On the Selection of the Best Retrieval Result per Query – An Alternative Approach to Data Fusion

Juárez-González, Alejandro; Móntes-y-Gómez, Manuel; Villaseñor-Pineda, Luis; Ortiz-Arroyo, Daniel

Published in:
Proceedings of International Conference on Flexible Question Answering Systems 2009

Publication date:
2009

Document Version
Early version, also known as pre-print

On the Selection of the Best Retrieval Result Per Query –An Alternative Approach to Data Fusion–

Antonio Juárez-González[1], Manuel Montes-y-Gómez[1], Luis Villaseñor-Pineda[1], and Daniel Ortíz-Arroyo[2]

[1] Laboratory of Language Technologies, Department of Computational Sciences, National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico
{antjug,mmontesg,villasen}@inaoep.mx
[2] Department of Computer Science and Engineering, Aalborg University, Esbjerg, Denmark
do@cs.aaue.dk

Abstract. Some recent works have shown that the "perfect" selection of the best IR system per query could lead to a significant improvement on the retrieval performance. Motivated by this fact, in this paper we focus on the automatic selection of the best retrieval result from a given set of results lists generated by different IR systems. In particular, we propose five heuristic measures for evaluating the relative relevance of each result list, which take into account the redundancy and ranking of documents across the lists. Preliminary results in three different data sets, and considering 216 queries, are encouraging. They show that the proposed approach could slightly outperform the results from the best individual IR system in two out of three collections, but that it could significantly improve the average results of individual systems from all data sets. In addition, the achieved results indicate that our approach is a competitive alternative to traditional data fusion methods.

1 Introduction

The great amount of available digital content has motivated the development of several information retrieval (IR) systems, which help users to locate useful documents for specific information needs. All these systems differ one from another in various issues, such as the preprocessing process, the data representation, the weighting scheme as well as the similarity measure [3].

Recent evaluations [18, 23] have yielded some interesting findings. Their results evidence that there is not a leading IR method, and, on the contrary, that most existing systems are complementary. They mainly show that different systems could achieve the best performance for different queries and, at the same time, that different systems could retrieve distinct relevant documents for each particular query.

In relation to these findings, Kompaoré and Mothe [12] demonstrated that the "perfect" selection of the best IR system for each query could lead to a significant improvement on the retrieval performance. We confirmed this fact by an experiment

T. Andreasen et al. (Eds.): FQAS 2009, LNAI 5822, pp. 111–121, 2009.

considering three document collections and five different IR systems per collection[1]. Table 1 shows the mean average precision (MAP) results from this experiment, which clearly indicate that the selection of the best IR system per query is a better alternative than the use of one single system for all queries.

Table 1. Improvement on the retrieval performance by selecting the best IR system per query

Data set	MAP (From best global system)	MAP (Using best system per query)	Percentage of Improvement
GeoCLEF	0.263	0.332	26.0%
ImageCLEF	0.292	0.373	27.6%
RobustCLEF	0.359	0.390	8.6%

Motivated by these results, in this paper we propose an automatic approach for the selection of the best retrieval system for each given query. In particular, we tackle this problem from a posteriori perspective; that is, we attempt to select the *best retrieval result* from a given a set of results lists generated by distinct IR systems. For this purpose, we define five different *heuristic measures* to evaluate the relative relevance of each result list. These measures are mainly supported on the idea that a document occurring in several result lists has more probability for being relevant, and, therefore, that the list containing the major number of likely relevant documents at the very first positions is the one with the greatest probability for being the best retrieval result. Thanks to this solution perspective, the proposed approach is independent from the internal processes carried out at the IR stage, and, therefore, it is versatile enough to work with very different IR systems.

Preliminary results in three different data sets, and considering 216 queries, are encouraging. They show that the proposed approach could slightly outperform the results from the best individual IR system in two out of three collections, but that it could significantly improve the average results from all data sets. In addition, they also indicate that our approach is a competitive alternative to traditional data fusion methods, which aim is to combine a set of result lists into a –better– single retrieval result.

The rest of the paper is organized as follows. Section 2 discusses some related work about IR using several retrieval systems. Section 3 introduces the proposed approach and describes the heuristic measures used for evaluating the relative relevance of each result list. Section 4 shows the experimental results on three different data sets from the CLEF[2]. Finally, Section 5 presents our conclusions and discusses some ideas for future work.

2 Related Work

The existence of several IR systems has motivated the design of different methods for handling their combination. The purpose of this kind of methods is to improve the

[1] Section 4 gives further details about the collections, queries, IR systems, and evaluation measures used in this experiment.

