Analysis of Combining Multiple Query Representations with Varying Lengths in a Single Engine

Abdur Chowdhury
America Online, Inc.
cabdur@aol.com

Steven Beitzel, Eric Jensen
Information Retrieval Laboratory
Department of Computer Science
Illinois Institute of Technology
{steve,ej}@ir.iit.edu

Abstract

We examine the issues of combining multiple query representations in a single IR engine. Differing query representations are used to retrieve different documents. Thus, when combining their results, improvements are observed in effectiveness. We use multiple TREC query representations (title, description and narrative) as a basis for experimentation. We examine several combination approaches presented in the literature (vector addition, CombSUM and CombMNZ) and present a new combination approach using query vector length normalization. We examine two query representation combination approaches (title + description and title + narrative) for 150 queries from TREC 6, 7 and 8 topics. Our QLN (Query Length Normalization) technique outperformed vector addition and data fusion approaches by as much as 32% and was on average 24% better. Additionally, QLN always outperformed the single best query representation in terms of effectiveness.

Introduction

Multiple query representations have been used in the past to improve the effectiveness of a given system. Different query representations are used because there is a belief that they will retrieve different documents. Thus, when combined the overall effectiveness of the system is improved. In this paper, we examine the effects of combining multiple query representations. Specifically, we examine the combination of query representations with different lengths. Our hypothesis is that when query representations of differing length are combined, vector addition approaches or traditional data fusion approaches are not appropriate. To examine this, the TREC [1] topics for TREC 6, 7 and 8 were used. TREC divides queries or topics into three distinct representations: title, description and narrative. Each representation is of increasing length. We examined the combination of query representations resulting from the combination of title with description and title with narrative. These combinations were selected to maximize the difference in query lengths and to examine the effects of fusing the results of multiple query representations in terms of effectiveness. Table 2 provides query length statistics for each of the topic sets and query representations. The goal of this research is to find a good strategy for combining query representations of varied length.

Several approaches have been suggested for combining multiple query representations [2, 3] and more generally for fusing ranked sets, commonly called data fusion [4]. These multiple-evidence techniques (data fusion) are touted as a means to improve the effectiveness of information retrieval systems. They are based on the premise that repeated evidence increases the probability of relevance. Historically, multiple-evidence research has considered multiple query representations within identical retrieval frameworks to be combinable into a single query for evaluation. This was typically achieved by summing the vectors of each query representation. However, our research has led us to believe that this is not necessarily optimal, and, in fact, even prior result combination techniques, which attempt to correct for score differences with normalization, may not perform optimally. This hypothesis is based on a belief that results obtained from query representations of differing length do not benefit from combination approaches that use statistical normalization, and the vector addition approach over-emphasizes results obtained from longer query representations.

In the next section, we review some prior approaches to multiple query representation combination and prior approaches to combining result sets (data fusion). In Framework, we present motivations for

finding better ways of combining multiple query representations of varying length, and describe our experiments. In Equation 5: Overlap (R = Relevant, NR = Not Relevant)

Results & Analysis we examine our results. Lastly, in Table 1: Difference in Effectiveness

Conclusions we examine what can be concluded from this research and give an overview of future work.

Prior Work

Research in query multiple-evidence techniques has focused primarily on combining query representations or fusing representation-specific results to create a single result set. Many information retrieval systems combine multiple query representations combination by performing a sum of all query vectors prior to any retrieval. This involves the union of the sets of query terms along with the addition of term frequencies for co-occurring terms. Results of this strategy depend on the ranking algorithm to order the unified result set effectively. Turtle and Croft examined the effects of combining the evidence from two or more formal query requests in an inference engine [2].

