A Dynamic Fusion Method Applied to Information Retrieval

Antonio Juárez-González

Laboratory of Language Technologies, Department of Computational Sciences,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
antjug@inaoep.mx

1 Motivation

The great amount of available digital content has motivated the development of several Information Retrieval (IR) approaches, which help users to locate useful documents for their specific information needs. All these approaches differ one from another in many issues, such as the preprocessing of the data, data representation, weighting scheme, etc. [1]. However, any approach can satisfy all the possible information needs from users. For these reason, Data Fusion (DF) has been used in IR to combine different result sets of a group of retrieval systems to bring together the strengths of each individual IR system. DF is a process (acquisition, design, and interpretation) of combining information gathered by multiple agents (sources, schemes, sensors or systems) into a single representation (or result) [5]. The goal of Data Fusion is to obtain a result with a higher performance than the individual results to be fused. The most widespread use of DF consists to combine all the retrieval result lists for all queries without paying attention to the quality of each individual list. In this case, if one of the lists has a very low performance, the fusion performance drops significantly. If, for each topic, we could identify the lists that help DF to perform better, we could have a significant improvement over the fusion of all available lists for all topics. Table 1 shows the possible gain in *MAP* by selecting the *n* lists that maximize the fusion performance, considering the combMNZ [3] fusion method and four data sets from CLEF[1].

Table 1. Gain of the fusion of selected lists over the fusion of all lists.

	Ad hoc 2005		**GeoCLEF 2008**		**ImageCLEF 2008**		**RobustCLEF 2008**	
	MAP	*Gain*	*MAP*	*Gain*	*MAP*	*Gain*	*MAP*	*Gain*
All lists	0.275		0.244		0.302		0.341	
2 lists	0.337	**22.5%**	0.340	**39.3%**	0.397	**31.4%**	0.423	**24.0%**
3 lists	0.330	**20.0%**	0.323	**32.3%**	0.379	**25.4%**	0.417	**22.8%**
4 lists	0.305	**10.9%**	0.278	**13.9%**	0.359	**18.8%**	0.382	**12.0%**

The manual analysis of table 1 shows that the previous selection of lists for the fusion process, can improve the performance of the retrieval results up to 22.5%, 39.3%, 31.4% and 24.0% for each data set, respectively. This analysis motivated our

[1] *www.clef-campaign.org*

investigation about an automatic method to perform the selection of lists that improve the fusion process.

2 Previous works in the Area

Given a set of result lists, selecting the lists to be fused to ensure gain over the best global list in the set remains as open problem. To our knowledge, there are only few works addressing this problem. Vogt and Cottrell [7] performed an analysis to investigate the feasibility of predicting the performance of the fusion of list pairs, using a linear regression method. A total of 15 characteristics from the list pairs were used to train the model. They conclude that their method is suitable when: *i)* at least one of the lists has a good performance; *ii)* both lists have similar sets of relevant documents; *iii)* the lists contain different sets of non-relevant elements; and *iv)* the scores of the elements in the lists have a similar distribution but a different ranking of the relevant elements. Ng and Kantor [6] considers the problem of predicting when the fusion of two lists improves the performance of the best of them. In contrast to [7], they used only two characteristics from the list pairs (proportion of the lower performance over the best performance and the out of order pairs between lists) to train a logistic regression model. Their model could identify in the test set 69% of the positive cases. Wu and McClean [8] extends the investigation made in [6, 7] by considering three different objective variables: *i)* data fusion performance, *ii)* performance improvement over average performance, and *iii)* performance improvement over best performance. They used three data sets, fusions of 3 to 10 lists, and three data fusion methods (combSUM, combMNZ and roundrobin). A multiple regression model was constructed using 4 to 10 characteristics from each group of lists fused. They conclude that the independent variables considered are effective to predict the values of the objective variables, however, as some of those variables are dependent of relevance judgments, they need an automatic method to identify the relevant documents in the lists. Gopalan and Batri [4] presented an investigation for selecting the best *m* retrieval strategies and the best fusion method for a collection. All the possible fusions of two and three result lists were generated and used in a genetic algorithm which used the *MAP* of the fusions as fitness function. They could obtain a maximum gain of 8.4% over the best list by generating a final result list using the identified *m* strategies and fusion method by the genetic algorithm in the specific collection for all queries. Diamond and Liddy [2] proposed a dynamic data fusion model. In their analysis they observed that there is a considerably opportunity of gain in the retrieval performance (up to 34% considering *prec@30*) by applying a different linear weighted fusion function to each query instead using the same static linear weighted function to all queries.

