695	0.1806	36%	71%	0.23	7%
Trec7	PROB	0.1529	VSM	0.1677	0.1769	42%	73%	0.30	5%
Trec7	BOOL	0.1529	VSM	0.1677	0.1752	43%	74%	0.31	4%
Trec7	BOOL	0.1529	PROB	0.1796	0.1800	41%	74%	0.30	0%
Trec8	VSM	0.2301	PROB	0.2115	0.2583	36%	83%	0.23	12%
Trec8	VSM	0.2301	BOOL	0.1956	0.2577	31%	80%	0.16	12%
Trec8	PROB	0.2431	BOOL	0.1956	0.2685	30%	79%	0.16	10%
Trec8	PROB	0.2431	VSM	0.1993	0.2595	36%	81%	0.23	7%
Trec8	BOOL	0.2438	VSM	0.1993	0.2597	36%	82%	0.23	7%
Trec8	BOOL	0.2438	PROB	0.2115	0.2684	35%	82%	0.23	10%

4.1.3 Combinations of Query Representations with Varied Similarity Measures

Next, we assigned retrieval strategies to different query representations and combined those. Here we expect the greatest improvement from fusion because we expect the greatest variation in result sets. This is in fact what we observe in Table 5 with the maximum improvement from fusion reaching 12%.

4.1.4 Combinations of Submissions to TREC-8

In order to compare the impact of the original system variations, we obtained the result sets submitted to TREC-8 by the leading teams and combined those results. The overlap for these TREC-8 submissions is much lower than that of the TREC-3 submissions used in Fox's work [Fox94]. The leading submissions for TREC-3 have N-overlap around 30% and R-overlap around 80%. The trend toward more similar results may be due to the fact

that as techniques such as query expansion and term weighting are shared at the TREC conference, many participants adopt the most successful ones. The relatively low effectiveness of fusion for these TREC-8 'winners' suggests that it may be possible to do better fusion using a common environment than it is by using input from disparate systems. By controlling the variations, one may consciously maximize the good fusion characteristics.

Table 6: Fusion of TREC8 submissions – disparate systems

Queries	Method-1	Method-1 Avg Precision	Method-2	Method-2 Avg Precision	Avg Precision	N-overlap	R-overlap	Spearman	Change
T+D	at99atdc	0.3089	ok8amxc	0.3169	0.3290	60%	91%	0.56	**3.8%**
T+D	at99atdc	0.3089	pir9Attd	0.3207	0.3369	53%	89%	0.47	**5.1%**
T+D	ok8amxc	0.3169	pir9Attd	0.3207	0.3377	54%	89%	0.49	**5.3%**

4.2 Conclusions

A unified environment for fusing retrieval strategies has been introduced. Such a common environment eliminates the question of whether system variations impact fusion effectiveness. It allows control of the variations in input settings to optimize fusion effectiveness. Prior work focused on fusing result sets that were generated from very different parsers, stop word lists, and lexical analysis. We have shown that the combination of baseline vector space, probabilistic and extended Boolean similarity measures, *with no other system variations*, does not result in significant improvement in precision. Further, dissimilarity introduced through varying query representation or through varying query expansion techniques does result in significant improvement in precision.

Our results support the prior work suggestion that overlap is an excellent indicator of the potential for fusing retrieval models. Additionally, we tested various linear combinations of merging VSM with probabilistic and found that when one input set is consistently a lower performer, it can be discounted to good advantage. In the TREC-6 and TREC-8 queries, the description query always does worse than the Title query thus a scalar for the description between 0.5 and 0.7 yields the best fusion. Finally, we have removed any concerns that parsing and other lexical analysis might influence fusion results.

References

[Alaoui98] Alaoui, M. N. Goharian, M. Mahoney, A. Salem, O. Frieder. "Fusion of Information Retrieval Engines (FIRE)", *In Proceedings of the PDPTA*, 1998.

[Bartell94] Bartell, B. T., G.W. Cottrell, and R.K. Belew. "Automatic combination of multiple ranked retrieval Systems" *Proc. of the 17th Annual International ACM-SIGIR* 1994.

[Belkin94] Belkin, N.J., P. Kantor, C. Cool, and R. Quatrain. "Combining evidence for Information Retrieval," *TREC-2*, NIST p. 35-44, edited by D. Harmon, 1994.

[Fox 94] Fox, E. and J. Shaw. "Combination of Multiple Searches," *Proceedings of the Second Text Retrieval Conference (TREC2)*, NIST Special Publication 500-215, 1994.

[Grossman and Frieder 1998] Grossman, D. and O. Frieder. *Information Retrieval: Algorithms and Heuristics.* Kluwer Academic Publishers. Norwell, Mass. 1998.

[Harman98] Harman, D., *editor Proceedings of The Third Text Retrieval Conference (TREC-3)*, sponsored by NIST and Advanced Research Projects Agency, 1998.

[Lee 97] Lee,J.H. "Analysis of Multiple Evidence Combination" *Proceedings of the 20th Annual Int. ACM SIGIR Conference (SIGIR 97)* July 27-31, p. 267-276, 1997

[McCabe99] M. C. McCabe, A. Chowdhury, D. Grossman, O. Frieder, "A Unified Environment for Fusion of Information Retrieval Approaches", *ACM-CIKM*, November 1999.

[Robertson98] Robertson S., S. Walker and M. Beaulieu, "Okapi at TREC-7: Automatic Ad hoc, Filtering, VLC and Interactive," *Proceedings of the Seventh Text Retrieval Conference*, 1998.

[Singhal96] Singhal, A., C. Buckley, and M. Mitra. "Pivoted Document Length Normalization." *Proceedings of the 19th Annual International ACM SIGIR*, August 18-22, 1996.

[Thompson90] Thompson, P. A combination of Expert Opinion Approach to Probabilistic Information Retrieval, part I: The Conceptual Model. *Information Proc. and Mgt.* Vol. 26(3) 1990.

[Vogt 98] Vogt, C. and G. Cottrell., "Predicting the Performance of Linearly Combined IR Systems," *Proceedings of the 21st Annual International ACM SIGIR*, 1998.