Recent Results on Fusion of Effective Retrieval Strategies in the Same Information Retrieval System[1]

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury
David Grossman, Nazli Goharian, Ophir Frieder
Information Retrieval Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616

{steve,ej,abdur,dagr,nazli,ophir}@ir.iit.edu

ABSTRACT

Prior efforts have shown that data fusion techniques can be used to improve retrieval effectiveness under certain situations. Although the precise conditions necessary for fusion to improve retrieval have not been identified, it is widely believed that as long as component result sets used in fusion have higher relevant overlap than non-relevant overlap, improvements due to fusion can be observed. We show that this is not the case when systemic differences are held constant and different highly effective document retrieval strategies are fused within the same information retrieval system. Furthermore, our experiments have shown that the ratio of relevant to non-relevant overlap is a poor indicator of the likelihood of fusion's effectiveness, and we propose an alternate hypothesis of what needs to happen in order for fusion to improve retrieval when standard voting/merging algorithms such as CombMNZ are employed.

1 INTRODUCTION

Recently there has been much research done in the field of information retrieval concerning the various kinds of "fusion" and their applications. 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 typically utilized to improve retrieval effectiveness, and is most often applied to the task of ad-hoc retrieval. Data fusion techniques also have applications outside the realm of ad hoc retrieval, having a relevant place in the worlds of metasearch, and distributed information retrieval.

This paper summarizes some of our efforts to examine 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. This belief is supported by the

[1] The authors would like to note that a greatly expanded version of this work was published in the 2003 ACM Symposium on Applied Computing (ACM-SAC). Interested readers are invited to refer to [Beit03] for further details

supposition that different document retrieval strategies will rank documents differently, returning different sets of relevant and non-relevant documents. Several popular result combination algorithms, such as CombSUM and CombMNZ [Fox94] have been invented to take advantage of this property, using a combination of voting and merging to fuse results from several sources into one unified set. Although much research has been conducted, precise analysis of why and when data fusion techniques improve retrieval has not yet been undertaken. This may be because Lee's overlap correlation [Lee97] is generally accepted to be true. It states that fusion will generally improve effectiveness as long as the relevant overlap between component result sets is greater than non-relevant overlap. Because of this, researchers are more focused on using fusion as a utility to improve their systems than on discovering the details of what actually makes fusion a worthwhile endeavor.

We have examined the case of using data fusion techniques to fuse result sets created by different, highly effective modern retrieval strategies. A key difference between our approach and prior approaches is that we performed our experiments in a completely controlled environment; only retrieval strategy was varied across component result sets. All other systemic differences such as parsing rules, stemming rules, relevance feedback techniques, stopword lists, etc, were held constant. This approach has allowed us to examine if different retrieval strategies are actually returning different sets of documents, as generally believed. Our preliminary experiments have shown that in actuality, highly effective retrieval strategies tend to return result sets with a high degree of general overlap (ie, sets containing the same documents). In addition, we have found that Lee's overlap correlation does not hold true when highly effective strategies are used and systemic differences are held constant. When we examined the problem further, we found that in the cases where fusion is improving, it is not due to the agreement of several systems on what documents are relevant, but rather, it is due to the increase in recall of relevant documents that only appear in one component result set, and the insertion of these unique relevant documents into the final fused result set at a position of high rank.

The remainder of this paper will give a brief overview of prior work, followed by a description of our preliminary experiments and a discussion of our results. We close with a brief discussion of future work in this area, and provide references to more detailed analyses for the interested reader.

2 PRIOR WORK

There exists a very large body of prior work in the area of data fusion. Many different types of evidence of relevance 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. A variety of different techniques for utilizing this information in a data fusion strategy can be found in the literature, although the most common are Fox & Shaw's CombSUM and CombMNZ measures [Fox94]. 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.

One of the earliest studies in data fusion was performed by Belkin and colleagues [Belk93, Belk95]. They investigated the effect of fusing results from different query representations 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 data fusion. 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 [Lee95]. He found that when performing combinations in this matter, significant improvements could be achieved. 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 [Lee97]. 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 component result sets $S_1 ... S_n$ 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 correlation. 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 [Bart94]. Positive results were achieved, however, the experiments were performed using a very small test collection (less than 50MB). In addition, many others have experimented in this area and observed results that seem to agree with Lee's overlap correlation.

