A Unified Environment for Fusion of Information Retrieval Approaches

M. Catherine McCabe
George Mason University
cmccabe@gmu.edu

Abdur Chowdhury
IIT Research Institute
abdur@cs.iit.edu

David A. Grossman
Illinois Institute of Tech
dagr@ir.iit.edu

Ophir Frieder
Illinois Institute of Tech
ophir@cs.iit.edu

ABSTRACT

Prior work has shown that combining results of various retrieval approaches and query representations can improve search effectiveness. Today, many meta-search engines exist which combine the results of various search engines in the hopes of improving overall effectiveness. However, the combination of results from different search engines masks variations in parsers, and other indexing techniques (stemming, stop words, etc.) This makes it difficult to assess the utility of the fusion technique. We have implemented the two most prevalent retrieval strategies: probabilistic and vector space using the same parser and the same relational retrieval engine. First, we identified a model that enables the fusion of an arbitrary number of sources. Next, we tested various linear combinations of these two methods as well as various thresholds for identifying retrieved documents. Our results show some improvement of effectiveness, but they also provide us for a baseline from which we can continue with other retrieval strategies and test the effect of fusing these strategies.

Keywords

Information Retrieval, Fusion, Metasearch, Retrieval, Text

1. INTRODUCTION

Improving the effectiveness of Information Retrieval (IR) systems remains a key challenge. Over the last several years, the overall results from the Text Retrieval Conference (TREC) remain in the range of twenty to thirty percent average precision-recall [Harman98]. Several investigators have explored improving effectiveness by combining the results of different retrieval strategies and different query representations. The hope is that each strategy will retrieve very different sets of relevant documents, and combining the results will yield a better result than any of the individual strategies. This is somewhat intuitive and many meta-search engines on the web have been developed in the hopes of capitalizing on the notion that fusing techniques will result in improved effectiveness. However, it is difficult to assess the effectiveness of meta-search engines because they are not typically run against a standard document collection with known relevance results. Another problem is that implementation variations such as parsing rules can have a profound impact on results. Consider a parser whose stop word lists contains the word *after* and another parser that does not stop this word. A query such as "Find all reviews of the song The Morning After" will result in a very different set of retrieved documents for each of the search engines. Hence, when a meta-search is done with two search engines that have different parser, it is not possible to trace any effects on performance – they could be due to either the fusion technique or the parser.

While there is existing prior work for fusion of collections - where results are brought back from various document collections and fusion techniques are used to integrate the rankings, to our knowledge, there is no work where the fusion of various approaches against one collection is conducted in a pure environment. Our work explores the effectiveness of fusion ranking schemes against one collection, with no variation in parser, stoplist, term-weights, etc. Effectiveness of such techniques could then be used to generate better result sets against each collection for feeding into the collection-fusion task.

We have implemented the two most prevalent retrieval strategies: vector space and probabilistic in a common environment. Although this initial work focuses on only two strategies, our model is flexible and provides for the combination of any number of retrieval strategies. Further work could focus on incorporating other strategies into our existing framework. The key concern here is to identify the impact of combining two retrieval strategies with no variations in query representation or parser/indexing rules as well as to identify the best means of combining these two approaches. We show how each strategy can be implemented in the unchanged relational model – this furthers our prior work on using the relational model to implement relevance feedback and the vector space model. Next, we show how various fusion techniques work with a variety of different fusion parameters while holding constant all individual implementation decisions – i.e. using the exact same parser. We use the entire 2GB TREC7 collection for our experimentation.

Section 2 reviews prior work in the information retrieval fusion. Section 3 describes our framework for IR and demonstrates the implementation of two leading similarity measures. Section 4 explains our approach to fusion, our experimental design and results. Finally conclusions and future work are discussed in section 5.

