Analyses of Multiple-Evidence Combinations for Retrieval Strategies

Abdur Chowdhury, Ophir Frieder, David Grossman, Catherine McCabe
Information Retrieval Lab
Illinois Institute of Technology
{abdur, ophir, dagr, catherm}@ir.iit.edu

1 Introduction

Multiple-evidence techniques are touted as means to improve the effectiveness of systems. Belkin, et al. [1] examined the effects of various query representations. Fox, et al. [2] proposed several combination algorithms and found that combinations of the same types of runs (long and short queries within the vector space model) did not yield improvement and sometimes even degraded performance. He did achieve improvement over individual runs when merging different retrieval strategies (e.g., vector space and p-norm Boolean). Lee [3] further examined various combination algorithms for fusing result sets to improve effectiveness. He identified that, for multiple-evidence to improve system effectiveness, the retrieved sets must have higher relevance overlap than non-relevance overlap. Lee did not identify the exact difference needed to improve effectiveness. His results had a 125% difference in relevant and non-relevant overlap.

While Lee's experiments focused on different system result sets, we focus on effective ranking strategies removing systemic differences of parsers, stemmers, phrase processing and weighting factors. We show that the improvements shown by Lee were likely produced by fusing ranking strategies less tuned than today's measures, and current improvements are likely to be produced by systemic differences rather than ranking strategies.

$$ROverlap = \frac{R \cap S_s \cap S_s ... \cap S_n}{(R \cap S_1) \cup (R \cap S_2) \cup ... (R \cap S_n)}$$

$$NROverlap = \frac{NR \cap S_s \cap S_s ... \cap S_n}{(NR \cap S_1) \cup (NR \cap S_2) \cup ... (NR \cap S_n)}$$

Equation 1: Overlap (R = Relevant, NR = Not Relevant)

2 Experimental Framework

Many factors affect systems performance, namely, parsing, stemming, phrase processing, query representation, weighting of features, ranking strategies, feedback model and collection enrichment. By examining ranking strategies with these factors held constant, we assess the effects of varying ranking strategies towards effectiveness when fusion techniques are applied. Our hypothesis is that the best ranking strategies are more similar than previously thought and when systemic differences are removed, fusion combination approaches are unlikely to provide significant benefit. To test this hypothesis, we implemented four ranking strategies shown to be highly effective in the recent TREC meeting (Pivoted Document Length Normalization [4], BM25 [5], Self-Relevance [6], and IIT [7]).

For each strategy, we evaluated TREC 6, 7, and 8 topics and varied the query lengths, (i.e., title only and title + description). For each retrieved set, we examined the percentage of overlap, the percentage of relevant overlap, and the percentage of non-relevant overlap. Additionally, we fused the final sets with CombMNZ [3] to examine the improvements in terms of effectiveness.

3 Results

We used four ranking strategies (two vector space: pivoted doc length, IIT and two probabilistic BM25, Self-Relevance); each strategy was used to rank the TREC 6, 7, and 8 topics. The first four results in Table 1 and Table 2 show the effectiveness of the various ranking systems for short queries (title only) and longer queries (title + description). While the effectiveness of the various ranking strategies is close, the documents retrieved are not the same.

	TREC-6	TREC-7	TREC-8
Pivoted Doc Len Norm	21.60%	16.37%	22.50%
IIT	23.03%	17.68%	24.58%
BM25	22.86%	17.47%	24.22%
Self-Relevance	22.15%	16.74%	24.44%
Average	22.41%	17.07%	23.94%
CombMNZ	22.98%	17.40%	24.24%
Improvement	2.54%	1.96%	1.27%
Improvement Best	-0.22%	-1.58%	-1.38%

Table 1: Title Only

Large differences in the retrieved document sets and a large difference in the relevant to non-relevant overlap ratio contradicts our hypothesis. If those factors are present then fusion techniques will improve the effectiveness of the system by simply varying the ranking algorithm. Otherwise, improvements gained by varying the ranking method are not improved via fusion.

	TREC-6	TREC-7	TREC-8
Pivoted Doc Len Norm	24.33%	18.72%	24.57%
IIT	25.47%	20.17%	26.66%
BM25	25.84%	20.00%	26.37%
Self-Relevance	24.71%	19.41%	27.03%
Average	25.09%	19.58%	26.16%
CombMNZ	25.57%	20.01%	26.69%
Improvement	1.92%	2.22%	2.04%
Improvement Best	-1.04%	-0.79%	-1.26%

Table 2: Title + Description

With systemic differences removed how similar are the various strategies? In Table 3, we illustrate that the overlap (OLAP) of the four strategies is quite high. As the query length increases (title +

description), the differences decrease a small amount. Additionally, the difference between relevant overlap and non-relevant overlap is only 23-33% where Lee found a 15% similarity in overlap and a 125% difference between R and NR overlap. Since the overlap is high and the R and NR differences are low, as shown in Table 1 and Table 2, the likelihood of improvements in effectiveness from fusion is low.

The results of the four sets with CombMNZ fusion are illustrated in the lower three rows of Table 1 and Table 2. As expected, the improvement with fusion is small over the average precision of the retrieved sets and results in lower scores than the best system. While this does not conclusively prove our hypothesis it leads us to several questions, namely, "Why are the various ranking strategies not producing good fusion sets?" and "Why did Lee see improvements in his experiments?".

