Experiments on Data Fusion Using Headline Information

Abstract

This poster describes initial work exploring a relatively unexamined area of data fusion: fusing the results of retrieval systems whose collections have no overlap between them. Many of the effective meta-search/data fusion strategies gain much of their success from exploiting document overlap across the source systems being merged. When the intersection of the collections is the empty set, the strategies generally degrade to a simpler form. In order to address such situations, two strategies were examined: re-ranking of merged results using a locally run search on the text fragments returned by the source search engines; and re-ranking based on cross document similarity, again using text fragments presented in the retrieved list. Results, from experiments, which go beyond previous work, indicate that both strategies improve fusion effectiveness.

Introduction

Data fusion is the merging of ranked document lists from *source search engines* into a unified list. Several merge methods have been proposed, each relying on different conditions or forms of information being returned by search engines. Lee described CombMNZ, which multiplied the sum of the scores returned by each search engine by the number of engines with non-zero scores (Lee, 1997). Aslam & Montague (2001) proposed a Borda-fuse voting model where documents were considered to have been 'voted' for by the search engines that retrieved them. The votes a document received determined its fused rank position. Manmatha et al (2001) reported a double distribution fit model, which was almost as effective as CombMNZ, despite ignoring document overlap. However, the method was only tested on source ranks of 1,000 documents containing a good number of relevant documents. The method's effectiveness on more common situations of short rankings with few relevant documents remains to be demonstrated. The three strategies rely on collection overlap and/or score information being returned by search engines. Overlap is expected when fusing different retrieval strategies, and is common for meta-search engines on the Web, but when merging the results returned by a set of dispirit digital libraries, overlap is much less likely. In addition, source search engines rarely return a similarity score when presenting a retrieved set.

Taking an approach that does not require such conditions, Lawrence & Giles performed a *local search* on a collection formed by downloading all documents retrieved by the source search engines (1998). Despite working well, full text downloading is slow and on some digital libraries may incur cost. Addressing this problem, Tsikrika & Lalmas (2001) reported on a local search method using only title & summary text as well as source rank position when fusing[1]. Tested on 10 queries, the method was found to be as effective as Lawrence and Giles's 'full text' local search, a somewhat striking result.

For our experiments, we chose to test methods that do not rely on overlap or document score information being present as it is our contention that such a situation is common, and is relatively unexplored. Our methods test the success of Tsikrika & Lalmas's local search on a much wider range of queries, combinations of retrieval systems, and sizes of text returned including much shorter texts such as document title only. In addition, we examined a novel approach inspired by CombMNZ and Borda-fuse: approximating document overlap by measuring the similarity of documents across the source rankings to be merged. As with the overlap-based methods, which ranked higher documents that were duplicates of each other, our *cross rank* approach, places documents higher in the fused ranking if they are similar to each other.

The rest of this short paper describes our fusion methods in more detail and means by which we tested them. Results are presented next, before concluding.

Search methods

Local search: Here, the document titles returned by the source search engines were formed into a new collection and a search (based on BM25 weighting) was performed locally with the resulting single ranking being evaluated for effectiveness. To establish an upper bound on our experimental results, local search on full document text (a la Lawrence & Giles) was also conducted. **Lower bounds – random and round robin**: To establish a lower bound on performance, the effectiveness of a *round robin* technique was measured: ranking the fused documents based solely on their rank position from source search engines. An additional lower bound based on randomly sorting the fused ranking was also measured. **Cross Rank Similarity Comparison**: Here, we examined how similar a document was to other retrieved documents. If there were more documents similar to a specific document, then it was ranked higher. Cross rank similarity comparison was computed by summing the (BM25-based) similarity score of a source document headline with the headlines of the other source documents. The similarity scores were normalised and documents were fused into a single ranking sorted by their normalised score.

Test Data Set

The target of our experiments on data fusion tried to simulate a realistic situation where each source engine searched a different non-overlapping document collection. For our experiments, we decided to examine fusing the top 20 documents returned by 4 non-overlapping search engines. Our fusion models worked on ranking the 80 resultant documents. As with other fusion work (e.g. Aslam, Lee, Savoy), TREC data was used in our experiment. We examined sets of 4 rankings chosen from the 21 systems that submitted runs to the short topic category-A adhoc part

[1] It is our understanding that a similar idea was reported at a SIGIR workshop some years ago, though not published.

of TREC-5. In order to test fusion of non-overlapping collections, the rankings for each source system was filtered so that each only returned documents from a ¼ of the TREC collection. We additionally restricted our collection to the newspapers/wires of TREC-5 alone: FT, WSJ, and AP. The average length of the headlines in collection was 12 words. Four search engines associated with the 4 sub collections were selected from the 21 available. Based on each source search engine's *mean average precision* (MAP) calculated by TREC, they were ranked from 1 to 21. Different combinations of search engines were chosen in order to observe the methods in a range of fusion situations. Fifty queries (Topic 251-300 from trec-5) were used in our experiments.

