Selecting the N-Top Retrieval Result Lists for an Effective Data Fusion

Antonio Juárez-González[1], Manuel Montes-y-Gómez[1], Luis Villaseñor-Pineda[1],
David Pinto-Avendaño[2] and Manuel Pérez-Coutiño[3]

[1] Laboratory of Language Technologies,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{antjug, mmontesg, villasen}@inaoep.mx
[2] Faculty of Computer Science,
Autonomous University of Puebla (BUAP), Mexico.
dpinto@cs.buap.mx
[3] Vanguard Engineering Puebla (VEng), Mexico.
mapco@v-eng.com

Abstract. Although the application of data fusion in information retrieval has yielded good results in the majority of the cases, it has been noticed that its achievement is dependent on the quality of the input result lists. In order to tackle this problem, in this paper we explore the combination of only the *n*-top result lists as an alternative to the fusion of all available data. In particular, we describe a heuristic measure based on redundancy and ranking information to evaluate the quality of each result list, and, consequently, to select the presumably *n*-best lists per query. Preliminary results in four IR test collections, containing a total of 266 queries, and employing three different DF methods are encouraging. They indicate that the proposed approach could significantly outperform the results achieved by fusion all available lists, showing improvements in mean average precision of 10.7%, 3.7% and 18.8% when it was used along with Maximum RSV, CombMNZ and Fuzzy Borda methods.

1 Introduction

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 several issues such as the preprocessing process, the data representation, the weighting scheme and the similarity measure [3]. Evaluation exercises (see for instance [1, 23]) have evidenced that there is not a leading IR approach from all this variety, and, furthermore, that the performance of IR is highly affected by the nature and complexity of collections and queries. In particular, they have shown that different methods may achieve the best performance scores for different queries as well as they may retrieve distinct relevant documents for the same query.

The above situation explains why data fusion (DF), which goal is to enhance the retrieval results by taking advantage of the strengths of several methods, has become one of the most used strategies in IR. Particularly, the last two decades have witnessed a lot of work concerning the design and development of different DF methods specially suited for IR tasks [4, 8, 12, 13, 16, 21].

Although the application of DF in IR has yielded good results in the majority of the cases, it has been noticed that its achievement is dependent on the quality of the input result lists [5, 8, 17, 22, 24]. This dependence is mainly because the widespread use of DF consists in combining all available results lists obtained for a specific query without considering any information about them. Evidently, under this scenario, the presence of some poor-quality lists (containing very few relevant documents) may cause a significant drop in the fusion performance.

In order to tackle the above problem, in this paper we consider the combination of only the *n*-best result lists as an alternative to the fusion of all available data. In particular, we describe a heuristic measure to evaluate the quality of each result list, and, consequently, to select the presumably *n*-top lists per query. The proposed measure attempts to estimate the quality of result lists based on the assumption that a document occurring in several lists has more probability for being relevant, and, therefore, that the lists containing the major number of likely relevant documents at the very first positions are the ones more suitable for being combined.

Preliminary results in four data sets, considering a total of 266 queries, and employing three different DF methods are encouraging. They indicate that in scenarios including lists of diverse qualities, the proposed approach could significantly outperform the results achieved by fusion all result lists, showing improvements in mean average precision that range from 6% to 62.2%.

The rest of the paper is organized as follows. Section 2 describes the related work in DF applied to IR. It mainly discusses some efforts regarding the improvement of DF results. Section 3 introduces the method proposed for estimating the quality of the result lists and for their subsequent selection. Section 4 describes the experimental setup, whereas, Section 5 shows the results regarding the fusion of only the *n*-top result lists per query, obtained using four different data sets. Finally, Section 6 presents our conclusions and exposes further research directions.

