On Fusion of Effective Retrieval Strategies in the Same Information Retrieval System

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder, Nazli Goharian
Information Retrieval Laboratory, Illinois Institute of Technology, Chicago, IL 60616
{steve,ej,abdur,dagr,ophir,nazli}@ir.iit.edu

Prior efforts have shown that under certain situations, retrieval effectiveness may be improved via the use of data fusion techniques. Although these improvements have been observed from the fusion of result sets from several distinct information retrieval systems, it has often been thought that fusing different document retrieval strategies in a single information retrieval system will lead to similar improvements. In this study, we show that this is not the case. We hold constant systemic differences such as parsing, stemming, phrase processing, and relevance feedback, and fuse result sets generated from highly effective retrieval strategies in the same information retrieval system. From this, we show that data fusion of highly effective retrieval strategies alone shows little or no improvement in retrieval effectiveness. Furthermore, we present a detailed analysis of the performance of modern data fusion approaches, and demonstrate the reasons why they do not perform well when applied to this problem. Detailed results and analyses are included to support our conclusions.

Introduction

Recently there has been much research done in the field of information retrieval concerning the various kinds of "fusion" and their applications. One method of fusion, called collection fusion, is the fusing of results from multiple, autonomous collections of data, and is often used in the context of metasearch and distributed Information Retrieval. An excellent introduction to the collection fusion problem can be found in (Voorhees, Gupta, & Johnson-Laird, 1994). Data Fusion is the combination of multiple pieces of evidence of relevance, such as different query representations, different document representations, and different retrieval strategies used to obtain a measure of similarity between a query and a document. This combination is then utilized to improve retrieval effectiveness, and is most often applied to the task of ad-hoc retrieval.

This paper examines some long-held beliefs about common, effective data fusion techniques. Prior work demonstrates that significant improvement is often seen when using standard data fusion algorithms on an arbitrary collection of result sets from different information retrieval systems (Lee, 1997). This belief is supported by the supposition that different document retrieval strategies will rank documents differently, returning different sets of relevant and non-relevant documents. If this is true, it follows that voting algorithms that boost score and rank of a relevant document that is agreed upon across component systems should indeed improve retrieval (via increasing average precision), and merging algorithms can also be used to increase the recall of relevant documents in the merged set. This has led to the development of common techniques that employ both voting and merging, such as the widely-used CombMNZ approach (Aslam & Montague, 2001; Fox & Shaw, 1994; Montague & Aslam, 2001, 2002). Past researchers have also attempted to predict whether or not fusion will yield improvements based on some properties of either the component retrieval systems used in the fusion process, or their result sets. Unfortunately, there was little understanding of exactly why fusion techniques consistently brought effectiveness improvements. Lee addressed this research question in a study that proposed a correlation between the overlap of the relevant and non-relevant document result sets, and the expected improvement from common data fusion techniques (Lee, 1997). Specifically, Lee stated that for data fusion techniques to improve retrieval effectiveness, the retrieved result sets from each approach being fused must have a greater overlap of relevant documents than of non-relevant documents. Although an optimal ratio of these overlap parameters was not provided by Lee, following research did seem to substantiate his claims

(McCabe, Chowdhury, Grossman, & Frieder, 1999; Vogt & Cottrell, 1998, 1999). We examine this study and related work in detail, and provide an explanation for why, although seemingly intuitive, the conclusions reached in that study are not applicable to the case of fusing only highly effective retrieval strategies. Our goal in this study is to examine data fusion of highly effective document retrieval strategies within the same information retrieval system, while holding all other systemic differences constant. Here we define a retrieval "strategy" as a method of assigning similarity between a query and a document (typically a ranking algorithm), where as a retrieval "system" is an entirely autonomous Information Retrieval Engine with its own independent systemic properties. These systemic properties include parsing rules, stemming rules, relevance feedback algorithms, phrase processing, query processing, document representation, etc. We hold these variables constant to observe the effect of fusing only different retrieval strategies and nothing more. Under these conditions we show that Lee's overlap hypothesis is actually a poor indicator of fusion's ability to improve effectiveness. Furthermore, we examine the fusion process in detail and show that any improvements gained from fusion techniques such as CombMNZ are likely due to an increase in the recall of highly ranked relevant documents in the fused result set.

This paper is organized as follows: in section 2 we present a detailed review of the literature in the area of data fusion techniques. Section 3 explores our motivations and experimental methodology. Our hypothesis is discussed in detail, and the experiments we have designed to prove our hypothesis are described. Section 4 contains our experimental results and analysis. Section 5 contains our conclusions, and ideas for promising future work in this area.

Prior Work

Data Fusion refers to a set of techniques whereby multiple pieces of evidence of relevance are utilized to improve the effectiveness of information retrieval systems. Prior work has shown that many different types of evidence have been utilized in an attempt to improve retrieval, including different query representations, different document representations and indexing strategies, and different retrieval strategies, or methods of finding a measure of similarity between a query and a document. From the literature, a variety of techniques that have been developed for performing data fusion can be found. These techniques are useful in several different applications of information retrieval, including the ad-hoc retrieval task commonly associated with the annual Text Retrieval Conference (TREC), as well as tasks in the area of distributed information retrieval and metasearch on the web. What follows is a comprehensive survey of prior work in this area, which will help to illuminate the motivations for this study.

