Disproving the Fusion Hypothesis: An Analysis of Data Fusion via Effective Information Retrieval Strategies

Steven M. Beitzel Eric C. Jensen Abdur Chowdhury
Ophir Frieder David Grossman Nazli Goharian
Information Retrieval Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
{steve,ej,abdur,frieder,grossman,goharian}@ir.iit.edu

ABSTRACT

Many prior efforts have been devoted to the basic idea that data fusion techniques can improve retrieval effectiveness. Recent work in the area suggests that many approaches, particularly multiple-evidence combinations, can be a successful means of improving the effectiveness of a system. Unfortunately, the conditions favorable to effectiveness improvements have not been made clear. We examine popular data fusion techniques designed to achieve improvements in effectiveness and clarify the conditions required for data fusion to show improvement. We demonstrate that for fusion to improve effectiveness, the result sets being fused must contain a significant number of unique relevant documents. Furthermore, we show that for this improvement to be visible, these unique relevant documents must be highly ranked. In addition, we present a comprehensive discussion on why previous assumptions about the effectiveness of multiple-evidence techniques are misleading. Detailed empirical results and analysis are provided to support our conclusions.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models.*

General Terms

Experimentation

Keywords

Data Fusion, Multiple Evidence, Information Retrieval

1 INTRODUCTION

In recent years, the category of work known as data fusion described a range of techniques in information retrieval whereby multiple pieces of information are combined to achieve improvements in retrieval effectiveness. These pieces of information can take many forms including different query representations, different document representations, and different retrieval strategies used to obtain a measure of relationship between a query and a document. Several researchers have used combinations of different retrieval strategies to varying degrees of success in their systems [2, 8]. Belkin, et al. examined the effects of combining several different query representations to achieve improvements in effectiveness [3, 4]. Chowdhury and colleagues used Query Length Normalization to maximize usage of similarity scores taken from multiple query representations using a single engine [7]. Lee examined the effect of using different weighting schemes to retrieve different sets of documents using a single query and document representation, and a single retrieval strategy [10].

Our goal in this study was to examine data fusion of highly effective strategies in an attempt to create a fused result set that has better mean average precision than the most effective single system. This approach differs from the usual goal of data fusion applied to metasearch or distributed retrieval. In these cases, fusion is used to determine which result documents to select for an integrated result set. We are trying to use fusion solely for improving retrieval effectiveness with highly effective retrieval strategies.

When work on data fusion was initially done, there was little understanding as to why the techniques mentioned above helped to bring about an improvement in effectiveness. This persisted until Lee performed an analysis of data fusion techniques [11]. In that work, he examined several multiple-evidence combination approaches and concluded that improvements in retrieval effectiveness due to fusion were directly related to the level of overlap present in the results from each approach being combined. Specifically, Lee hypothesized that for multiple-evidence techniques to improve effectiveness, the retrieved sets from each approach must have a higher relevant overlap than non-relevant overlap, although an optimal ratio of these quantities is not provided. The formulas for calculating relevant overlap and non-relevant overlap are shown in Equation 1. The experimentation provided by Lee shows significant improvements for fused result sets, thus appearing to support Lee's original hypothesis. Unfortunately, there are two key points that Lee did not account for, which limit the conclusions that can be safely drawn from his work. In his experiments, Lee *did not use the most effective result sets available*, but rather, selected his test sets at random. Furthermore, he used result sets from *entirely different* information retrieval systems. *This does not simply vary the retrieval strategy used for the experiments, but all retrieval utilities and other systemic differences.* These include things such as parsing rules, stemming, phrase processing, relevance feedback

SAC 2003, Melbourne, Florida, USA

techniques, etc. The failure to account for these points in the experimentation makes it difficult to isolate the factors that are directly contributing to the effectiveness of data fusion techniques. We have corrected for the ambiguities in Lee's experimentation by performing fusion on result sets from highly effective separate systems, and on result sets from highly effective retrieval strategies inside the same system. We show that fusion is only effective when it causes an increase in recall of highly ranked documents. This occurs when the component result sets contain a large number of relevant documents that are unique across the component sets, and those documents are highly ranked. The remainder of this paper is organized as follows: in Section 2, we give a detailed discussion of prior work and clarify the motivations for our hypothesis. In Section 3 we discuss our methodology. In Section 4 we present our experimental results and analysis. Finally, we summarize our conclusions in Section 5.