3 Research objectives

Our research goes a step further than analyses in [6, 7, 8] by proposing an automatic method to select the lists that could improve fusion avoiding dependence of the relevance judgments. In contrast to [4] we won´t identify the best lists for all the queries in a determined collection, we will select the lists that could improve the fusion results considering each topic individually, independent of the data set. Our method won’t need previous knowledge about the retrieval systems as in [2], instead it will use the similarities of the lists to assign them a relevance value. Additionally, our method won't need to perform fusion until the best lists are identified, will use an unsupervised approach which will allows its use in any set of lists. Our research has the next objective: *Develop an automatic method to perform the Dynamic Fusion of Results of Information Retrieval Systems*. In our main objective, we consider *Dynamic Fusion of Results* as the process of *i)* identifying the more adequate lists to be fused, and *ii)* selecting the fusion method to be applied.

4 Methodology

Our methodology follows the next main steps: *i) proposal of characteristics*, where several characteristics based in redundancy an ranking of the elements in the lists will be extracted to try to capture the relevance of each result list; *ii) relevance measurement*, where the most useful characteristics for measure the relevance of the lists will be selected; *iii) selection of the n more relevant lists per query*, where the characteristics will be used to select and fuse a fixed number of lists; *iv) selection of a variable number of lists per query*, where the characteristics will be used to select and fuse an unfixed number of lists; and *v) selection of the fusion method*, where an analysis of the data in the previous steps will be analyzed to select a fusion method for a determined list group.

5 State of the research and preliminary results

Our research has completed the first three main steps of the methodology. So far we have proved several characteristics to measure the relevance of the lists for the fusion process. The idea behind this measure is that the relevance of a list increments by the presence of common documents at the very first positions. Table 2 shows the global performance of our method with four fusion methods, and it is compared with the performance of the fusion of all lists (baseline).

We can observe that our method is able to perform better than the baseline fusion in almost all cases. Fusion with previous List Selection obtains a gain over the fusion of all lists up to 6.0%, 9.0% and 10.4% for data set Ad hoc 2005 for maximum RSV, combMNZ and Fuzzy Borda fusion methods, respectively, using n=2 for all cases; for data set GeoCLEF 2008 the gains are up to 18.8%, 12.2% and 14.7% for maximum RSV (n= 2), combMNZ (n= 3) and Fuzzy Borda (n= 3); for ImageCLEF 2008 the gains obtained are up to 23.5%, 12.5% and 7.4% for maximum RSV (n= 2), combMNZ (n= 3) and Fuzzy Borda (n= 3); finally, for RobustCLEF 2008 we could obtain gains up to 24.6%, and 62.2% for maximum RSV and Fuzzy Borda (both with n= 2).

Table 2. Performance in MAP of Data Fusion and Fusion with List Selection

Method	Ad hoc 2005	GeoCLEF 2008	ImageCLEF 2008	RobustCLEF 2008
Fusion of all lists				
maximum RSV	0.231	0.180	0.251	0.231
combMNZ	0.275	0.244	0.302	0.341
Fuzzy Borda	0.267	0.251	0.321	0.167
List Selection *n* = 2				
maximum RSV	**0.245**	**0.214**	**0.310**	**0.288**
combMNZ	**0.300**	0.233	**0.333**	0.334
Fuzzy Borda	**0.295**	**0.266**	**0.341**	**0.271**
List Selection *n* = 3				
maximum RSV	0.229	**0.188**	**0.303**	**0.263**
combMNZ	**0.281**	**0.274**	**0.340**	0.328
Fuzzy Borda	**0.285**	**0.288**	**0.345**	**0.261**
List Selection *n* = 4				
maximum RSV	0.225	0.177	**0.287**	**0.246**
combMNZ	0.274	**0.261**	**0.323**	0.324
Fuzzy Borda	**0.278**	**0.286**	**0.335**	**0.223**

The evaluation results obtained so far, allow us to establish the following conclusions: *i) p*erformance of Data Fusion can be improved with a previous lists selection; *ii)* the list relevance measure proposed allow us to select lists that help DF to perform better; *ii)* the fusion methods considered tend to perform better when few lists are used; *ii)* we need to address the case where there is an empty intersection.

6 References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval, Addison Wesley, 1999.
2. Diamond, T., Liddy, E.D.: Dynamic data fusion. In: Proceedings of the TIPSTER Text Program: Phase III. Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, Maryland, USA, 1998.
3. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Proceedings of The Second Text REtrieval Conference, TREC-2, 1994.
4. Gopalan, N.P., Batri, K.: Adaptive Selection of Top-m Retrieval Strategies for Data Fusion in Information Retrieval. In: International Journal of Soft Computing, 2(1):11-16, 2007.
5. Hsu, D.F., Taksa, I.: Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval. In: Information Retrieval 8(3):449-480, 2005.
6. Ng, K.B., Kantor, P.B.: Predicting the effectiveness of naive data fusion on the basis of system characteristics. In: Journal of American Society for Information Science, 51:1177–1189, 2000.
7. Vogt, C. C., Cottrell, G. W.: Predicting the performance of linearly combined IR systems. In Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.
8. Wu, S., McClean, S.: Performance prediction of data fusion for information retrieval. In: Information Processing and Management, 42(4):899–915, 2006.