Fox and Shaw [4] proposed several result combination algorithms and found that combinations of different ranking strategies yielded improvements in overall system effectiveness. One of their approaches was the CombSUM algorithm in which results from each strategy had their scores (min-max) normalized and documents that were present in multiple result sets had their scores summed. Their CombMNZ algorithm was based on the same normalization, but also attempted to account for the value of multiple evidence by multiplying the sum of the scores of a result by the number of result sets in which it was present. Belkin, et al. [3] examined the effects of these algorithms by combining various manually generated query representations.

Lee [5] further examined various combination algorithms for fusing result sets to improve effectiveness. In combining results from varying retrieval systems, he identified that for multiple-evidence to improve system effectiveness the retrieved sets must have higher relevant overlap than non-relevant overlap. Lee did not identify the exact difference needed to improve effectiveness. Additionally, his results had a 125% difference in relevant and non-relevant overlap.

The importance of overlap was also identified in [11], which showed that when combining results from multiple engines, the combination of the most different engines was most effective. A comparison of their combination approaches to CombSUM and CombMNZ also showed the importance of accounting for actual scores returned by the various engines, as opposed to only their rankings or score distributions.

Montague further examined data fusion. He compared the effectiveness of the min-max score normalization to other approaches such as z-score normalization [6]. He showed that when combining of results from multiple systems, z-score normalization methods that were less sensitive to boundary and outlier values were more effective. Our research builds upon this in the specific domain of merging query representations within a single system.

Framework

We hypothesize that as query representations differ increasingly in length and terms, the probability that the final result sets will have high overlap is lower, thus data fusion techniques will not produce the improvements seen in the literature. Additionally, vector addition approaches do not account for the varied query length in an appropriate manner and thus the overall improvements gained by using multiple query representations is not optimal. To explore this we first examine ranking strategies to see how the combination of multiple query representations via vector addition may affect the final ranking. We then examine the issues involved when applying data fusion techniques to results from multiple query representations of varying length.

$$\frac{\sum_{j=1}^{t} w_{qj} d_{ij}}{\sqrt{\sum_{j=1}^{t} (d_{ij})^2 \sum_{j=1}^{t} (w_{qj})^2}}$$

Equation 1 : Cosine Measure [7]

Where:

- w_{qj} = *weighting factor for query term j*
- d_{ij} = *weighting factor for document i's term j (both typically some combination of tf*idf)*

Relevance ranking began by considering the frequency of each query term in the retrieved documents. As ranking algorithms matured, they added term weighting, and began to model queries and documents as vectors whose similarity could be measured by calculating the distance between them on the Euclidean plane. It was realized that this distance misrepresented the similarity of many query-document relationships because queries are generally much shorter than documents. This led to the progression from using distance as a similarity measure to using the cosine of the angle between the vectors, which effectively normalized similarity scores based on the lengths of the vectors, removing length from consideration. In Equation 1 we

show an example of cosine ranking where the query's length is a factor in the ranking of a given document.

However, when executing a query against many documents using a single query representation, $\sqrt{\sum_{j=1}^{t}(w_{qj})^2}$ is constant for each document *i*. Since single query representations are the traditional form of querying, newer similarity measures ignore this aspect of document ranking and focus on improved length normalization for documents. This can be seen by examining several effective ranking strategies like pivoted document length normalization [8] (vector space) and Okapi BM25 (probabilistic) [9].

$$\sum\left(\frac{1+\ln(1+\ln(tf))}{(.8+.2*(docsize/avgdocsize))}\right)*idf*qtf$$

Equation 2: Pivoted Document Length Normalization

$$\sum\log\left(\frac{(N-n)+.5}{(n+.5)}\right)*\left(\frac{2.2*tf}{.3+(.75*docsize/avgdocsize)+tf}\right)*qtf$$

Equation 3: Okapi BM25

Where:

- *tf = frequency of occurrences of the term in the document*
- *qtf = frequency of occurrences of the term in the query*
- *docsize = document length*
- *avgdoclength = average document length*
- *N = is the number of documents in the collection*
- *n = is the number of documents containing the word*
- *idf = log(N/n)*

Vector addition approaches to combine multiple query representations simply add the query term vectors together. When a term occurs in multiple representations, its frequency is updated for the ranking strategy. This has the effect of over-emphasizing the longer query representations' retrieved documents. Documents retrieved from a longer query have a higher chance of obtaining a greater score, while the documents retrieved from shorter representations, which may be very relevant, are unfairly compared. This can be seen when there is low overlap in retrieved documents from each representation, as this is when normalization is most important.