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

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

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

2 ANALYSIS OF PRIOR APPROACHES

Prior to Lee's analysis, there was a great deal of speculation as to what conditions might lead to an optimal effect of fusion on retrieval effectiveness. Most notably, Belkin and his colleagues were among the first to postulate that there is a relationship between results overlap and success with fusion. Belkin based his arguments on the premise that differently-configured experiments (those using different query or document representations or different retrieval strategies) are likely to return different sets of relevant documents and different sets of non-relevant documents. The development of this rationale is discussed in detail in [11], along with further discussion that, when taken in conjunction with Turtle and Croft's analysis [14] of multiple query representations, leads to a corollary of Belkin's original postulate. Namely, this corollary states that improvements might be seen from fusion even when the result sets are similar, as long as the difference in relevant overlap is greater than the difference in non-relevant overlap. Given this, it is possible to design result set combination algorithms that increase the score of a document based on repeated evidence of its relevance, as done by Fox and Shaw in [8]. One of the algorithms designed by Fox and Shaw, CombMNZ, has proven to be a simple, effective method for combining result sets. It was used by Lee in his fusion experiments, and has become the standard by which newly developed result combination algorithms are judged. More recent research in the area of metasearch engines has led to the proposal of several new result combination algorithms, making use of training data and techniques such as voting algorithms and Bayesian inference [1, 13]. Although these algorithms have been shown to behave comparably and occasionally superior to CombMNZ, they did not exist when Lee performed his initial experiments, and so in-depth analysis of their performance is left to the scope of future work.

As stated above, Lee showed significant improvements when fusing random, heterogeneous result sets with common data fusion algorithms. More recent work by Montague, et al., provides experimentation performed under similar conditions to Lee's work, and shows similar results [12]. Given that results showing fusion to be effective exist, there is a surprising lack of detail surrounding the analysis of *why* it is effective, save for Lee's basic assumptions about overlap. To date, no detailed analysis exists in the literature of exactly how factors such as overlap and systemic differences affect the performance of fusion.

Having reviewed the evolution of the prior work and the basis for Lee's assumptions, we must examine the key limitations with Lee's experiments to truly investigate the reasons behind fusion's reported benefits. Lee's experiments proceed under the assumption that as long as result sets involved in fusion have greater relevance overlap than non-relevance overlap, there will be an improvement. To justify this, his experiments used a series of result sets that had merely 15% overlap, and a 125% difference in relevant and non-relevant overlap. In addition, the result sets used for the experiments were chosen at random, and were not the most-effective result sets from the available pool (the third Text Retrieval Conference). Lee's work contained no analysis of the relative effectiveness of fusion when using random sets versus using the most effective available sets, and it contained no comparison between the effectiveness of his fused results and the effectiveness of the best system at TREC-3. As stated above, another key issue with Lee's experimental environment is that his experiments were performed via the fusion of result sets from entirely different information retrieval engines. Performing fusion in this way varies important systemic differences, thereby introducing more than one independent variable and making it difficult to derive sound conclusions from the data. To truly study the effects of fusing retrieval strategies alone, systemic differences must be held constant. Lee did not analyze the effect of varied systemic differences on fusion's effectiveness in his work. Given these points, it is difficult to generalize based on Lee's experiments, and it is clear that a fully controlled environment with the best possible result sets must be used to fully evaluate the effectiveness of data fusion techniques.

Some recent work [16] has focused on analyzing the effects of using average systems vs. highly effective systems for fusion. Soboroff, et al. developed a system to generate pseudo-relevance judgments for a document collection based on pooling and ultimately found that although their model proved effective in predicting the behavior of average retrieval systems, it fared quite poorly in predicting the behavior of very good retrieval systems. This tends to suggest that highly effective retrieval strategies retrieve different relevant documents. Chowdhury, et al. [6] began an investigation of fusing highly effective retrieval strategies. While their data was limited, they formed initial conclusions suggesting that fusion of highly effective strategies does not tend to improve effectiveness. This shows motivation for further work in this area, and lends merit to more detailed investigation of the causative factors of improvement from data fusion.