2 Related Work

Broadly speaking, data fusion (DF) is the process of combining information gathered by multiple agents (sources, schemes, sensors or systems) into a single representation or result [10]. In IR it has been used to combine results from several retrieval approaches into a "better" single result list. In particular, in this area DF methods differ one from another in the way they compute the final score of documents. Some methods directly use the retrieval status values of the documents across the lists [4, 11, 21], other consider their rank [8, 13], and others their probability of occurring in a predefined segment of the lists [12, 14]. In addition, some recent methods are based on the Social Choice Theory [16, 18], and use pair wise contests of documents to determine their final score.

The application of DF in IR has shown relevant results in the majority of the cases; nevertheless, it has been noticed that it is sensitive to several factors. On the one hand, its performance is affected by the quality of the input lists, and, on the other hand, the selection of the appropriate DF method depends on characteristics such as the redundancy and complementarity of the lists.

Regarding these problems, Gopalan and Batri [9] proposed a supervised method for selecting the m-best retrieval approaches and the best DF method for a given target document collection, and Diamond and Liddy [5] introduced the idea of learning a different linear weighted fusion function for each query instead of using the same static function to all queries.

More recently, some works have focused on investigating the feasibility of predicting the performance of the fusion of a given set of result lists [17, 22, 24]. To some extent, they have demonstrated that an appropriate selection of the input lists may result in a significant improvement of the DF process. However, given that these works consider the relevance judgments as central information for their predictions, they can only be considered as insightful studies about this phenomenon, but cannot be applied as automatic selection procedures.

Supported on the results of these studies, in this paper we consider the combination of only the n-top result lists as an alternative to the fusion of all available data, and, going a step forward, we propose an unsupervised method for selecting the presumably n-best lists per query. The major differences of our method in comparison to previous approaches are that it considers the selection the n-best lists for each individual query, and it does not depend on user relevance judgments nor on a priori information about the used IR methods.

3 Selecting the N-Top Result Lists

As we previously mentioned, the performance of DF is commonly affected by the quality of the input lists. Motivated by this situation, in this paper we explore the idea of combining only the n-top result lists as an alternative to the fusion of all available data. Under this proposal, the major problem is the selection of the n-top result lists for each query, which can be defined as the problem of determining the set of lists having the greatest relevance values in accordance to a specified measure.

More formally, given a set of m result lists $R = \{L_1, L_2, \dots, L_m\}$, where L_i indicates a list of documents (i.e., $L_i = \{d_1, d_1, \dots, d_{|L_i|}\}$), and a relevance measure Q, the problem of selecting the n-top result lists consists in identifying the set of n lists $T \subset R$ with the greatest relevance values, such that:

$$\forall (L_i \in T, L_j \notin T) \ Q(L_i) > Q(L_j) \tag{1}$$

Due to our intention about developing a fusion strategy that does not depend on the user relevance judgments nor consider information of the IR methods, we decided to design a measure that evaluates the relevance of the lists according to their inter-similarities, by using information about the redundancy and ranking of documents across them. In particular, we relied on the idea that the relevance of a list must be incremented by the presence of common documents at the very first positions.

Formula 2 shows the proposed relevance measure, where $q(d_k, L_i)$ denotes the contribution of document d_k to the relevance or quality of list L_i, and $r(d_k, L_i)$ indicates the position (rank) of d_k in the list L_i.

$$Q(L_i) = \sum_{\forall d_k \in I} q(d_k, L_i) \tag{2}$$

$$q(d_k, L_i) = 1 - \frac{ln(r(d_k, L_i))}{ln(|L_i|)} \tag{3}$$

It is important to comment that our first attempt to measure the value of q was $q(d_k, L_i) = 1/r(d_k, L_i)$. Nevertheless, using this direct formula was not possible to achieve satisfactory results, since it severely castigated the contribution of most documents to the global relevance value. In order to reduce the enormous differences in the values of consecutive documents in the lists, especially at the very first positions, we modified this formula by including a smoothing factor as showed in Formula 3. With this modification the values of the first five documents are 1, 0.9, 0.85, 0.8 and 0.77 respectively[1], instead of 1, 0.5, 0.33, 0.25 and 0.2.