	OLAP	R-OLAP	N-OLAP	DIFF
T6 - Title	65.26%	90.26%	72.38%	24.70%
T7 - Title	67.80%	92.30%	74.62%	23.70%
T8 - Title	65.98%	91.50%	71.87%	27.32%
T6 - T+D	61.69%	89.07%	68.32%	30.38%
T7 - T+D	63.95%	91.62%	70.62%	29.74%
T8 - T+D	62.86%	91.73%	68.84%	33.25%

Table 3: Ranking Strategy Similarity

To answer the above questions, we downloaded the result sets from NIST for TREC3 and reproduced Lee's results for fusion. Additionally, we downloaded the best three systems for TREC 3 and fused those results. In Figure 1, we show that as the various results used by Lee were combined, the effectiveness increased over both the average of the runs and the best of the six runs.

When the best three TREC 3 runs were combined (Figure 2), the improvement due to fusion is small. Why are the most effective systems not good candidates for fusion? The six sets chosen by Lee have a 125% difference between the relevant and non-relevant overlap and only a 15% overlap of documents (Table 4). Thus, when combined they produced a 39% increase in effectiveness, although this improvement was not better than the best system alone (40.12%). When the best techniques were fused with a 52% difference of R and NR and a 42% overlap, the improvement was only 4% over the best system. This tends to back our hypothesis that the benefits from fusion when used against highly effective systems in not significant.

To further evaluate our hypothesis, we implemented three variations of tf-idf and two SMART (ann, lnc-ltn) ranking strategies to simulate the various systems that were being experimented with during TREC 3. An overlap of 23% was observed, and a 72% difference in R and NR overlap was noted. When the sets were fused, an improvement of 23.6% in effectiveness was achieved (TF*IDF,T6 - Table 4). When this improved set is compared to the most effective rankings it was still less effective than the current ranking algorithms. We believe that as the ranking strategies are more effective the benefit of fusion decreases. We also examined the three best runs from TREC 8 and observed similar results of convergence minimizing the benefits of fusion for the best systems.

4 Summary and Conclusions

We removed systemic differences and compared effective ranking strategies in the context of multiple evidence systems. We show that when systemic differences are removed and effective ranking strategies are used, improvements in average precision due to fusion are negligible. Furthermore, prior results indicating gains from fusion did so using strictly relatively poor ranking strategies. Similar improvements obtained when using poor ranking strategies were observed with phrase processing improvements [8]. With today's higher quality ranking strategies, it is not at all clear that fusion of similar query representations provides significant benefit.

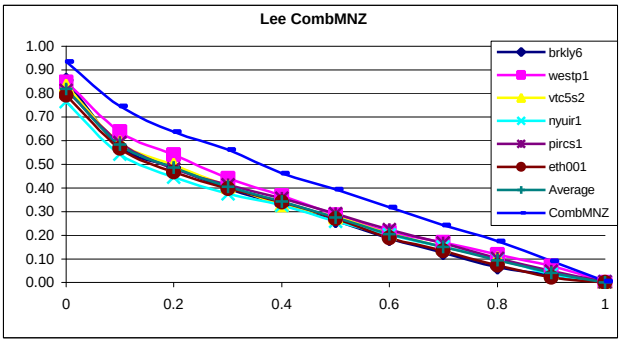

Figure 1: TREC 3 Lee Experiments

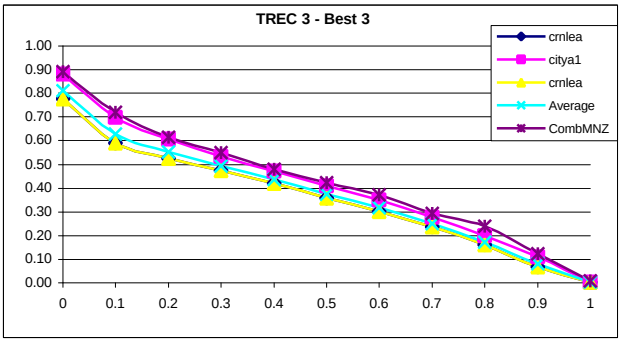

Figure 2: TREC 3 Best 3 runs

	OLAP	R-OLAP	N-OLAP	DIFF	AVG P/R	IMP
Lee - 6 way	15.18%	71.49%	31.68%	125.67%	28.84%	38.35%
T3 - Best 3	42.14%	84.11%	55.02%	52.87%	36.97%	13.26%
TF*IDF, T6	21.63%	79.22%	45.98%	72.30%	12.46%	23.68%
T8 - Best 3	43.83%	86.77%	58.26%	48.94%	31.80%	9.01%

Table 4: Lee Experiments

[1] N.J. Belkin, et al., "Combining the evidence of multiple query representations for information retrieval," Information Processing & Management, 31(3), pp. 431-448, 1995.

[2] E.A. Fox and J.A. Shaw, "Combination of multiple searches," NIST TREC-2, pp. 243-252, 1994.

[3] J. Lee, "Analyses of multiple evidence combination", ACM-SIGIR, pages 267-276, Philadelphia, 1997.

[4] A. Singhal, et al., "Pivoted document length normalization", ACM-SIGIR, 1996.

[5] S. Robertson, et al., "Okapi at TREC-4", NIST TREC-4, November 1995.

[6] K. Kwok, et al., "TREC-7 Ad-Hoc, High precision and filtering experiments using PIRCS", NIST TREC-7, November 1998.

[7] A. Chowdhury, et al., "Improved query precision using a unified fusion model", NIST TREC-9, November 2000.

[8] A. Turpin, and A. Moffat, "Statistical phrases for Vector-Space information retrieval", ACM-SIGIR, 1999: 309-310.