One of the earliest efforts in data fusion is detailed in (Nicholas. J. Belkin, Cool, Croft, & Callan, 1993; Nicholas J. Belkin, Kantor, Fox, & Shaw, 1995). Belkin and his colleagues were among the first researchers to investigate the effect of data fusion techniques on retrieval effectiveness. They noted that several research efforts prior to their own made the empirical observation that using different retrieval techniques, query representations, or document representations often led to result sets with surprisingly little overlap (Katzer, McGill, Tessier, Frakes, & Dasgupta, 1982; McGill, Koll, & Norreault, 1979; Saracevic & Kantor, 1988). This was also discussed in a theoretical nature within the realm of probabilistic retrieval and inference networks (Robertson, 1977; Turtle & Croft, 1991). Many of these early research efforts made attempts to use the low-overlapping result sets from different experiment configurations for improving retrieval effectiveness.

Saracevic & Kantor used independently-generated query representations to create a number of result sets, and found that a document was increasingly more likely to be judged relevant as the number of retrieved sets in which it appeared increased (Saracevic & Kantor, 1988). Turtle & Croft performed similar experiments using an inference network, and found that combining different query representations led to increased retrieval effectiveness over any single representation (Turtle & Croft, 1991). Foltz & Dumais found similar improvements by combining multiple retrieval strategies, lending credence to Data Fusion as a general technique, not tied merely to query representations alone (Foltz & Dumais, 1992). Based on this information, Belkin and his colleagues set out to expand on the prior work using the large (for the time) 2GB TREC collection from TREC-1. They created several different Boolean query representations, and tracked effectiveness improvements over a large number of combinations. Ultimately they found improvements consistent with the prior work of the time, and concluded that combining multiple pieces of evidence was nearly a surefire way to increase retrieval effectiveness,

suggesting that as more evidence of relevance becomes available for combination, greater improvement can be expected.

Belkin's conclusions led to further research in the area of fusion, and most notably the development, by Fox & colleagues, of several result combination algorithms that use both voting and merging principles to combine evidence from several different sources (Fox & Shaw, 1994). One of the algorithms defined in this study, CombMNZ, has become the standard method of combining results from multiple searches in data fusion experiments. The development of these result combination algorithms further stimulated data fusion as a viable research topic, and in recent years a large number of studies have been devoted to it. Armed with Belkin's postulate of "more is better" and a standard method of experimentation in CombMNZ, researchers turned their focus to optimizing the improvements gained from data fusion and isolating the conditions required for data fusion to be most beneficial.

Lee did some initial work in trying to maximize effects gained from data fusion by exploring the effectiveness of combining the results from several term-weighting schemes with different properties in order to retrieve more types of relevant documents (Lee, 1995). Lee classified documents in his collection into several types, and combined term-weighting schemes that were each engineered to bring back separate types of documents using Fox & Shaw's CombSUM results combination method. He found that when performing combinations in this matter, significant improvements could be achieved. Although no overlap analysis is given, the improvements in this study were most probably due to a general increase in recall, given that the combinations were specifically designed to retrieve documents of different types. Unfortunately it is not always feasible to examine and classify the target document collection, therefore it is difficult to generalize the effectiveness of this technique to pure ad-hoc retrieval. Lee furthered his efforts on data fusion with another study that proposed a correlation between the level of difference between relevant and non-relevant overlap among component systems and the degree of improvement that can be expected from voting/merging fusion techniques such as CombMNZ (Lee, 1997). Specifically, Lee stated that as long as the component systems being used for fusion had greater relevant overlap than non-relevant overlap, improvement would be observed, although an optimal ratio of these quantities was not provided. The formulae for calculating relevant overlap and non-relevant overlap for result sets S_1 through S_n are shown in Equation 1.

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

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

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

The experimentation provided in the study shows significant improvements for fused result sets, thus appearing to support the overlap hypothesis. Unfortunately, there are two key points that were left unaccounted for, which limit the conclusions that can be safely drawn from this study. In the experiments, the result sets being fused *were not the most effective result sets available*; they were selected at random from a large pool of result sets from TREC-3. Furthermore, the result sets used were 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 differences include things such as parsing rules, stemming, phrase processing, relevance feedback 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. In more recent research, Lee also investigated the effects of fusing different *query representations,* obtained from various relevance feedback methods, in the same information retrieval system. (Lee, 1998). While this method showed significant improvements in retrieval effectiveness, it is an entirely separate approach from fusing document retrieval strategies since in this case different sets of terms are used to retrieve the documents that form each component result set. As a result, this study suggests that fusing result sets formed from different query representations may bring improvement, but it does nothing to clarify the effects of fusing different document retrieval strategies while holding other systemic differences constant.

Although Lee's retrieval strategy study only provided limited experimental data to support the overlap hypothesis, and didn't specify a particular threshold around which fusion would definitely become worthwhile. Much of the following research proceeded under the assumption that the overlap to improvement correlation was well founded. Another popular avenue for optimizing data fusion improvements gave even more weight to Lee's

proposed overlap correlation. A series of studies was performed using linear combinations of sources - essentially giving a weight of confidence in the quality of a source before fusing with a common results combination algorithm like CombMNZ. Bartell and colleagues were responsible for some of the first work done in linear combinations (Bartell, Cottrell, & Belew, 1994). They used numerical optimization techniques to determine optimal scalar values for a linear combination of source systems. Positive results were achieved, however, the experiments were performed using a very small test collection (less than 50MB). Similar work was done more recently using web search engines, achieving moderate success (Alaoui-Mounir, Goharian, Mahoney, Salem, & Frieder, 1998). A highly detailed study performed by Vogt did extensive examination of using linear combination fusion for a very large number of component systems and reached results that seemed to agree with Lee's overlap correlation (Vogt & Cottrell, 1998, 1999). Similar results were obtained using a relational approach to the IR problem in (McCabe et al., 1999).