Section 5 presents the DF results achieved in four different data sets when the proposed measure was used to select the *n*-top result list for each query.

4 Experimental Setup

In order to evaluate the proposed DF approach, we used four different data sets from the CLEF[2]. In particular, we considered a total of 189,477 documents, 266 queries, and three different DF methods. The following sections give further details about these data sets and the used evaluation measure.

Table 1. Data sets used in our experiments

Data set	**Queries**	**Supported Queries**	**Number of Documents**	**Relevant docs per query** (*average*)
Ad-hocCLEF	50	50	169,477	39
GeoCLEF	25	24	169,477	26
ImageCLEF	39	39	20,000	60
RobustCLEF	160	153	169,477	28

[1] This values were calculated under the assumption that $|L_i| = 1000$.

[2] Cross-Language Evaluation Forum (*www.clef-campaign.org*).

4.1 Data Sets and Result Lists

We used four data sets corresponding to the following CLEF tracks: 2005 Ad-hoc English retrieval [6], 2008 Geographic IR [15], 2008 Image Retrieval [2], and 2008 Robust IR [1]. Table 1 describes some data about these collections. It is important to clarify that in the experiments we only considered the set of supported queries, that is, the queries that have at least one relevant document in the reference collection.

Given that our goal was to evaluate the DF process, we consider five retrieval result lists per query for each data set. In all cases, five different retrieval systems were used to retrieve the result lists. In particular, for the GeoCLEF data set, we used some IR systems developed in [20], which differ one from another in the use of different relevance feedback and ranking refinement techniques. For the ImageCLEF data set, the result lists were retrieved using different combinations of visual and textual features [7]. Finally, for the ad-hoc English track and RobustCLEF data sets, we used five distinct retrieval strategies implemented in the Lemur IR toolkit[3]; these strategies considered different retrieval models as well as different weighting schemes, such as the vector space model and the probabilistic model with Boolean and Frequency-based weightings.

4.2 Data Fusion Methods

In order to obtain general conclusions about the proposed method, we considered three different DF methods: *Maximum RSV* (from linear combination methods), *CombMNZ* (from positional methods), and *Fuzzy Borda Count* (from social choice theory methods). We did not consider probabilistic-based fusion methods because they imply a previous training and our approach is aimed to be fully unsupervised.

Following we present a brief description of the used DF methods. For more details on linear combination fusion methods refer to [4, 11, 21], on CombMNZ go to [8, 13], and on Fuzzy Borda consult [18].

4.2.1 Maximum RSV

This method sorts all documents in the lists by their normalized retrieval status value (RSV), computed independently from each IR system. In the case of repeated documents, the one with the highest value is considered for the final list.

Formally, let $R = \{L_1, L_2, \dots, L_m\}$ be the set of m result lists, $L_i = \{d_1, d_1, \dots, d_{|L_i|}\}$ a list of retrieved documents, and $D = \bigcup_i L_i$ the set of all different documents in the lists. Then, the final score for each document $d_k \in D$ is computed as defined in (4), where $v(d_k, L_i)$ is the normalized RSV of document d_k in the list L_i.

$$MaxRSV(d_k) = \max_{\forall L_i \ni d_k} \left(v(d_k, L_i) \right) \qquad (4)$$

[3] www.lemurproject.org

4.2.2 CombMNZ

Using the same notation from the previous section, this DF method sorts documents from *D* in decreasing order according to the following score.

$$combMNZ(d_k) = \left(\sum_{\forall(L_i \in R)} e(d_k, L_i)\right)\left(\sum_{\forall(L_i \in R, L_i \ni d_k)} |L_i| - r(d_k, L_i) + 1\right) \quad (5)$$

$$e(d_k, L_i) = \begin{cases} 1 & \text{if } L_i \ni d_k \\ 0 & \text{if } L_i \not\ni d_k \end{cases}$$

where $e(d_k, L_i)$ indicates the existence of document d_k in the list L_i, and $r(d_k, L_i)$ its rank in the list.