3 METHODOLOGY

The first step in a detailed analysis of why fusion works is an examination of Lee's key claim that the effectiveness of fusion is directly related to the difference between the relevant and non-relevant overlap of the component result sets. Specifically, Lee indicated that higher differences between relevant and non-relevant overlap would lead to greater effectiveness improvements. To explore this claim, it is necessary to first ensure that the experimental environment is properly controlled.

We implemented several highly effective retrieval strategies for fusion in the same system, thereby removing any interference from systemic differences, and also fused the top three systems over several years of the Text Retrieval Conference to get a clear representation of the behavior of fusion on highly effective result sets. We performed an analysis of the difference between the relevant and non-relevant overlap for these fused sets and found that, in fact, Lee's assumption does not hold true. A summary of these experiments is presented in detail in Section 4.

After failing to validate Lee's primary assumption about overlap, we returned to the primary motivation for performing data fusion: *fusing multiple strategies should enable improved retrieval effectiveness over the best available single-system.* Going back again to Lee's experiments, it can be seen that although he came close to the top system, he fell short of exceeding its effectiveness. The remainder of our study focused on what conditions must exist to get fusion to elevate effectiveness past the level of the best approach available.

The next step in our study was to expand on the implications of the work done in [16], namely that highly effective retrieval strategies tend to return different relevant documents. If this is indeed true, we reasoned that fusion of highly effective retrieval strategies might yield improvement if recall were improved, particularly for relevant documents at the top of the component result sets used in fusion. If a relatively large number of unique relevant documents are found during the fusion process, and ranked highly in the fused result set, it would raise average precision. The converse of this theory would also explain why multiple-evidence combinations of retrieval strategies that retrieve many of the same documents show little improvement (as found in the study by [6]). If the majority of the documents, relevant and non-relevant alike, are shared between component result sets, then multiple-evidence combination algorithms such as CombMNZ will simply scale the scores of most of the documents, which will not lead to an improvement in effectiveness. We designed several experiments to prove this hypothesis; they are discussed in detail in Section 4.

4 RESULTS AND ANALYSIS

For our experiments, we implemented three modern retrieval strategies that were recently shown to be highly effective, one Vector-Space and two Probabilistic (IIT [5], BM25 [15], Self-Relevance [9]). A single information retrieval engine was then used with each of these retrieval strategies to evaluate query topics from the ad-hoc track at TREC 6, 7, and 8, and also query topics from the web track at TREC-9 and TREC-10. All of our experiments used only the title field of the TREC topics.

Our first set of experiments were designed to determine the validity of Lee's assumption about improvements in effectiveness due to fusion being dependent on the difference between relevant and non-relevant overlap. According to Lee, the larger the difference in relevant and non-relevant overlap, the greater the improvements from fusion should be. To examine this, we first fused each of our three highly effective retrieval strategies inside the same information retrieval system, and compared the effectiveness of these fused result sets to the effectiveness of the best of the three systems. We illustrate this in Table 1.

Table 1: Improvement of Same-System Retrieval Strategies

	Trec6	Trec7	Trec8	Trec9	Trec10
Best	0.1948	0.1770	0.2190	0.1847	0.1949
Fused	0.1911	0.1751	0.2168	0.1671	0.1935
Imp/Best	-1.90%	-1.07%	-1.005	-9.53%	-0.72%

We then performed a detailed overlap analysis of these results, shown in Table 2.

Table 2: Overlap of Same-System Retrieval Strategies

	Trec6	Trec7	Trec8	Trec9	Trec10
Overlap	62.76%	61.14%	59.42%	61.61%	59.17%
R Overlap	89.52%	89.90%	90.23%	88.61%	85.88%
NR Overlap	72.93%	72.82%	72.03%	71.49%	68.94%
%Diff R/NR	22.75%	23.46%	25.27%	23.95%	24.57%

The second set of experiments in which we test Lee's overlap claim involves fusing the three best result sets from distinct TREC competitors for all years with title-only results available. The improvement and overlap analysis are shown in Table 3 and Table 4.