Recently, research in data fusion has led to the development of models that depart from the more traditional CombMNZ approach. As collections with document link information (such as the world wide web) have become more prevalent, there has been some research attempting to take advantage of any evidence of relevance that might be present in document link structures and use it to improve retrieval effectiveness. Picard developed a fusion model for this problem and performed some preliminary experiments using the CACM collection in (Picard, 1998). While these results were initially promising, it is difficult to know how such a technique would perform using large modern collections. Picard was limited to the CACM because at the time it was the only collection containing link data between documents.

Manmatha, et al., developed an unsupervised probabilistic method for combining search results from separate information retrieval systems (Manmatha, Rath, & Feng, 2001). Their hypothesis is that all "good" text search engines will exhibit similar characteristics, and their model is based on using available relevance judgments as training data to model the score distributions of relevant and non-relevant documents on a per-query basis. Models are created for each component search engine, and a mixture model is generated that uses evidence from each component model to try and obtain optimal results. For a query with no training data, this mixture model produces a probability of relevance based on relevance scores for the results of that query from the component search engines. Reasonable results were obtained from preliminary experiments that fused the top result sets from different systems at TREC-3, although it should be noted that entirely separate systems were used for fusion in these experiments, therefore systemic differences were not accounted for.

Some recent work (Soboroff, Nicholas, & Cahan, 2001) 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 *systems* can retrieve different relevant documents, although there is no discussion of what factor in a retrieval system is most likely causing these different relevant documents to be retrieved.

Chowdhury, et al. (A. Chowdhury, O.Frieder, Grossman, & McCabe, 2001) began an investigation of fusing highly effective retrieval strategies, while keeping everything else constant in the system. While their data was limited, they formed initial conclusions suggesting that fusion of highly effective strategies does not tend to improve effectiveness. This work was continued by Beitzel and colleagues (Beitzel et al., 2003), who tried to investigate and identify the conditions that are required for fusion to show an improvement when highly effective strategies are involved. They concluded that for fusion of highly effective strategies to improve effectiveness there must be a significant number of unique relevant documents merged into the fused set. This study shows motivation for further work in determining exactly what takes place when highly effective retrieval strategies are fused; we must determine if observed improvements are due to the effectiveness of the retrieval strategies alone or if they are due to the variation of systemic differences between the component result sets.

Recent work by Montague, et al., provides experimentation performed under similar conditions to Lee's work, and shows similar results (Montague & Aslam, 2001, 2002). 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.

In summary, there exists a very large body of research in the area of data fusion. In spite of this, the precise reasons and conditions under which data fusion will help to improve retrieval have not been precisely specified. Lee comes closest to identifying a possible indicator for when fusion is a worthwhile approach, however, there is a lack of research exploring the specific case of fusing results from highly-effective document retrieval strategies while holding systemic differences constant. The remainder of this study focuses on this very problem, and examines the accuracy of Lee's overlap hypothesis as an indicator of fusion performance under these circumstances.

Motivations & Methodology

The focus of this study is to examine in detail what conditions are required in order to show improvement when using data fusion. Having reviewed the evolution of the prior work in the area, it is clear that the most logical starting place for this analysis is to examine Lee's study on fusing retrieval strategies, and the validity of his hypothesis about the correlation between the difference in relevant and non-relevant overlap, and improvements due to fusion. To review, Lee's experiments proceed under the assumption that as long as the component result sets involved in fusion have greater relevant overlap than non-relevant overlap, there will be an improvement. To justify this, the experiments used a series of result sets that had low general overlap (15%), 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). The study contained no analysis of the relative effectiveness of fusion when using random sets versus using the most effective available sets, no comparison between the effectiveness of the fused results and the effectiveness of the best system at TREC-3, and most importantly, no evaluation of fusion's effectiveness when varying only retrieval strategies in the same information retrieval system. Performing fusion of entirely different systems or of result sets that are generated using highly different retrieval utilities (different parsers, stemmers, phrase lists, stopwords, relevance feedback methods, etc.) introduces more than one independent variable and makes it very difficult to derive sound conclusions about the effects of different retrieval strategies from the data. To truly study the effects of fusing retrieval strategies alone, systemic differences must be held constant. Lee's study did not analyze the effect of varied systemic differences on fusion's effectiveness. Given these points, it is difficult to generalize based on these 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.

Our goal is to discover if retrieval strategies alone are responsible for the effectiveness improvements observed from data fusion. Furthermore, we wish to target this examination towards the fusion of modern, highly effective retrieval strategies. To analyze this problem, we must identify the cases where fusion techniques are able to provide improvements in retrieval effectiveness. Then, the likelihood of these conditions occurring when fusion of highly effective retrieval strategies is performed while all other factors are held constant must be examined. As stated above, data fusion techniques can improve retrieval in two ways. First, voting can be employed in order to boost the rank of documents that are common amongst component result sets. This point of benefit makes clear the source of Lee's statements regarding overlap. If the percentage of relevant overlap is significantly higher than the percentage of non-relevant overlap, the voting mechanisms should be more likely to boost the ranks of relevant documents, thereby improving retrieval effectiveness. However, when considering the case of highly effective retrieval strategies, we believe that voting is actually far more likely to hurt retrieval effectiveness (Montague & Aslam, 2002). The reasoning for this lies in the fact that, because the component strategies are known to be highly effective, it is fair to assume that the ranking they provide for their results is already of fairly high quality (i.e., relevant documents are likely to already be ranked higher than non-relevant documents). Given this, voting is more likely to boost a common non-relevant document to a higher rank than a common relevant document. If this occurs enough times, any improvements gained from boosting relevant documents may be cancelled out, and retrieval effectiveness may even be degraded. This leads us to establish the first part of our two-part hypothesis: when fusing highly effective retrieval strategies, the voting properties of multiple-evidence techniques such as CombMNZ will not improve effectiveness.