Given that results exist which show data fusion to be effective, 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. This question is what led us to examine the data fusion problem in greater detail.

3 METHODOLOGY

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.

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. The reasoning for this lies in the fact that, because the component strategies are 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 CombMNZ-like fusion techniques 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. 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, the goal of this work becomes clear: if there are no improvements to effectiveness when all systemic differences are held constant and only retrieval strategies are varied, any improvements observed from data fusion *cannot* be due to the retrieval strategies.

To prove our hypothesis we designed many experiments that measure the effectiveness of both the voting and merging properties of data fusion using CombMNZ. If it can be shown that neither beneficial property of fusion is bringing improvement when holding constant all systemic differences and only varying retrieval strategies, then we have proved our hypothesis.

4 RESULTS

For our experiments, we implemented three modern retrieval strategies that were recently shown to be highly effective in the TREC forum, one Vector-Space and two Probabilistic (IIT [Chow00], BM25 [Robe95], Self-Relevance [Kwok98]). 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 experiments were designed to determine the validity of Lee's overlap correlation. It dictates that as long as there is a difference in relevant and non-relevant overlap, fusion will likely improve effectiveness. To examine this, we first used CombMNZ to fuse the results of each of our three highly effective retrieval strategies inside the same information retrieval system, and compared the effectiveness of this fused result set to the effectiveness of the best-performing single retrieval strategy out of the three. We illustrate this with average precision values in Table 1.

Table 1: Fusion of Effective Retrieval Strategies in the Same System

	Trec6	Trec7	Trec8	Trec9	Trec10
Best Strategy	0.1948	0.1770	0.2190	0.1847	0.1949
Fused Results	0.1911	0.1751	0.2168	0.1671	0.1935
% Imp. of Fused over 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 Analysis for Same-System Fusion

	Trec6	Trec7	Trec8	Trec9	Trec10
Overlap	62.76%	61.14%	59.42%	61.61%	59.17%
Rel Overlap	89.52%	89.90%	90.23%	88.61%	85.88%
NRel 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 testing the overlap correlation involves fusing the three best result sets from distinct TREC competitors for all years with title-only results available. A key difference in these experiments is that these result sets were all generated by separate information retrieval systems – they were not guaranteed to have used the same parsing rules, stemming rules, relevance feedback algorithms, etc, so it is clear that more is being varied here than simply retrieval strategy. The average precision values for improvement, and the overlap analysis are shown in Table 3 and Table 4.

Table 3: Fusion of Best-Performing TREC Systems

	Trec6	Trec7	Trec8	Trec9	Trec10
Best TREC System	0.2876	0.2614	0.3063	0.2011	0.2226
Fused Results	0.3102	0.2732	0.3152	0.2258	0.2441
% Imp. Of Fused over best	7.86%	4.51%	2.91%	12.28%	9.66%

Table 4: Overlap Analysis for Best-TREC fusion

	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 (the TREC systems), we do see small to moderate improvements with fusion, however, if the overlap correlation were true, and if our resilts were to be consistent with those found by Lee, our effectiveness improvements should have been more substantial, and generally increased as the difference in relevant and non-relevant overlap increased. This is clearly not the case, as can be seen from Tables 1-4 above. Generally, overlap is lower (30-43% - see Table 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 test our hypothesis, we examined 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.