2. PRIOR WORK

2.1 Initial Fusion of Result Sets

Initial work on fusion was done by Fox, Shaw and Thompson as early as TREC-1 [Fox94]. With TREC2, there was an

CIKM '99 11/99 Kansas City, MO, USA
© 1999 ACM 1-58113-146-1/99/0010...$5.00

opportunity to use knowledge of individual system performance at TREC1 to select systems for merging [Fox94, Thompson90]. It seemed intuitive to take the five good approaches/systems from TREC-1, merge them using some reasonable technique, weight them and run them for TREC-2. In Thompson's work, he applied a weight to the individual results based on their prior performance. Thompson considered each result set an 'expert' and used an approach used previously for combining experts. The idea is that some experts should be considered more applicable than others. Thompson's merged results were not significantly different from simply using the best of the sources. Fox did a straight merge of the results using various combination algorithms. Fox found that combinations of the same types of runs (long and short queries with vector space for instance) did not achieve improvements and sometimes degraded performance. However, he achieved improvement over individual runs when merging different paradigms - vector space versus p-norm Boolean [Shaw95]. Note that Fox developed a few key merging approaches that have been used frequently. These are:

COMBMNZ = COMBSUM * number of runs with document
COMBSUM = sum of the individual measures
COMBMIN = minimum of the individual measures
COMBMAX = maximum of the individual measures
COMBAVG = average of the individual measures

2.1 Characteristics for Fusion

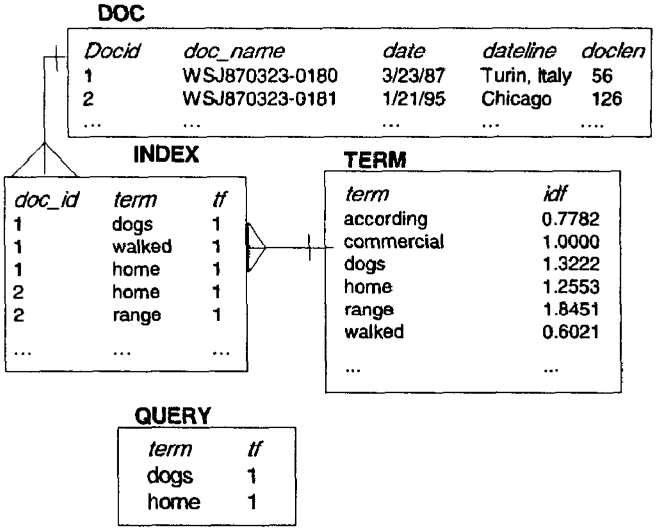
Using several result sets from TREC-3 submissions, Lee began exploring the characteristics of result sets used for fusion [Lee97]. He found that the overlap of the result sets was an important factor in the effect of fusion. The overlap is computed as the Relevant and Nonrelevant overlaps shown in Equations 1 and 2. All of the Lee result sets had a high Relevant overlap and a low Nonrelevant overlap and performed well with fusion.

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

$$N_{overlap} = \frac{N_{common} \times 2}{N_{vsm} + N_{PROB}} \qquad (2)$$

2.2 Fusion with Linear Combinations

In combining sources, it is natural to wonder if one source should be weighted more heavily than another. The idea is to apply a scalar to the similarity measures and to identify the best combinations of result sets. Bartell used numerical optimization techniques including a variation of Guttman's Point Alienation, a statistical measure of ranking correlation, and Conjugate Gradient to determine optimal scalars for a linear combination of results [Bartell94]. He achieved good results when running on very small collections (less then 50 MB). Unfortunately, these tests were done on such a small collection, it is not clear how well these results would scale to a larger collection. More recently, the FIRE system used the factors of result sets individual performance and dissimilarity of result sets to determine scalars for a linear combination. They experimented with merging of different web search engines on a subset of the TREC collection and found a 6.3% improvement in average precision over the best individual run when merging the top documents [Mounir98]. Vogt tested numerous linear combinations of several results from TREC-5 [Vogt98]. He examined 36,600 result pairs. A linear regression of several potential indicators was performed to determine the potential for result sets to be fused. Vogt considered thirteen factors. Some were measures of the individual inputs such as average precision/recall and some pairwise such as overlap and unique document counts. He found that the characteristics for effective fusion are 1) at least one result has high precision/recall, 2) high overlap of relevant documents, 3) low overlap of nonrelevant documents, 4) both distribute scores similarly, and 5) each rank relevant documents differently. Vogt found an average improvement of 34% on average precision/recall when using his model for merging two result sets.