The second way that fusion techniques like CombMNZ can positively affect retrieval is if they are

able to merge relevant documents that are unique to a single component system into the final fused result set. This increases recall, and may increase average precision if the new relevant documents are inserted into the fused result set at high enough ranks, thereby bringing improvements to retrieval effectiveness. A caveat of this is that when the component result sets have a high degree of relevant overlap, the likelihood of merging in unique relevant documents, especially at high ranks, will tend to be very small. In fact, when all other systemic differences are held constant, the retrieval strategies being used for fusion are going to be using the same terms to produce their ranked sets. Given that they are all highly effective, the strategies are most likely going to return highly similar document sets in this case, and the only differences are likely to be in the *ranking* of each set, rather than in the content of the sets themselves, except in cases of some documents "falling off the end" of the result set, as TREC result sets are truncated at 1000 documents retrieved. This leads to the second part of our hypothesis, which states that highly effective retrieval strategies tend to retrieve the same relevant documents, and therefore it is very unlikely that unique relevant documents will be merged into the final result set, and effectiveness will not be improved. When both points of our hypothesis points are taken together, they illuminate an important fact about data fusion of highly effective strategies: if there are no observable effectiveness improvements when all systemic differences are held constant and only retrieval strategies are varied, any improvements observed from data fusion techniques using CombMNZ cannot be due to retrieval strategies; rather, they must be due to the effect of one or more systemic differences. We hypothesize that Lee's overlap correlation is in fact a poor indicator of the performance benefits available from data fusion when highly effective result sets are used; in fact, we hypothesize that any observed improvements are due to an increase in the recall of highly-ranked unique relevant documents in the fused result set.

To prove our hypothesis we designed experiments that measure the effectiveness of both the voting and merging properties of data fusion using CombMNZ. CombMNZ was chosen because it has repeatedly been shown to outperform the other combination variants developed by Fox & Shaw, and it is designed to incorporate both voting and merging. In addition, we measured the difference of relevant and non-relevant overlap across the highly-effective component result sets used in our experiments, as Lee did. In our experiments, we will show that neither voting nor merging is bringing improvement when fusing highly effective strategies in the same system and that having greater relevant overlap than non-relevant overlap does not necessarily correspond to an automatic benefit from fusion. Additionally, we show that improvements only occur when enough relevant documents appearing in only one component result set are given high rank in the fused result set.

Results

Our experimental goals are clear: we must show that both possibly beneficial properties of fusion will not improve effectiveness when fusing highly effective retrieval strategies in the same information retrieval system. To do this, experiments must be conducted over a large number of queries, and under a controlled environment that holds constant any systemic differences. Furthermore, we had to ensure that the retrieval strategies being fused were highly effective. To achieve these goals, we implemented three modern retrieval strategies that have recently been shown to be highly effective, one Vector-Space and two Probabilistic: IIT (Abdur Chowdhury et al., 2000), BM25 (Robertson, Walker, Beaulieu, Gatford, & Payne, 1995), Self-Relevance (Kwok, Grunfeld, Chan, Dinstl, & Cool, 1998). It should be noted that there are other approaches to document retrieval such as Language Modeling and Inference Networks that are not represented by the set of retrieval strategies selected for these experiments. As stated above, this should not affect the experimental outcome, since any highly effective retrieval strategies are likely to retrieve highly similar sets of documents.

A single information retrieval engine was used with each of the selected retrieval strategies to evaluate all query topics from the ad-hoc track at TREC 6, 7, and 8, and also all query topics from ad-hoc task of the web track at TREC-9 and TREC-10. We carefully ensured that the component result sets were all generated with the same set of systemic properties (parsers, stemmers, phrase lists, stopwords, etc.), and that the only variable parameter in the component result sets was the retrieval strategy used. Relevance Feedback was not used for any of these experiments. We then fused the component result sets and evaluated the fused set for improvements in retrieval effectiveness. Finally, we performed an overlap analysis for each group of component sets to facilitate examination of Lee's overlap correlation. Once this was complete, we used several techniques to examine each

beneficial feature of data fusion to see how it was affecting effectiveness.

The first set of experiments we ran were the fusion of each retrieval strategy in the same system making use of the title-only topic descriptions. These results are shown in Table 1. As stated above, all systemic differences remained constant. All measurements are Mean Average Precision unless otherwise specified.

Strategy	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
IIT	0.1900	0.1718	0.2190	0.1778	0.1704
BM25	0.1948	0.1770	0.2186	0.1847	0.1949
Self-Relevance	0.1708	0.1558	0.2065	0.1560	0.1639
Best	0.1948	0.1770	0.2190	0.1847	0.1949
Fused	0.1911	0.1751	0.2168	0.1671	0.1935
Improve/Best	-1.90%	-1.07%	-1.00%	-9.53%	-0.72%

Table 1: Improvement of Same-System Fused Retrieval Strategies - Title-Only, No Relevance Feedback

We also performed an overlap analysis on these runs, which is given in Table 2.