4.2.3 Fuzzy Borda Count

This DF method considers the set of lists *R* as a set of experts that establish their preference for different alternatives (i.e., documents) by means of pairwise contests. It mainly sorts documents from *D* in decreasing order according to the following score:

$$FuzzyBorda(d_k) = \sum_{\forall(L_i \in R, L_i \ni d_k)} p(d_k, L_i) \quad (6)$$

$$p(d_k, L_i) = \sum_{\forall d_j \in L_i} c_{L_i}(d_k, d_j)$$

$$c_{L_i}(d_k, d_j) = \begin{cases} \dfrac{v(d_k, L_i)}{v(d_k, L_i) + v(d_j, L_i)} & \text{if } v(d_k, L_i) \geq v(d_j, L_i) \\ 0 & \text{Otherwise} \end{cases}$$

where $v(d_k, L_i)$ is the normalized retrieval status value of d_k in the list L_i, $c_{L_i}(d_k, d_j)$ indicates how much expert *i* (list L_i in this case) prefers d_k to d_j, $p(d_k, L_i)$ corresponds to the degree of preference of d_k by L_i, and finally, the total score indicates the general preference of d_k by all lists.

4.3 Evaluation Measure

The evaluation of results was carried out using a measure that has demonstrated its pertinence to compare IR systems, namely, the Mean Average Precision (MAP). It is defined as the norm of the average precisions (AveP) obtained for each query. The AveP for a given query is calculated as follows:

$$AveP = \frac{\sum_{r=1}^{m} P(r) \times rel(r)}{n} \tag{7}$$

where *P*(*r*) is the precision at the first *r* documents, *rel*(*r*) is a binary function which indicates if document at position *r* is relevant or not for the query; *n* is the number of relevant documents for the query that exist at the entire document collection; and *m* is the number of relevant documents retrieved. In all the experiments, we computed the MAP taking into account the first 1000 retrieved documents.

In addition, in all experiments we evaluated the statistical significance of results by means of the *paired student's t-test* considering a confidence level $\alpha = 0.05$, which is extendedly used in IR tasks [19].

5 Experimental Results

5.1 Baseline Results

As we previously mentioned, the traditional DF approach consists in combining all available results lists obtained for a specific query without considering any information about them. Based on this fact, Table 2 presents the MAP results corresponding to the combination of the entire set of five result lists per query, using three different DF methods and four different data sets. In general, these results indicate that methods taking advantage of the document's redundancies, such as CombMNZ and Fuzzy Borda, are more robust than the ones based on information complementarities, such as MaxRSV.

In addition, the last row of Table 2 shows the average performance rate (i.e., the average MAP results) from the five individual IR methods considered for fusion. The comparison of these results against those from DF reveals that in many cases, but not all, DF results are higher. This is an important result since it indicates that in a real IR scenario, where there is no a priori information about the available IR methods, it is a better alternative to apply a DF method, particularly the CombMNZ method, than randomly select one IR method.

Table 2. Baseline results obtained by combining all results lists

DF Method	**Ad hoc 2005**	**GeoCLEF 2008**	**ImageCLEF 2008**	**RobustCLEF 2008**
MaxRSV	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
IR systems ***Average Performance***	0.250	0.233	0.238	0.265

5.2 Results of the Proposed Approach

The proposal of this paper is the combination of only the n-top result lists per query. Therefore, our experiments were designed to confirm the hypothesis that the combination of only the presumably n-best list per query (determined by a proposed heuristic quality measure) allows achieving better results than the combination of all available data. In order to carry out these experiments we proceed as follows:

1. Calculate the quality value (Q) for each one of the given result lists as described by Formula 2.
2. Select the set of n list having the greatest quality values. In particular, given that our interest was to combine the selected set of lists, we considered the following cases: $2 \leq n < |R|$.
3. Perform the DF process using the three methods, namely, Maximum RSV, CombMNZ and Fuzzy Borda.