DOC

Docid	doc_name	date	dateline	doclen
1	WSJ870323-0180	3/23/87	Turin, Italy	56
2	WSJ870323-0181	1/21/95	Chicago	126
...	...	...	...	....

INDEX

doc_id	term	tf
1	dogs	1
1	walked	1
1	home	1
2	home	1
2	range	1
...	...	...

TERM

term	idf
according	0.7782
commercial	1.0000
dogs	1.3222
home	1.2553
range	1.8451
walked	0.6021
...	...

QUERY

term	tf
dogs	1
home	1

Figure 1: The Design of the Scalable IR Engine

3. A COMMON ENVIRONMENT FOR FUSION

The basis for our fusion environment is the relational platform for information retrieval described in [Grossman97] and implemented in the Scalable Information Retrieval Engine (SIRE). The SIRE system has ranked well at the Text Retrieval Conference (TREC) over the past seven years. The vector space approach with various cosine-based similarity measures was used within SIRE for the TREC experimentation. An example of such SQL is shown in SQL1, using the cosine similarity measure.

SQL 1:

```
SELECT d.DocName, SUM((i.tf * t.idf * q.tf *t.idf)/d.doclen)
    FROM Index i, Doc d , Query q, Term t
        WHERE d.Docid = i.Docid
        AND q.term = i.term
        AND t.term = q.term
    GROUP BY d.DocName
    ORDER BY 2 DESC;
```

Modifications to the SUM() element permit implementation of most leading similarity measures. For instance, with the additional computation and storage of some document statistics, (log of the average term frequency), some collection statistics (average document length and the number of documents) and term statistics (document frequency), the following measures are implemented. In the pivoted normalization measure, the constant .20 is the pivot value proposed in [Singhal96]. In the OKAPI measure, the constants are the values described in [Robertson98].

SQL 2: Pivoted normalization measure

SUM(((1 + LOG(i.tf)) / ((d.LogAvgTF) * (AvgDocLen + (0.20 * d.DocLen)))) * (t.idf * ((1 + LOG(q.tf)) / (q.LogAvgTF))))

SQL 3: OKAPI Probabilistic measure

SUM(LOG((((NumDocs - t.df) + 0.5) / (t.df + 0.5)) * ((2.2*i.tf) / (.3 + ((.9 * d.DocLen)/AvgDocLen) + i.tf))))

The Boolean operator TAND (threshold AND) is used in Information Retrieval to require a certain number of the specified query terms to be present in a document for it to qualify as relevant. This feature can be quite complex to implement in retrieval systems. However, with SQL, TAND is easily achieved by adding a HAVING COUNT(*) >= threshold_num_of_terms.

The flexibility and power of the relational platform make it possible to bring together several retrieval strategies in one common environment. This leads to using one index of the collection, created with a single parser, stop list and other parameters so that the only variation in approaches is the similarity measure. This isolates the retrieval strategy for examination of the effectiveness of merging various retrieval strategies. In current work, we have limited our experimentation to examining the fusion of the result sets of the two retrieval strategies. Interesting future work would be to merge the strategies within the SQL – essentially getting a fusion of similarity measures rather than a fusion of ranked lists.

4. EXPERIMENTATON AND RESULTS

We implemented the vector space model (VSM) and the probabilistic model (PROB) in SIRE, using the SQL shown above. We conducted baseline runs for fusion input using whole terms and two-term phrases from the title-only version of the TREC topics. We used no stemming or query expansion for our baseline runs. We then examined several enhanced runs, again keeping the parser and all system settings identical while using techniques based on prior calibrations for relevance feedback [Lundquist99] and mandatory concepts [Holmes98] to enhance the retrieval performance. Our relevance feedback run added the best terms from the top 20 documents (based on N*nidf, where N is the number of top documents the term is found in.) For all retrieval runs, we used identical query terms for both vector space and probabilistic retrievals. For our fusion, we merged results using the best fusion approach from the five proposed by Fox and Shaw, the CombMNZ approach (See Section 2) [Fox 94].