	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
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%

Table 2: Overlap of Same-System Retrieval Strategies

For our second set of experiments, we fused the top three result sets from each TREC-year. This was done in order to compare the effectiveness improvements gained from fusing highly-effective strategies in the same system to the improvements gained from fusing separate, highly effective systems. These results are given in Table 3, and their overlap analysis is given in Table 4. Note that, while we were careful to ensure that relevance feedback was not used in the creation of any of our result sets, it is possible (and likely) that these results, over separate TREC systems, did make use of it, and other precision-enhancing utilities as part of their systemic differences. This accounts for the difference in the ranges of scores reported in Table 1 and Table 3.

	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
System 1	0.2876	0.2614	0.3063	0.2011	0.2226
System 2	0.2556	0.2488	0.2876	0.1970	0.2105
System 3	0.2481	0.2427	0.2853	0.1812	0.2084
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 3: Improvement of Best TREC Systems

	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
Overlap	34.43%	39.31%	42.49%	30.09%	33.75%
R Overlap	83.08%	80.84%	84.63%	85.85%	81.87%
NR Overlap	53.33%	56.36%	57.13%	51.26%	54.01%
% Diff R/NR	55.78%	43.44%	48.14%	67.48%	51.58%

Table 4: Overlap of Best TREC Systems

From Table 1 and Table 2 we can see that the fused strategies never outperform the best single strategy, rather, fusion appears to hurt retrieval effectiveness (although decrease in performance over the best is likely an artifact of CombMNZ). Furthermore, Table 1 and Table 2 illustrate that improvements are not guaranteed when fusing effective strategies in the same system, even when relevant overlap is greater than non-relevant overlap. This indicates that Lee's overlap correlation is a poor indicator of the likelihood of fusion to improve effectiveness. When examining the more complex case of fusing different retrieval systems, slight improvement is observed, however Table 3 and Table 4 show that improvements from fusion do not generally increase as relevant overlap becomes increasingly greater than non-relevant overlap. This makes it difficult to find a threshold of difference in Relevant and Non-Relevant overlap past which fusion is likely to improve effectiveness. This also shows that the overlap correlation is a poor indicator of fusion's expected performance.

Given that the overlap correlation appears to be a poor indicator of potential effectiveness improvements to be had from fusion, we developed several methods to test the effect of the voting and merging properties of the CombMNZ algorithm. This was done in order to try and obtain a clear idea of exactly what is happening to the result sets when they are being fused, and to isolate conditions under which fusion may be helpful. To begin, we addressed our assumption that highly effective retrieval strategies will produce similar, high quality rankings, therefore minimizing the possible positive effects of boosting on agreement. To show similarity we calculated the Spearman Rank Correlation (Everitt, 2002) between each pair of ranked

component sets used in our fusion experiments. These results are shown for highly effective, same-system strategies in Table 5 and for the best TREC systems in Table 6.

Pair	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
IIT/BM25	0.8808	0.8725	0.8934	0.8357	0.7425
IIT/Self-Relevance	0.6392	0.7128	0.6261	0.7263	0.6117
BM25/Self-Relevance	0.5921	0.6172	0.5673	0.6239	0.52

Table 5: Spearman Rank Correlations for Highly Effective Strategies in the Same System

Pair	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
S1/S2	0.4182	0.4918	0.6053	0.4759	0.4647
S1/S3	0.4148	0.4887	0.5848	0.4091	0.4636
S2/S3	0.4735	0.5011	0.5401	0.3674	0.4501

Table 6: Spearman Rank Correlations for the best TREC Systems

Table 5 shows that there is a positive Spearman correlation between all ranking pairs of highly effective strategies fused in the same system. In most cases, these correlations would be considered "moderate" to "strong" positive correlations, which show that the rankings are highly similar. By contrast, Table 6 shows that the correlations for the best TREC systems are not as strong, therefore they have lower agreement on ranking. Referring back to Table 1 and Table 3, it can be seen that fusion of the best TREC systems shows greater effectiveness improvements over the best single system than the fusion of highly effective strategies in the same system. This favors our hypothesis that when the rankings are very similar, any positive effects due to boosting are minimized.

To further illustrate that the boosting qualities of CombMNZ are likely to have negative effects, we calculated the average change in rank of relevant and non-relevant documents from their original position in the component sets to their final position in the fused set. Documents that failed to appear in one or more component sets were assumed to be the lowest ranked documents in those sets. We call this value the *Rank Displacement Coefficient*. Clearly, the most optimal behavior for best fusion results would be for relevant documents to have a high positive coefficient, meaning they moved far up in rank in the final fused set, and for non-relevant documents to have a low or negative coefficient, meaning that they didn't move much, or even moved down in rank. The behavior of the Rank Displacement Coefficient is displayed in Figure 1. In this example, the values of the Rank Displacement Coefficient are -2 for Relevant (one relevant document dropped two spots in rank) and +1 for Non-Relevant (two non-relevant documents rose one spot each in rank). These are clearly undesirable values. The Rank Displacement Coefficients for the relevant and non-relevant documents from same-system and best-TREC fusion experiments are shown in Table 7 and Table 8.