The results from these experiments are shown in Tables 3 to 5. These tables also include in the last row the baseline results obtained by the combination of all lists (traditional DF approach). In them, the numbers in bold indicate that our method could outperform the baseline results, and the asterisks (*) next to the MAP scores indicate that the achieved improvement was statistically significant.

Table 3. Data fusion results using the *n*-top lists and the Maximum RSV method

Number of selected lists	Ad hoc 2005	GeoCLEF 2008	ImageCLEF 2008	RobustCLEF 2008
n = 2	**0.245***	**0.214**	**0.310***	**0.288***
n = 3	0.229	**0.188**	**0.303***	**0.263***
n = 4	0.225	0.177	**0.287***	**0.246***
Combining all lists	0.231	0.180	0.251	0.231

Table 4. Data fusion results using the *n*-top lists and the CombMNZ method

Number of selected lists	Ad hoc 2005	GeoCLEF 2008	ImageCLEF 2008	RobustCLEF 2008
n = 2	**0.300***	0.233	**0.333***	0.334
n = 3	**0.281**	**0.274***	**0.340***	0.328
n = 4	0.274	**0.261***	**0.323***	0.324
Combining all lists	0.275	0.244	0.302	0.341

Table 5. Data fusion results using the *n*-top lists and the Fuzzy Borda method

Number of selected lists	Ad hoc 2005	GeoCLEF 2008	ImageCLEF 2008	RobustCLEF 2008
n = 2	**0.295***	**0.266**	**0.341***	**0.271***
n = 3	**0.285***	**0.288***	**0.345***	**0.261***
n = 4	**0.278***	**0.286***	**0.335**	**0.223***
Combining all lists	0.267	0.251	0.321	0.167

In general, we consider that these results are encouraging, because they show that in all cases, except one configuration, the proposed method could outperform the baseline results. This behavior was particularly clear for the ImageCLEF data set, where we obtained very good results using all DF methods and considering any number of lists. We believe this was because this dataset contains more relevant documents per query (60 as showed in Table 1) than the other three collections.

On the other hand, we cannot formulate a definitive conclusion about the adequate value of *n*, since its selection depends on the used DF method and on the characteristics of the target document collection. However, from the results, it is possible to observe that *n* = 3 and *n* = 2 tended to generate the best results, indicating somehow that it is better to select the best lists than eliminate the worst(s).

6 Conclusions and Future Work

This paper proposed a new DF approach based on the combination of only the *n*-top result lists per query as an alternative to the fusion of all available data. The selection of the top result lists relies on an unsupervised quality measure that uses information about the redundancy and ranking of the documents across the lists. This approach differs from previous proposals in that it does not depend on any a priori knowledge about the IR methods nor on the user relevance judgments.

The evaluation results in four IR test collections, considering a total of 266 queries, and employing three different DF methods are encouraging. They indicate that the proposed approach could significantly outperform the results achieved by fusion all result lists, considering the MAP scores. They also show that the approach may be successfully used in conjunction with several DF methods given that it could achieve average improvements of 10.7%, 3.7% and 18.8% when was used along with Maximum RSV, CombMNZ and Fuzzy Borda respectively. In addition, we could observe relevant results with several data sets of different characteristics, obtaining average improvements over the baseline of 3.9%, 7.9%, 11.8% and 20.7% for the Ad-hoc, GeoCLEF, ImageCLEF and RobustCLEF collections.

Finally, supported by the presented experimental results, we plan to focus our future work in two main issues. On the one hand, the selection of the most appropriate DF method for a given set of lists and, on the other hand, the dynamic choice of the value of *n* (number of lists to be combined) based on the redundancy and complementarity characteristics of the given result lists.

Acknowledgments. This work was done under partial support of CONACYT (project grants 83459 and 82050, and scholarship 165499). We would also like to thank the CLEF organizing committee for the resources provided.