$$CombMNZ = (nsim_{VSM} + nsim_{PROB}) \times num\_found\_in \quad (3)$$

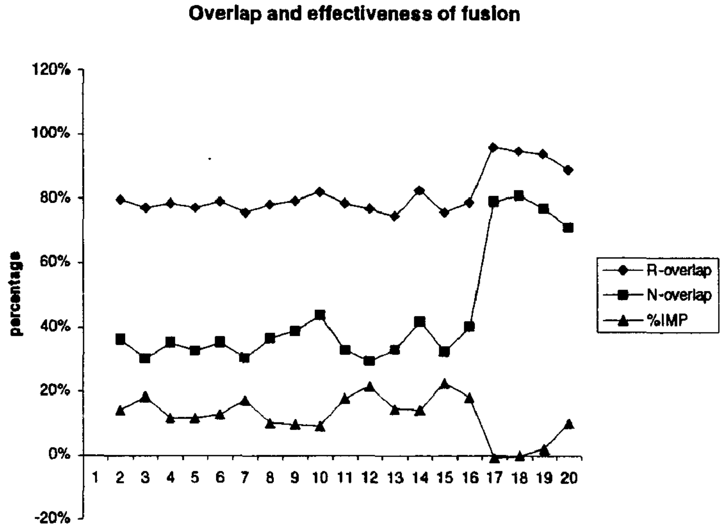


Figure 3 The impact of overlapping nonrelevant documents

This approach takes the sum of the normalized similarity of each input run and multiplies that sum times the number of runs containing the document. This benefits documents found in both runs (by a factor of 2) and works best when there is more overlap among relevant documents than nonrelevant. We analyzed the overlap in our result sets using Rcommon, Roverlap, Ncommon and Noverlap as described in [Lee 97]. A summary of our results is shown in table 1.

Figures 3 - 5 show that the two retrieval strategies are very similar in their effectiveness. The Overlap shown in Table 1 shows that the two approaches bring back very much the same documents – both relevant and non relevant.

Retrieval Runs	VSM	PROB	Common	Fused	Overlap
Baseline – avg. p/r	0.1626	0.1628		0.1611	
Relevant	2063	2079	1978	1934	0.96
Nonrelevant	42566	42550	33409	34985	0.79
Concepts –avg. p/r	0.209	0.1784		0.2083	
Relevant	2290	2208	2145	2246	0.95
Nonrelevant	38353	38662	31308	38593	0.81
Title w Rel.Feed.	0.1892	0.1829		0.1936	
Relevant	2397	2358	2233	2369	0.94
Nonrelevant	45196	44458	34377	45298	0.77
TREC6 runs					
Baseline-Title only	0.1700	0.1627		0.1877	
Relevant	1907	2018	1741	1956	0.89
Nonrelevant	41946	40564	29357	41804	0.71
Lee's Prior Work	westp1	vtc5s2			
Avg P/R	0.3157	0.2941		0.3734	
Relevant	6237	6077	4748	6857	0.77
Nonrelevant	43763	43923	13193	43143	0.30