Table 3: Improvement of Best TREC Systems

	Trec6	Trec7	Trec8	Trec9	Trec10
Best	0.2876	0.2614	0.3063	0.2011	0.2226
Fused	0.3102	0.2732	0.3152	0.2258	0.2441
Imp/best	7.86%	4.51%	2.91%	12.28%	9.66%

Table 4: Overlap of Best TREC Systems

	Trec6	Trec7	Trec8	Trec9	Trec10
Overlap	34.43%	39.31%	42.49%	30.09%	33.75%
Rel Overlap	83.08%	80.84%	84.63%	85.85%	81.87%
NRel Overlap	53.33%	56.36%	57.13%	51.26%	54.01%
% diff R/NR	55.78%	43.44%	48.14%	67.48%	51.58%

When fusing separate systems, we do see small to moderate improvements with fusion, however, if Lee's claim were true, one would expect effectiveness improvements to increase monotonically as difference in relevant and non-relevant overlap increases. This is clearly not the case, as can be seen from Tables 1-4 above. Generally, overlap is lower (30-43% - see Table 4Table 4) in cases where there is some improvement over the best system (2.9-12.3% - see Table 3), as opposed to cases where little or no improvement (and occasionally loss) is observed (59-63% - see Table 1 and Table 2).

To further the examination of our hypothesis, we first decided to test our supposition that fusion only yields improvement when the component result sets contain a relatively large number of unique relevant documents. To measure this, we took each component result set and merged them such that the top X documents were

examined, and any document appearing in more than one result set was discarded. This was done for various values of X so that we could observe the number of unique relevant documents present at different depths of the component result sets. The above experiments were done both for fusion of the best TREC systems and for the fusion of the three highly effective retrieval strategies in the same system. We plotted out the results in a series of graphs, one per TREC-Year. Each graph shows the percentage of uniquely relevant documents present at various depths of examination. Two curves are shown on each graph: one representing the fusion of the top three TREC systems for that year (marked as "best"), and a second curve representing the fusion of the three highly effective strategies in the same information retrieval system.