References

1. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C. CLEF 2008: Ad Hoc Track Overview. In: Working Notes for the CLEF 2008 Workshop. Aarhus, Denmark, 2008.
2. Arni, T., Clough, P., Sanderson, M., Grubinger, M. Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task. In: Working Notes for the CLEF 2008 Workshop. Aarhus, Denmark, 2008.

3. Baeza-Yates, R., Ribeiro-Neto, B. Modern Information Retrieval. Addison Wesley, 1999.
4. Bartell, B.T., Cottrell, G.W., Belew, R.K. Automatic Combination of Multiple Ranked Retrieval Systems. In: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 1994.
5. 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.
6. Di Nunzio, G.M., Ferro, N., Jones, G.J.F., Peters, C. CLEF 2005: Ad Hoc Track Overview. In: Working Notes for the CLEF 2005 Workshop. Vienna, Austria, 2005.
7. Escalante, H.J., González, J.A., Hernández, C.A., López, A., Montes, M., Morales, E., Sucar, L.E., Villaseñor, L. TIA-INAOE's Participation at ImageCLEF 2008. In: Working Notes for the CLEF 2008 Workshop. Aarhus, Denmark, 2008.
8. Fox, E.A., Shaw, J.A. Combination of Multiple Searches. In: Proceedings of The Second Text REtrieval Conference (TREC-2). Gaithersburg, Maryland, USA, 1994.
9. 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.
10. 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.
11. Kantor, P. B. Decision level data fusión for routing of documents in the TREC3 context: A best case analysis of worst case results. In Proceedings of The Third Text REtrieval Conference (TREC-3). Gaithersburg, Maryland, USA, 1995.
12. Lebanon, G., Lafferty, J. Cranking: Combining rankings using conditional probability models on permutations. In Proceedings of the Nineteenth International Conference on Machine Learning. Sydney, Australia, 2002.
13. Lee, J.H. Analyses of Multiple Evidence Combination. In: Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Philadelphia, PA, USA, 1997.
14. Lillis, D., Toolan, F., Collier, R., Dunnion, J. Probfuse: A probabilistic approach to data fusion. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, 2006.
15. Mandl, T., Carvalho, P., Gey, F., Larson, R., Santos, D., Womser-Hacker, C. GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In: Working Notes for the CLEF 2008 Workshop. Aarhus, Denmark, 2008.
16. Montague, M., Aslam, J.A. Condorcet fusion for improved retrieval. In Proceedings of the 11th International Conference on Information Knowledge and Management (CIKM-ACM). McLean, VA, USA, 2002.
17. 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.
18. Perea J.M., Ureña L.A., Buscaldi D., Rosso P. TextMESS at GeoCLEF 2008: Result Merging with Fuzzy Borda Ranking. In: Working Notes for the CLEF 2008 Workshop. Aarhus, Denmark, 2008.
19. Smucker, M.D., Allan, J., Carterette, B. Agreement Among Statistical Significance Tests for Information Retrieval Evaluation at Varying Sample Sizes. In: Poster session for The 32nd Annual ACM SIGIR Conference (SIGIR 09). Boston, MA, USA, 2009.
20. Villatoro-Tello, E., Montes-y-Gómez, M., Villaseñor-Pineda, L. INAOE at GeoCLEF 2008: A Ranking Approach based on Sample Documents. In: Working Notes for the CLEF 2008 Workshop. Aarhus, Denmark, 2008.
21. Vogt, C., Cottrell, G., Belew, R., Bartell, B. Using relevance to train a linear mixture of experts. In Proceedings of The Fifth Text REtrieval Conference (TREC-6). Gaithersburg, Maryland, 1997.

22. 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.
23. Vorhees, E.M. Overview of TREC 2007. In: Proceedings of the sixteenth Text Retrieval Conference (TREC 2007). Gaithersburg, Maryland, USA, 2007.
24. Wu, S., McClean, S. Performance prediction of data fusion for information retrieval. In: Information Processing and Management, 42(4):899–915, 2006.