General multiple-evidence techniques such as data fusion approaches strive to reconcile the difference in ranked sets by normalizing score ranges from each set in order to combine results. Our technique works towards this goal by examining the scenario of combining multiple query representations where each query vector is considered separately, but executed in the same retrieval engine using the same strategies. In this case, score variances can be defined solely in terms of the lengths of the queries. This specific focus allows us to address the problem of result combination as an application of query length normalization. We believe that normalizing document scores based on query length will achieve a more appropriate interleaving of documents when results are fused.

Data fusion techniques use normalized weights to interleave results from retrieved document sets. The goal of this normalization is to remove any non-semantic differences in the properties of each query representation and corresponding result set from consideration by the ranking process. This prevents cases where results from a longer query representation are ranked higher than others simply because they have a higher score due to the length of their query. In addition, multiple-evidence can be used as an additional ranking factor, as in the case of CombMNZ.

However, traditional normalization is based only on statistical analysis of the scores in the result sets (min-max, mean, etc.). Statistical normalization is inappropriate when all retrieval strategies and utilities are held constant for all queries. In this case, the actual values of the scores of the various result sets only differ because of the corresponding query representations themselves. This makes utilization of these values during result comparison plausible and indeed desirable (since they are the original form of similarity data). Since the underlying causes of the score differences are not considered by statistical approaches, scores of result sets must be normalized so that only their distribution remains. Clearly, some vital ranking information is being lost.

The solution to this problem is to normalize query representation scores with a scheme that accounts for the underlying causes of their differences, thereby making their results comparable without loss of similarity data. We approach this by focusing on the most primary difference between query representations: their length.

Our approach consists of normalizing scores by selecting the shortest query representation as the base score range, and multiplying scores of longer query representations by the ratio of the length of the shortest query representation to the longer query representation as shown in Equation 4.

$$score(d)_{total} = score(d)_{shortest} + \sum_{i=1}^{n-1}\frac{length(shortest)}{length(i)} * score(d)_i$$

Equation 4: Query Length Normalization

Where:

- *d = a result document*
- *score(d) = similarity score for document d*
- *shortest = query representation with the least distinct terms*
- *length(q) = number of terms in query representation q*
- *n = total number of query representations*

Overlap is examined as a function of how well fusion techniques will perform. Relevant overlap and non-relevant overlap are calculated as in Equation 5. Additionally, we examine the total overlap of returned documents. Since we believe that different query representations return different documents, then overlap may be a good indicator of how fusion techniques will perform. Lastly, we examine effectiveness of the given query representations. Since, our technique uses a single ranking strategy and a normalization method that preserves original scores while removing differences in query length, we believe this approach will be less prone to degradation in overall effectiveness when a single representation is less effective.

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

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

Equation 5: Overlap (R = Relevant, NR = Not Relevant)

Results & Analysis

Throughout, we use the title, description, and narrative query representations from the TREC 6, 7, and 8 query sets. The title query consists of a short keyword query, which averages 2.5 terms in length (after stop word removal) over all three query sets, the description is a sentence which averages 16 terms in length, and the narrative is a paragraph which averages 47 terms in length. All three represent the same information need.

In order to support our hypothesis, we designed and executed a series of experiments using the various query representations on the ad-hoc query sets of TREC-6, TREC-7, and TREC-8. All experiments were performed using our lab's IR engine, AIRE [10]. All relevant systemic differences (retrieval strategy, stemming, parsing, phrasing, etc.) were held constant throughout all experiments so that we could focus our efforts on analyzing the behavior of data fusion of results from various query representations.