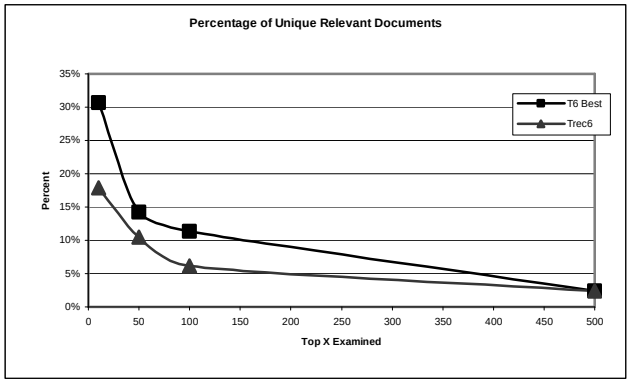

Figure 1: TREC-6 Unique Relevant Document Analysis

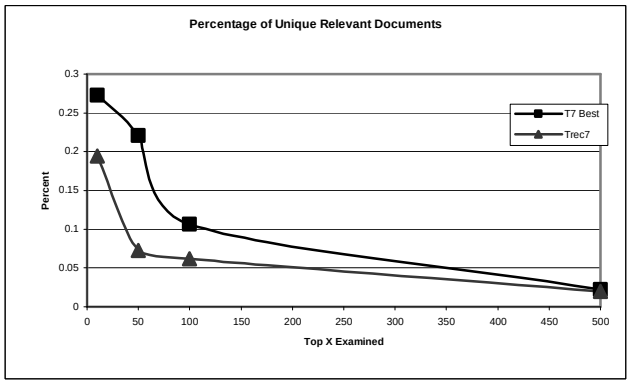

Figure 2: TREC-7 Unique Relevant Document Analysis

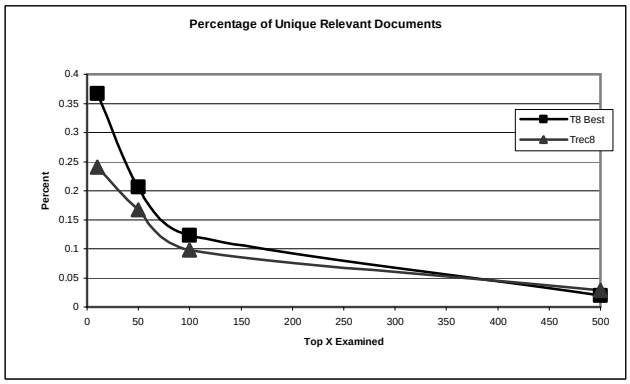

Figure 3: TREC-8 Unique Relevant Document Analysis

These graphs above clearly show that for each TREC year, the fusion of the top three systems has a higher percentage of unique relevant documents in its result set for a given depth X. It is particularly interesting to note that the percentage of unique relevant documents is always greatest near the top of the result set. This means that recall is improved for the highest ranked documents. If our hypothesis about the relationship between percentage of unique relevant documents and effectiveness improvements is correct, then according to the graphs above we would expect to see that the fusion of the top 3 systems always yield a greater improvement over the best single system.

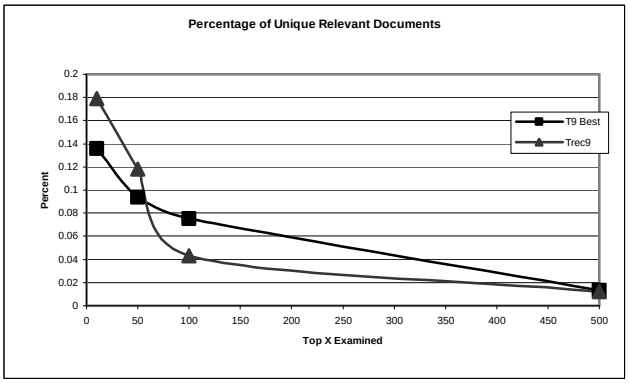

Figure 4: TREC-9 Unique Relevant Document Analysis

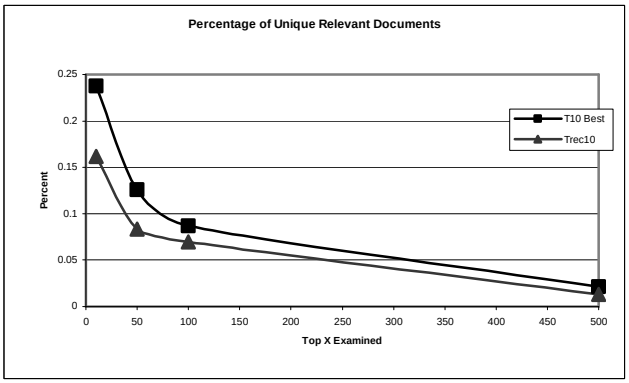

Figure 5: TREC10 Unique Relevant Document Analysis

Referring back to Table 1 and Table 3 shows us that our data concurs with this expectation. To explain this we can first refer back to the earlier observation that the percentage of unique relevant documents in the result set was always at its highest when examining only the top documents in each component set. Therefore, when this is true, the probability of having a noticeable effect on average precision is high since fusion is allowing recall to improve by merging in different relevant documents at the highest ranked positions in the result set. Greater clarity can be achieved by examining the average number of unique (across component sets) relevant and non-relevant documents added to the result set at various depths by fusion.

Table 5: Avg. # Unique R & NR added in same-system fusion

Depth	R	NR	Ratio
10	0.72	3.18	0.23
50	1.29	11.83	0.11

100	1.53	21.97	0.07
500	1.60	89.84	0.02

Table 6: Avg. # Unique R & NR added in TREC-best fusion

Depth	R	NR	Ratio
10	1.49	4.30	0.35
50	3.46	19.77	0.17
100	3.93	36.63	0.11
500	3.19	157.61	0.02

It can be seen from the tables above than in cases where fusion shows improvement (TREC-best), the average number of relevant documents added to the highly ranked documents (depth = 10) is roughly doubled over the same-system case, while the average number of non-relevant documents is only increased by 25%.

It is still desirable to explain why multiple-evidence alone is not enough to yield significant improvement for fusion over the best single system when fusing highly effective systems or retrieval strategies. The reason for this is simply because fusing sets of documents that are very highly similar (i.e., they have high general overlap), then multiple-evidence techniques will simply scale the scores of the majority of the documents and will not help in separating relevant documents from non-relevant ones. Consequently, when general overlap is high, the number of unique (non-repeated) documents will be lower, and improvements due to fusion will be very unlikely.