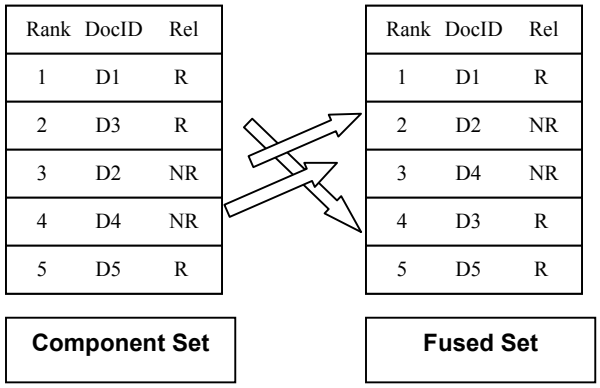

Figure 1: Rank Displacement Coefficient

Type	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
Relevant	1.078	0.959	1.174	0.655	1.833
Non-Relevant	22.038	20.498	19.433	27.771	37.11

Table 7: Rank Displacement Coefficients for Highly Effective Strategies in the Same System

Type	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
Relevant	2.777	2.656	1.969	2.176	3.120
Non-Relevant	47.306	32.953	31.925	61.532	74.624

Table 8: Rank Displacement Coefficients for the best TREC Systems

Table 7 shows us that for fusing in the same system, the Rank Displacement Coefficients are not friendly to fusion. Non-relevant documents have a very high positive coefficient, meaning that they tended to move up in rank by a large degree, while relevant documents had very small coefficients, meaning that on average, their ranks did not change very much. Table 8 shows that when fusing the best TREC systems the rank displacements were even more volatile than when fusing in the same system, with the displacements being nearly twice as large. However, when normalizing for the difference in magnitude, the relevant rank displacement is slightly higher in the case of the best TREC systems, which is to be expected, given that improvement due to

fusion is higher for the best TREC systems. From these results, it is clear that voting is highly detrimental to fusion in the case of fusing highly effective retrieval strategies in the same system. We believe this is because highly effective strategies produce similar, effective rankings to begin with, and the voting techniques employed by CombMNZ are more likely to hurt effectiveness than help it.

These first experiments have shown that the voting properties of CombMNZ are likely to be detrimental to effectiveness when fusing highly effective retrieval strategies in the same system. To fully explore the conditions under which fusion may help effectiveness, we also designed a series of experiments for investigating the effect of merging as well. We have hypothesized that merging is highly unlikely to help improve effectiveness because highly effective retrieval strategies have high relevant overlap, and have high-quality rankings, therefore the probability of merging in relevant documents, particularly at high ranks where they would be very beneficial, is very small. To illustrate this, we examined documents that were not in the intersection of all component result sets used for fusion, and calculated the portion of these that were relevant and the portion that were non-relevant. This is shown in Table 9 for highly effective strategies in the same system, and in Table 10 for the best TREC systems.

Type	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
Relevant	585	560	710	351	779
Non-Relevant	38518	38266	39905	39444	43737
% Relevant	.015%	.014%	.017%	.009%	.017%

Table 9: Relevant and Non-Relevant Documents outside the intersection for Highly Effective Strategies in the Same System

Type	Trec-6	Trec-7	Trec-8	Trec-9	Trec-10
Relevant	1320	1214	1113	979	1403
Non-Relevant	76124	68676	66030	84231	77496
% Relevant	.017%	.017%	.017%	.011%	.018%

Table 10: Relevant and Non-Relevant Documents outside the intersection for the best TREC Systems

From Table 9 and Table 10 it can clearly be seen that the probability of a document outside the intersection being relevant is smaller when fusing highly effective retrieval strategies in the same system. This is in favor of our hypothesis that merging will not, on average, be able to improve retrieval effectiveness when the result sets being merged already have high overlap and high quality. These results, taken together with the above discussion on the effects of voting, provide strong evidence that the voting and merging properties of CombMNZ will not bring improvement when fusing highly effective strategies in the same system.

Finally, we also wished to examine the reasons why fusing separate but highly effective retrieval systems, such as the best systems from TREC, usually results in some improvement. We hypothesized that improvement is most likely to be achieved when new relevant documents are merged into the final result set at high rank, or, more formally, that an increase in recall of relevant documents for which there was no agreement across component result sets, and the placement of these "unique" relevant documents at high ranks in the fused set plays a factor in bringing effectiveness improvements. 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 for both the best TREC systems and the highly effective 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 "unique" 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 highly effective strategies in the same information retrieval system.