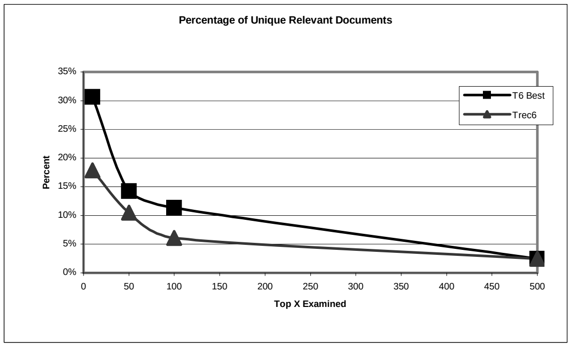

Figure 1: TREC-6

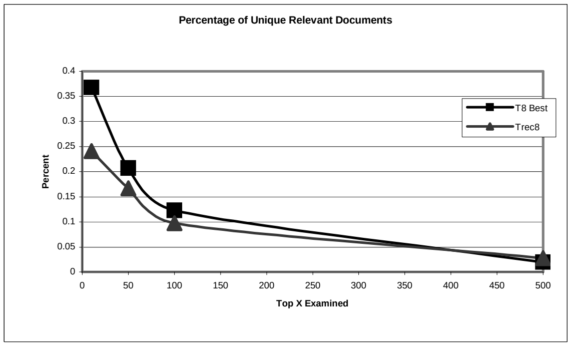

Figure 3: TREC-8

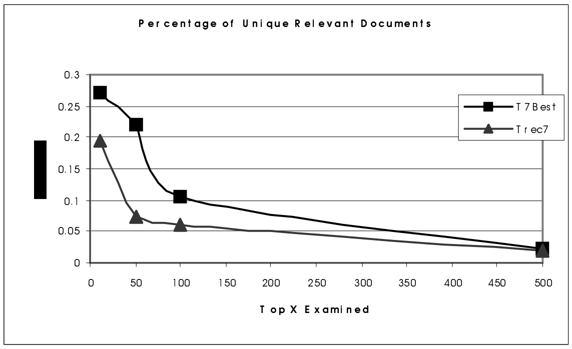

Figure 2: TREC-7

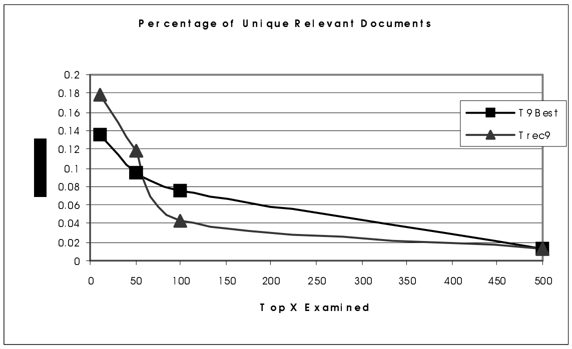

Figure 4: TREC-9

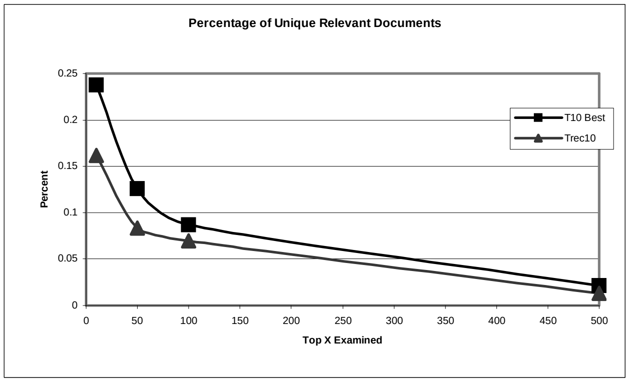

Figure 5: TREC10

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.

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. # of Unique Rel & NRel documents 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. # of Unique Rel & NRel documents 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

[Bart94] 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.

[Beit03] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, O. Frieder, N. Goharian, "Disproving the Fusion Hypothesis: An Analysis of Data Fusion via Effective Information Retrieval Strategies", *Proceedings of the 2003 ACM Symposium on Applied Computing (ACM-SAC)*, Melbourne, FL, March 2003.

[Belk93] 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.

[Belk95] 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.

[Chow00] A. Chowdhury, et al., "Improved query precision using a unified fusion model", Proceedings of the 9th Text Retrieval Conference (TREC-9), 2000.

[Fox94] 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.

[Kwok98] 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.

[Lee95] J.H. Lee, "Combining Multiple Evidence from Different Properties of Weighting Schemes," Proceedings of the 18th Annual ACM-SIGIR, pp. 180-188, 1995.

[Lee97] J.H. Lee, "Analyses of Multiple Evidence Combination," Proceedings of the 20th Annual ACM-SIGIR, pp. 267-276, 1995.

[Robe95] S. Robertson, et al., "Okapi at TREC-4", Proceedings of the 4th Text Retrieval Conference (TREC-4), 1995.