5 CONCLUSIONS AND FUTURE WORK

We have experimentally shown that multiple-evidence alone is not enough to ensure effectiveness improvements when fusing highly effective retrieval strategies. In order to use data fusion techniques for improving effectiveness, there must be a large percentage of unique relevant documents added to the fused set as highly ranked results, not a simple difference between relevant and non-relevant overlap as previously thought. We investigated and identified the relationship between overlap of result sets and fusion effectiveness, demonstrating that fusing result sets with high overlap are far less likely to yield a large improvement than fusing those with low overlap, if the sets being fused are highly effective. We also identified that varying systemic differences amongst result sets tends to bias improvements that have been seen in fusion experiments from the prior work, and shown that when these differences are removed, causation factors of fusion are more easily studied. For future work, we plan to investigate the specific effects that various systemic variations have on fusion effectiveness, and research the development and performance of new and existing intelligent data fusion algorithms that might overcome the limitations of those commonly used today.

6 REFERENCES

[1] J. Aslam and M. Montague, et al., “Models for Metasearch”, Proceedings of the 24th Annual ACM-SIGIR, 2001.

[2] B.T. Bartell, G.W. Cottrell, and R.K. Belew, “Automatic Combination of multiple ranked retrieval systems,” Proceedings of the 17th Annual ACM-SIGIR, pp. 173-181, 1994.

[3] N.J. Belkin, C. Cool, W.B. Croft and J.P. Callan, “The effect of multiple query representations on information retrieval performance,” Proceedings of the 16th Annual ACM-SIGIR, pp. 339-346, 1993.

[4] N.J. Belkin, P. Kantor, E.A. Fox, and J.A. Shaw, “Combining evidence of multiple query representation for information retrieval,” Information Processing & Management, Vol. 31, No. 3, pp. 431-448, 1995.

[5] A. Chowdhury, et al., "Improved query precision using a unified fusion model", Proceedings of the 9th Text Retrieval Conference (TREC-9), 2000.

[6] A. Chowdhury, et al., “Analyses of Multiple-Evidence Combinations for Retrieval Strategies”, Proceedings of the 24th Annual ACM-SIGIR, 2001.

[7] A. Chowdhury, S. Beitzel, E. Jensen, "Analysis of Combining Multiple Query Representations in a Single Engine", Proceedings of the 2002 IEEE International Conference on Information Technology - Coding and Computing (ITCC), Las Vegas, April 2002.

[8] E.A. Fox and J.A. Shaw, “Combination of Multiple Searches,” Proceedings of the 2nd Text Retrieval Conference (TREC-2), NIST Special Publication 500-215, pp. 243-252, 1994.

[9] K. Kwok, et al., "TREC-7 Ad-Hoc, High precision and filtering experiments using PIRCS", Proceedings of the 7th Text Retrieval Conference (TREC-7), 1998.

[10] J.H. Lee, “Combining Multiple Evidence from Different Properties of Weighting Schemes,” Proceedings of the 18th Annual ACM-SIGIR, pp. 180-188, 1995.

[11] J.H. Lee, “Analyses of Multiple Evidence Combination,” Proceedings of the 20th Annual ACM-SIGIR, pp. 267-276, 1995.

[12] M. Montague, et al., “Relevance Score Normalization for Metasearch”, Proceedings of ACM-CIKM, 2001.

[13] M. Montague, et al., “Condorcet Fusion for Improved Retrieval”, Proceedings of ACM-CIKM, November 2002.

[14] H. Turtle and W.B. Croft, “Evaluation of an inference network-based retrieval model,” ACM Transactions on Information Systems, Vol. 9, No. 3, pp. 187-222, 1991.

[15] S. Robertson, et al., "Okapi at TREC-4", Proceedings of the 4th Text Retrieval Conference (TREC-4), 1995.

[16] I. Soboroff, et al., “Ranking Retrieval Systems without Relevance Judgments”, Proceedings of the 24th Annual ACM-SIGIR, 2001.