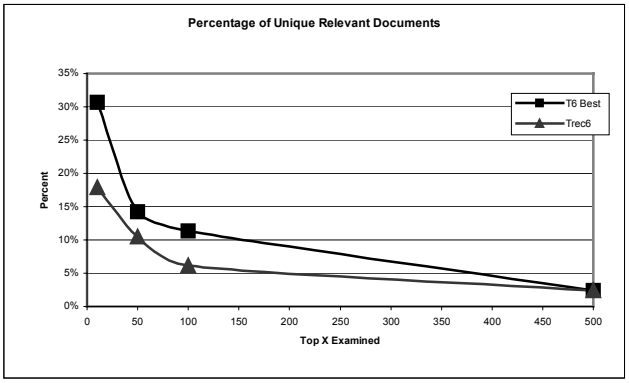

Figure 2: TREC-6 Unique Relevant Document Analysis

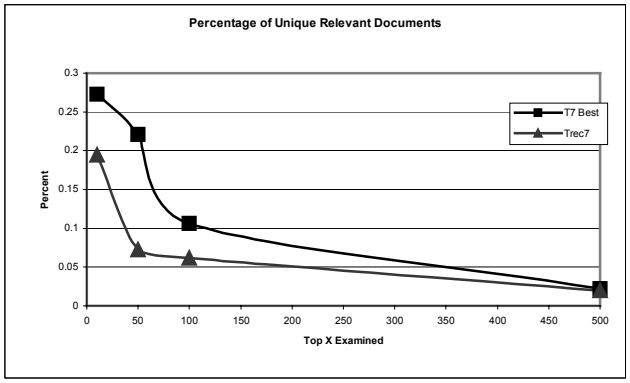

Figure 3: TREC-7 Unique Relevant Document Analysis

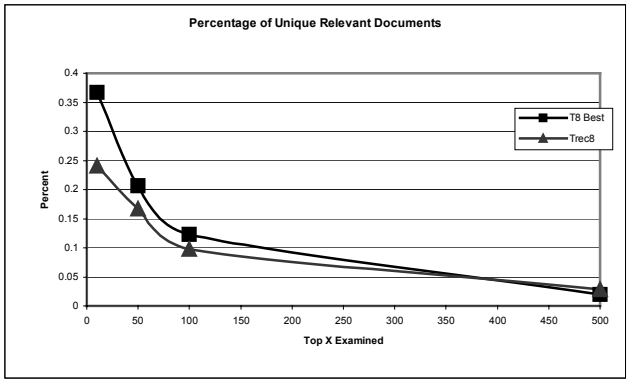

Figure 4: TREC-8 Unique Relevant Document Analysis

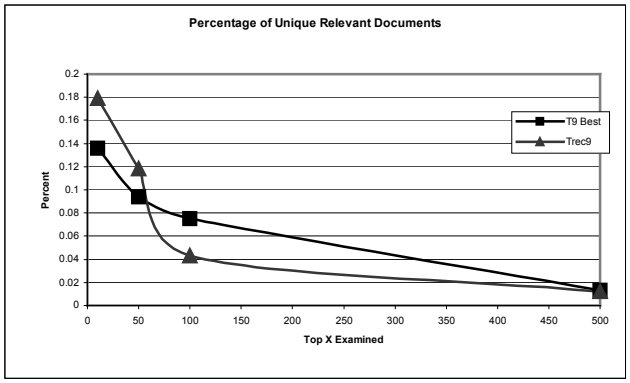

Figure 5: TREC-9 Unique Relevant Document Analysis

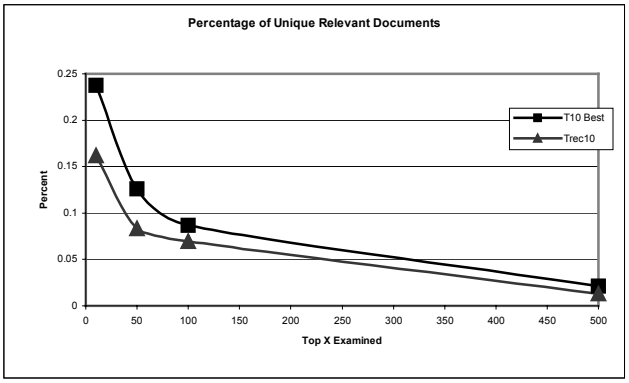

Figure 6: TREC10 Unique Relevant Document Analysis

These graphs above clearly show that for each TREC year, the fused set from the top three TREC systems contains a higher percentage of these no-agreement "unique" relevant documents in its final result set for any given depth X (recall that X is the depth to which component result sets are examined for "unique" relevant documents). 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 these "unique" relevant documents are being inserted at high ranks in the fused result set. 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. Note that there is one outlier in the analysis, located in Figure 5. We speculate that this outlier is due to the fact that Figure 5 represents data from TREC-9, which was the first year that the standard Ad-hoc track on SGML news data was discontinued in favor of an Ad-hock task on Web data. Because of this, it is likely that competing systems, heavily tuned on the news data from the previous years, produced inconsistent results. We note that by the following year, (Figure 6) also performed on web data, these anomalies would seem to have been corrected, as the outlier is not present.

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 the topmost 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.

Depth	Relevant	Non-Relevant	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 11: Avg. # Unique R & NR added in same-system fusion

Depth	Relevant	Non-Relevant	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

Table 12: Avg. # Unique R & NR added in TREC-best fusion

It can be seen from Table 11 and Table 12 that 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%.

In summary, these experiments have shown that in general, improvements cannot be expected when fusing highly effective retrieval strategies while holding systemic differences such as parsers, stemmers, and relevance feedback algorithms constant. We have shown that the high initial quality of the rankings in component result sets coupled with the high degree of overlap makes it very difficult for voting/merging fusion techniques such as CombMNZ to improve effectiveness. Finally, we have shown that, when fusing high quality result sets from different systems, such as the best systems from TREC, improvements are most likely to be seen when relevant documents having minimal agreement (i.e., appear in only one component set) are merged into the fused set at a high rank.

Conclusions & Future Work

In this paper we have thoroughly explored the long believed precept that fusing different, highly effective retrieval strategies in the same system is a reliable and successful method of improving retrieval effectiveness. We have found that in fact, this is not the case. Through a series of comprehensive experiments, we have illustrated exactly what happens when result sets from highly effective strategies are fused, and have identified the reasons and conditions under which significant improvements are not likely to be observed – most notably, we have shown that fusing highly effective retrieval strategies does not guarantee effectiveness improvements, and that the difference between relevant and non-relevant overlap of component result sets is a poor indicator of the effectiveness of fusion. In fact, we have shown that improvements are never observed when highly effective strategies are fused and all other factors are held constant. Additionally, we have attempted to improve on some of the prior work in this area by taking great care with the control of our test environment to ensure that choice of retrieval strategy was the only independent variable in our experiments, thereby ensuring the sound nature of our conclusions. For future work we intend to explore alternative methods of fusion that may be able to take advantage of some of the flaws in CombMNZ, in particular its naïve approach to boosting on agreement. We would like to investigate a more effective way of modeling the relationship between agreement across result sets, local rank of a document in its component result set, and probability of relevance. Additionally, now that we have shown effectiveness cannot be improved by fusing highly effective retrieval strategies, we would like to examine other system-dependant factors of the test environment such as parsing, relevance feedback, and phrase processing to determine if one of these techniques may benefit fusion. This may include an extension of the experiments with relevance feedback fusion that Lee began in (Lee, 1998).