Table: 1 Average P/R, R-overlap and N-overlap

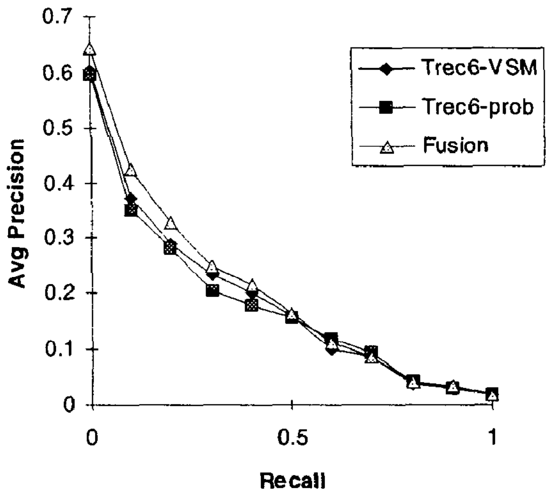


Figure 3 Average Precision/Recall on Trec6

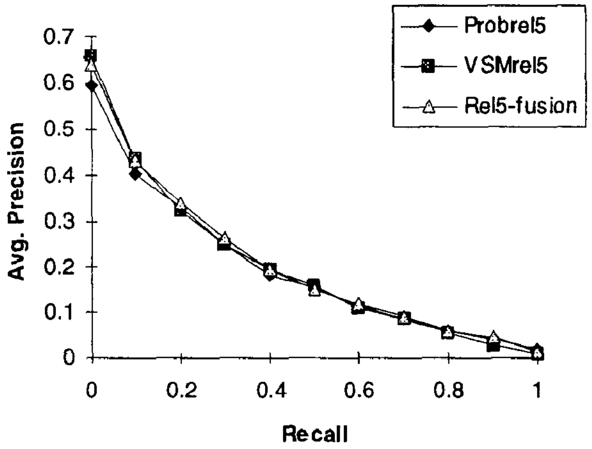


Figure 4 Average Precision/Recall on Trec7 with relevance feedback

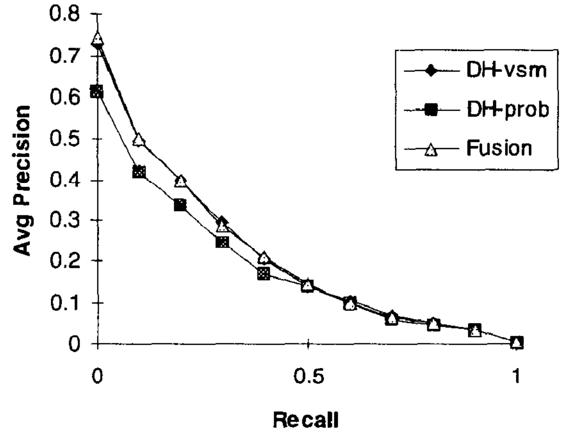


Figure 5 Average Precision/Recall on Trec7 with concepts

Prior work indicates that it is best to maximize the overlap of relevant documents and minimize the overlap of nonrelevant documents in order to achieve improvements with fusion [Lee97, Fox94]. Our results support this conclusion in that these two strategies have very high overlap of nonrelevant documents and they do not achieve much improvement with fusion. Lee's results (his best one is shown in Table 1) all had around 30-40% overlap Our input sets had much higher overlap in both areas. Figure 2 shows the relationship of the overlap of nonrelevant documents (N-overlap), the overlap of Relevant documents (R-overlap) and the percentage improvement of the fusion run's average precision/recall compared to run for our runs and all of the COMBMNZ runs from Lee. The negative relationship between N-overlap and fusion effectiveness is seen in the way the two lines diverge simultaneously. This is consistent with the analysis conducted by Vogt [Vogt98]. For each fusion run we experimented with scalars for weighting the approaches as shown in Equation 4 where α is the scalar for the vector space measure and β is the scalar for probabilistic.

$$CombMNZ_S = (\alpha(nsim_{VSM}) + \beta(nsim_{PROB})) \times number\_of\_runs\_found \quad (4)$$

This linear combination is an experimental approach to the weighting of experts based on performance by Thompson and to the use of a training set and numerical optimization for automatic learning of weights conducted in [Bartell 94]. Our results, shown in Figure 6 show that the scalars do not significantly impact the overall effectiveness of the combination. This is consistent with Thompson's work combining experts with weights for the individual inputs.

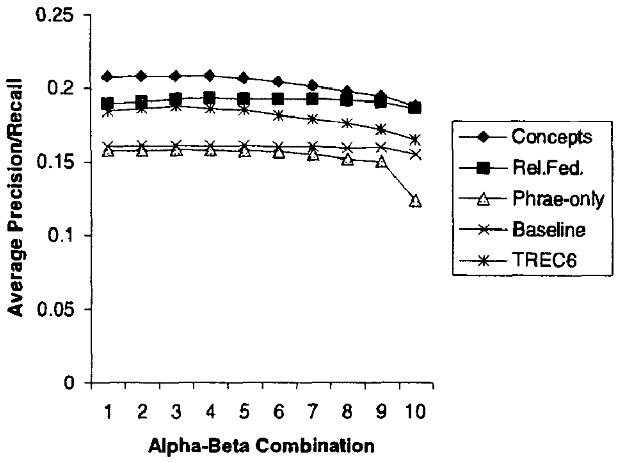


Figure 6 Linear scalars do not significantly impact fusion results

5. CONCLUSIONS AND FUTURE WORK

We have introduced a unified environment for fusing several retrieval strategies. Prior work has focused on fusing result sets that were generated from very different parsers, stop word lists, and lexical analysis. Granted, some prior work did use the same parser, but this work did not focus on the probabilistic and the vector space models – the two most common used models in the past few years at TREC. To our knowledge, we have not seen work done with fusion of these two approaches within a framework that ensures that no other factor affects the measurement of the effectiveness of fusion.