To examine our hypothesis we ran a series of experiments where each query representation was used by itself as a baseline. Then the following combination approaches were executed: Vector Addition, CombSUM, CombMNZ, and our Query-Length Normalization technique. We theorized that our QLN technique would outperform the most effective combination technique for cases when overlap was high, and that QLN would perform comparably to fusion approaches in cases where overlap was low. Additionally, we believed that when there was great difference in the effectiveness of the single representations, QLN would outperform the other techniques.

The full results of our experiments are given in Table 4 and Table 5 and the overlap analysis is given in Table 3. Examining Table 3 we see that the overlap between title and description or narrative query representations is very low, thus fusion techniques like CombMNZ are predicted to not do much better than fusion techniques like CombSUM. Table 4 and Table 5 demonstrate that the average precision is comparable between CombSUM and CombMNZ when fusing results from title + description and title + narrative as expected. What these approaches have in common is that each result set is normalized first and then fused. What we find was the overlap was NOT a good indicator of how well fusion would perform. Lee's 125% difference in relevant and non-relevant documents did not exist in our different query representations. Thus, CombSUM and CombMNZ performed equivalently for our experiments.

When examining the vector addition approach, we see that for TREC 6 and TREC 8 title + description and title + narrative it performs worse than the best of the single query representations. Examining this further, we see that as the difference in effectiveness of the query representations (Table 1) grows, the vector addition approach does consistently worse. This is best illustrated in the TREC 6 title + narrative run where there was a 43% difference in effectiveness between the individual representations and the effectiveness of the vector addition approach is 12% worse than the best single representation.

Lastly, we examine our vector normalization QLN technique. We believe that this allows for more appropriate merging of results. This technique produced a better final result for all six of the experiments when compared to the most effective single query representation. When compared to the vector addition approach (CombV), vector normalization did better in five of the six experiments and on average did 18% better in terms of overall effectiveness. When comparing QLN (Vnorm) to CombMNZ, we see that it performed better in four out of the six experiments and on average did 24% better. As seen with vector addition, when the effectiveness of each of the individual representations is greatly different CombMNZ does not perform as well as our vector normalization approach.

Avg P	Title	Description	Diff
T6	23.03%	16.11%	42.95%
T7	17.68%	18.85%	6.62%
T8	24.58%	22.05%	11.47%

Table 1: Difference in Effectiveness

Conclusions

Our results have shown that query length normalization is an effective method of dealing with the problem of normalizing results from multiple query representations to facilitate their combination. They have also shown that overlap is not an effective indicator of when fusion techniques will or will not help. Our technique outperformed both vector addition (CombV) and traditional data fusion with min-max normalization (CombSum and CombMNZ) on average. It was always better than any single query representation and was never significantly worse than any combination approach. It performed best when difference in effectiveness of individual query representations (Table 1) was highest. This was especially evident when comparing it to Vector Addition (CombV), which we believe behaved poorly in these cases due to negative influences of query terms from the worse representation (either the union of the terms themselves or a disadvantageous TF boost). The largest improvements over fusion techniques also occurred in this situation. We believe this is because of the loss of information when equalizing the ranges of scores from different query representations. When query representations retrieve very different results, perhaps some are better than others and they should not be treated equally.

	Title			Description			Narrative		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
T6	1	5	2.7	5	62	20.4	17	142	65.3
T7	1	3	2.5	5	34	14.3	14	92	40.8
T8	1	4	2.5	5	32	13.8	14	75	35.5