Experiment Results

The results of our experiments are listed in Table 1, presenting MAP, rows are sorted by round robin score. From the table, it can be seen that cross rank works slightly better than round robin, which is encouraging since it ignored source rank position and was based on inter-headline similarity alone. Local search on headlines was markedly better than the other two regardless of the systems being fused. The upper bound of local full text search was the best by far; quite a different result from Tsikrika & Lalmas who reported local searching on the different amounts of source text resulting in similar effectiveness. We note that in their work, local search only improved on round robin by 10% where here, improvement was 159%. We speculate that the known success of BM25 ranking used in our system may be the reason. Tsikrika & Lalmas used an alternative ranking method.

Search Engine rank	**Random Sort**	**Round Robin**	**Cross Rank**	**Local Headline**	**Local Full Text**
1,3,4,5 (best)	0.1518	0.2025	0.1747	0.2337	0.4149
1,8,15,21	0.1353	0.1857	0.1691	0.2224	0.3997
4,9,14,17	0.1702	0.1829	0.1866	0.2325	0.4020
1,3,5,21	0.1364	0.1708	0.1769	0.2464	0.3960
11,12,13,14	0.1297	0.1634	0.1629	0.2216	0.4317
2,3,18,19	0.1596	0.1594	0.1788	0.2197	0.4305
1,19,20,21	0.1252	0.1337	0.1714	0.2299	0.4525
18,19,20,21 (worst)	0.1109	0.1025	0.1179	0.2326	0.4501[2]
Average	**0.1399**	**0.1626**	**0.1673**	**0.2299**	**0.4222**
% change	-13.96%	0.0%	+2.89%	+41.39%	+159.66%

Table 1: Mean average precision for different fusion methods and different search engines

Other than the presented results, various other fusion models were tried such as cross rank comparison using different lengths of a document's opening section (i.e. 40, 50, 60, 80, 100 words), computing cross rank similarity across smaller document sets, mixture of cross rank and local headline search, mixture of cross rank and round robin method, etc. In general, the more text used, the better fusion became. The performance of the mixture of cross rank and local headline search was in between applying these two models alone. Therefore, under these conditions, local search on retrieved headlines was found to be the most cost effective and efficient method.

Conclusion

We have presented 2 methods of data fusion that make minimal assumptions of the source systems being searched. Despite ignoring score information and document overlap and using only the minimal text returned in a ranked list, an effective means of merging has been produced. This is, however, preliminary work and we plan to extend our study examining further means of merging cross rank and local search. Also, we will explore improvements on the cross rank similarity calculation and local search ranking method: both computed *inverse document frequency* (IDF) scores of terms from the local 80-document collections, estimating IDF from a larger collection maybe beneficial.

References

Javed A, Aslam and Mark Montague; Models for Metasearch; Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), New Orleans, Louisiana, USA, pp 276-284

Steve Lawrence and C. Lee Giles; Context and page Analysis for Improved Web Search; IEEE Internet Computing, Volume 2, Number 4, pp 38-46, 1998

Joon Ho Lee; Analyses of Multiple Evidence Combination; Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), Philadelphia, Pennsylvania, UAS, July 1997, pp 267-275

R. Manmatha, T. Rath and F. Feng; Modelling Score Distributions for Combining the Outputs of search Engines; Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), New Orleans, Louisiana, USA, pp 267-275

Jacques Savoy, Anne Le Calvé and Dana Vrajitoru; report on the TREC-5 Experiment: Data Fusion and Collection Fusion; Proceedings of the TREC'5, NIST Publication 500-238, Gaithersburg (MD), November 1997, pp 489-502

Theodora Tsikrika and Mounia Lalmas; Merging Techniques for Performing Data Fusion on the Web; Proceedings of ACM CIKM'01 conference, November 5-10, 2001, Atlanta, Georgia, USA, pp 127-134

[2] The high MAP when fusing worse source rankings is due to fewer relevant source documents being retrieved.