We have shown the implementation of the two leading retrieval strategies in this environment. The use of a common environment enables isolation of the role of the similarity measure itself in fusion, eliminating differences due to parsers and other system variations. In our fusion results, we have observed a negative correlation between the overlap of nonrelevant document sets and the efficacy of fusion. This is consistent with prior work.

Additionally, we have tested various linear combinations of merging VSM with probabilistic and have found that there was very little improvement to effectiveness. Finally, we have focussed on why our merging have resulted in only a .02% improvement. We have found that Lee's overlap ratio is very high for this result - while, in other results published by Lee, merging was successful when the overlap was very low. Our work suggests that the overlap is an excellent indicator of the potential for fusing VSM and probabilistic models and we have removed any concerns that parsing and other lexical analysis might have influenced this result. Clearly, other retrieval strategies should be attempted for future work -particularly those that return very different result sets from VSM and probabilistic.

6. REFERENCES

[Bartell 94] Bartell, B. T., G.W. Cottrell, and R.K. Belew. "Automatic combination of multiple ranked retrieval Systems" SIGIR '94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994.

[Fox 94] Fox, E. and J. Shaw. "Combination of Multiple Searches," Proceedings of the 2nd Text Retrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215, 1994.

[Grossman97] Grossman, D., O. Frieder, D. Holmes and D. Roberts, "Integrating Structured Data and Text: A Relational Approach," Journal of the American Society for Information Science, January 1997.

[Harman95] Harman, D., editor Proceedings of The Third Text Retrieval Conference (TREC-3), sponsored by the National Institute of Standards and Technology and Advanced Research Projects Agency, 1995.

[Harman98] Harman, D., editor Proceedings of The Third Text Retrieval Conference (TREC-7), sponsored by the National Institute of Standards and Technology and Advanced Research Projects Agency, 1998.

[Holmes98] Holmes D., M. McCabe, D. Grossman, A. Chowdhury and O. Frieder, "Use of Query Concepts and Information Extraction to Improve Information Retrieval Effectiveness," Proceedings of the Seventh Text Retrieval Conference (TREC), sponsored by the National Institute of Standards and Technology and Advanced Research Projects Agency, 1998

[Lee 97] Lee, J.H. "Analysis of multiple evidence combination." SIGIR '97: Proceedings of the Twentieth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,1997.

[Lundquist99] C. Lundquist, O. Frieder, D. Holmes, and D. Grossman, "A Parallel Relational Database Management System Approach to Relevance Feedback in Information Retrieval," Journal of the American Society for Information Science, 50(5), April 1999

[Mounir 98] Mounir, S., N. Goharian, M. Mahoney, A. Salem, and O. Frieder. "Fusion of Information Retrieval Engines (FIRE)", The Proceedings of PDPTA, 1998.

[Robertson98] Robertson S., S. Walker and M. Beaulieu, "Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive," Proceedings of the 7th Text Retrieval Conference (TREC)7, 1998.

[Shaw 95] Shaw, J.A., E.A. Fox. "Combination of Multiple Searches". The Third Text Retrieval Conference (TREC 3), 1995. National Institute of Standards and Technology Special Publication.

[Singhal96] Singhal, A., C. Buckley, and M. Mitra, "Pivoted Document Length Normalization," Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996.

[Thompson 90] Thompson, P. A combination of Expert Opinion Approach to Probabilistic Information Retrieval, part I: The Conceptual Model. Information Processing and Management, vol 26(3) 1990.

[Vogt 98] Vogt, C. and G. Cottrell., "Predicting the Performance of Linearly Combined IR Systems," Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.