Table 2: TREC 6, 7 and 8 Query Length Statistics

	1000				10			
	Overlap	R-Overlap	N-Overlap	Distinct	Overlap	R-Overlap	N-Overlap	Distinct
T6 - T+D	25.556%	75.821%	60.854%	86.810%	21.200%	68.829%	60.028%	89.357%
T7 - T+D	41.940%	84.003%	67.771%	78.460%	41.000%	80.456%	62.203%	79.500%
T8 - T+D	37.316%	82.042%	62.952%	81.170%	33.600%	65.789%	62.478%	83.200%
T6 - T + N	19.28%	72.44%	58.86%	90.05%	15.00%	64.66%	54.61%	92.47%
T7 - T + N	22.24%	72.83%	59.04%	88.58%	19.20%	64.79%	54.81%	90.40%
T8 - T + N	22.75%	64.32%	57.92%	88.52%	16.20%	55.35%	61.62%	91.90%

Table 3: Overlap Analysis of Query Representations Ranked Retrieved Sets

T6	Title	Description	CombV	CombSUM	CombMNZ	Vnorm
Avg P	23.03%	16.11%	22.57%	22.64%	23.45%	25.47%
Imp / Best			-2.00%	-1.69%	1.82%	10.59%
Imp / Vector				0.31%	3.90%	12.85%
Imp / MNZ						8.61%
T7						
Avg P	17.68%	18.85%	20.88%	20.61%	20.70%	20.46%
Imp / Best			10.77%	9.34%	9.81%	8.54%
Imp / Vector				-1.29%	-0.86%	-2.01%
Imp / MNZ						-1.16%
T8						
Avg P	24.58%	22.05%	26.34%	26.27%	26.37%	26.50%
Imp / Best			7.16%	6.88%	7.28%	7.81%
Imp / Vector				-0.27%	0.11%	0.61%
Imp / MNZ						0.49%

Table 4: Title + Description Combination Experiments

T6	Title	Narrative	CombV	CombSUM	CombMNZ	Vnorm
Avg P	23.03%	16.09%	20.14%	24.00%	24.32%	26.67%
Imp / Best			-12.55%	4.21%	5.60%	15.81%
Imp / Vector				19.17%	20.75%	32.42%
Imp / MNZ						9.66%
T7						
Avg P	17.68%	15.29%	20.74%	21.79%	22.09%	21.66%
Imp / Best			17.31%	23.25%	24.94%	22.51%
Imp / Vector				5.06%	6.51%	4.44%
Imp / MNZ						-1.95%
T8						
Avg P	24.58%	15.05%	22.85%	24.73%	24.89%	26.31%
Imp / Best			-7.04%	0.61%	1.26%	7.04%
Imp / Vector				8.23%	8.93%	15.14%
Imp / MNZ						5.71%

Table 5: Title + Narrative Combination Experiments

References

[1] http://trec.nist.gov, *TREC Homepage*, November 20, 2001.

[2] H. Turtle, W. B. Croft, "Evaluation of an Inference Network-based Retrieval Model". *ACM Transactions on Information Systems*, 9:3, July 1991, pp. 187-222.

[3] N. Belkin, P. Kantor, E. Fox, J. Shaw, "Combining the evidence of multiple query representations for information retrieval", *Information Processing & Management*, 31:3, pp. 431-448, 1995.

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

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

[6] M. Montague, J. Aslam, "Relevance Score Normalization for Metasearch", *ACM-CIKM*, pages 427-433, Atlanta, 2001.

[7] G. Salton, C. Yang, A. Wong, "A Vector-Space Model for Automatic Indexing", *Communications of the ACM*, 18:11, pp. 613-620, 1975.

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

[9] S. Robertson, 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.

[10] A. Chowdhury, S. Beitzel, E. Jensen, M. Saelee, D. Grossman, O.Frieder, "IIT-TREC-9 - Entity Based Feedback with Fusion", *Proceedings of the Ninth Annual Text Retrieval Conference*, NIST, November 2000.

[11] S. Mounir, N. Goharian, M. Mahoney, A. Salem, O. Freider, “Fusion of Information Retrieval Engines (FIRE)”, *Int. Conf. on Parallel and Distrib. Proc. Tech. and Appl*., LV., 1998.