References

Alaoui-Mounir, S., Goharian, N., Mahoney, M., Salem, A., & Frieder, O. (1998). *Fusion of Information Retrieval Engines (FIRE).* Paper presented at the International Conference on Parallel and Distributed Processing Technologies and Applications (PDPTA-1998), Las Vegas, NV.

Aslam, J., & Montague, M. (2001). *Models for Metasearch.* Paper presented at the 24th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-2001), New Orleans, LA.

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1994). *Automatic Combination of Multiple Ranked Retrieval Systems.* Paper presented at the 17th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-1994).

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., Frieder, O., & Goharian, N. (2003, March 9-11). *Disproving the Fusion Hypothesis: An Analysis of Data Fusion via Effective Information Retrieval Strategies.* Paper presented at the 18th Annual ACM Symposium on Applied Computing (SAC-2003), Melbourne, FL.

Belkin, N. J., Cool, C., Croft, W. B., & Callan, J. P. (1993). *The Effect of Multiple Query Representations on Information Retrieval Performance.* Paper presented at the 16th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-1993).

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management, 31*(3), 431-448.

Chowdhury, A., Beitzel, S. M., Jensen, E. C., Saelee, M., Grossman, D., Frieder, O., et al. (2000, November 13-16). *IIT TREC-9 - Entity-Based Feedback with Fusion.* Paper presented at the 9th Annual Text Retrieval Conference (TREC-9), National Institute of Standards and Technology, Gaithersburg, MD.

Chowdhury, A., O.Frieder, Grossman, D., & McCabe, M. C. (2001, September 9-12). *Analyses of Multiple-Evidence Combinations for Retrieval Strategies.* Paper presented at the 24th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-2001), New Orleans, LA.

Everitt, B. S. (2002). *The Cambridge Dictionary of Statistics* (2 ed.): Cambridge University Press.

Foltz, P. W., & Dumais, S. T. (1992). Personalized Information Delivery: An analysis of information-filtering methods. *Communications of the ACM, 35*(12).

Fox, E. A., & Shaw, J. A. (1994). *Combination of Multiple Searches.* Paper presented at the 2nd Annual Text Retrieval Conference (TREC-2), NIST, Gaithersburg, MD.

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

Kwok, K. L., Grunfeld, L., Chan, M., Dinstl, N., & Cool, C. (1998). *TREC-7 Ad-Hoc, High Precision and Filtering Experiments using PIRCS.* Paper presented at the 7th Annual Text Retrieval Conference (TREC-7), NIST, Gaithersburg, MD.

Lee, J. H. (1995). *Combining Multiple Evidence from Different Properties of Weighting Schemes.* Paper presented at the 18th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-1995).

Lee, J. H. (1997). *Analyses of Multiple Evidence Combination.* Paper presented at the 20th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-1997).

Lee, J. H. (1998). Combining the Evidence of Different Relevance Feedback Methods for Information Retrieval. *Information Processing and Management, 34*(6), 681-691.

Manmatha, R., Rath, T., & Feng, F. (2001). *Modeling Score Distributions for Combining the Outputs of Search Engines.* Paper presented at the 24th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-2001), New Orleans, LA.

McCabe, M. C., Chowdhury, A., Grossman, D., & Frieder, O. (1999, November). *A Unified Environment for Fusion of Information Retrieval Approaches.* Paper presented at the 8th Annual ACM Conference on Information and Knowledge Management (CIKM-1999).

McGill, M., Koll, M., & Norreault, T. (1979). *An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems*. Syracuse: Syracuse University, School of Information Studies.

Montague, M., & Aslam, J. (2001). *Relevance Score Normalization for Metasearch.* Paper presented at the 10th Annual ACM Conference on Information and Knowledge Management (CIKM-2001).

Montague, M., & Aslam, J. (2002, November). *Condorcet Fusion for Improved Retrieval.* Paper presented at the 11th Annual ACM Conference on Information and Knowledge Management (CIKM-2002), Tyson's Corner, VA.

Picard, J. (1998). *Modeling and combining evidence provided by document relationships using probabilistic argumentation systems.* Paper

presented at the 21st Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-1998), Melbourne, Australia.

Robertson, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation, 33*(4).

Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., & Payne, A. (1995). *Okapi at TREC-4.* Paper presented at the 4th Annual Text Retrieval Conference (TREC-4), NIST, Gaithersburg, MD.

Saracevic, T., & Kantor, P. (1988). A Study of Information Seeking and Retrieving, III: Searchers, Searches, Overlap. *Journal of the American Society of Information Science, 39*(3).

Soboroff, I., Nicholas, C., & Cahan, P. (2001). *Ranking Retrieval Systems without Relevance Judgments.* Paper presented at the 24th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-2001), New Orleans, LA.

Turtle, H., & Croft, W. B. (1991). Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions on Information Systems, 9*(3), 187-222.

Vogt, C. C., & Cottrell, G. W. (1998). *Predicting the Performance of Linearly Combined IR Systems.* Paper presented at the 21st Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-1998), Melbourne, Australia.

Vogt, C. C., & Cottrell, G. W. (1999). Fusion via a Linear Combination of Scores. *Information Retrieval, 1*(3), 151-173.

Voorhees, E. M., Gupta, N. K., & Johnson-Laird, B. (1994). *The Collection Fusion Problem.* Paper presented at the 3rd Annual Text Retrieval Conference (TREC-3), NIST, Gaithersburg, MD.