[2] In particular, we considered the collections from the Geographic, Image and Robust IR tracks from the 2008 edition of the Cross-Language Evaluation Forum (*www.clef-campaign.org*).

retrieval performance by taking advantage from the strengths of different systems. In general, these methods can be clustered in two main approaches:

Data fusion. Its idea is to combine results from several IR systems into a –better– single result list [4, 5, 9, 11, 13, 19]. Methods from this approach are mainly unsupervised and are supported on two basic assumptions: first, relevant documents tend to occur in several result lists (known as chorus effect), and second, relevant documents tend to be ranked at the very first positions (known as skimming effect). Recent research using this approach has mainly concentrated on: (*i*) fusion for multimedia and multilingual retrieval [7, 11, 15]; (*ii*) the automatic selection of the result lists to include into the fusion process [8, 16, 22, 25]; and (*iii*) the choice of the most appropriate fusion method for a particular situation [6].

Selection of the best retrieval system. Methods from this approach focus on two different problems: on the one hand, the selection of the best retrieval system for each particular query [10, 12], and, on the other hand, the identification of the best global system for a given set of queries [17, 20, 24]. The former tend to use supervised techniques in order to learn a mapping between (kinds of) queries and systems, whereas, the later are mainly based on unsupervised techniques that take advantage of the redundancies across different result lists.

The method proposed in this paper focuses on a problem close to the selection of the best retrieval system, namely, the *selection of the best result list* for each particular query. Different to previous methods [10, 12], which rely on a supervised approach or require the participation of a user, our method is based on an automatic *unsupervised approach* that rank the result lists taking into consideration their relative relevance. In particular, we propose five different *heuristic measures* to evaluate the relative relevance of each result list. These measures recover some ideas from data fusion by including information about the *redundancy and ranking of documents* from each result list; nevertheless, in this case, we use this information to evaluate and select the lists and not as a criterion for their combination.

3 Selecting the Best Result List

Having n-different IR systems, it is possible to retrieve n-different result list for each given query. Therefore, under this scenario, the problem of selecting the best result list can be defined as the problem of determining the list that maximizes some specified relevance measure.

More formally, given a set of result lists $R = \{L_1, L_2, \ldots, L_n\}$, where L_i indicates an ordered list of documents (i.e., $L_i = <d_1, d_2, \ldots, d_m>$), and a relevance measure Q, the problem of selecting the best result list consists in identifying the list L_i such that:

$$Q(L_i, R) \geq Q\big(L_j, R\big), \forall j \neq i \tag{1}$$

The following section presents some heuristic measures for evaluating the relative relevance of each result list. As we mentioned, these measures are supported on the idea that a document occurring in several result lists has more probability for being relevant, and that the list containing the major number of likely relevant documents at the very first positions is the one with the greatest probability for being the best retrieval result.

In other words, they attempt to take advantage of the same effects considered for data fusion, namely, the chorus and skimming effects [22]. In particular, we have proposed five different measures that combine these effects in a slightly different way.

3.1 Heuristic Relevance Measures

First relevance measure. This measure only considers the chorus effect; it is based on the assumption that the relevance of a document is proportional to the number of lists that include it, and, therefore, that the relevance of a result list depends on how much it is intersected with the rest of the lists. This measure is computed as follows.

$$Q_1(L_i, R) = \sum_{\forall L_k \in R} |L_i \cap L_k| \tag{2}$$

Second relevance measure. It combines information about the redundancy and ranking of documents across the set of result lists (i.e., the chorus and skimming effects respectively). It mainly looks at the positions of the documents from the intersection of the lists. The idea behind this measure is that the relevance of a list increments by the presence of common documents at the very first positions. Considering that I represents the set of documents from the intersection of all result lists, and that $p(d_k,L_i)$ indicates the position of the document d_k in the list L_i, this measure is calculated as follows:

$$Q_2(L_i, R) = \sum_{\forall d_k \in I} \left(\frac{1}{p(d_k, L_i)} \right) \tag{3}$$

Third relevance measure. It is very similar to Q_2; nevertheless, in this case it emphasizes the punishment to final documents instead of the reward to the top documents. Q_3 is defined as follows:

$$Q_3(L_i, R) = \frac{1}{\sum_{\forall d_k \in I} p(d_k, L_i)} \tag{4}$$

Fourth relevance measure. It modifies the way that rank information is used in Q_2. It mainly introduces a smoothing factor that allows reducing the enormous differences in the values of contiguous documents, especially at the very first positions. This measure is calculated as follows:

$$Q_4(L_i, R) = \sum_{\forall d_k \in I} \tilde{p}(d_k, L_i) \tag{5}$$

$$\tilde{p}(d_k, L_i) = 1 - \frac{ln(p(d_k, L_i))}{ln(|L_i|)} \tag{6}$$

Fifth relevance measure. Following the same idea than Q_4, this measure modifies the way that rank information is used in Q_3. It mainly introduces a smoothing factor (refer to formula 6) that allows reducing the enormous differences in the values of contiguous documents, especially at the very last positions of the result lists. This measure is computed as follows:

$$Q_5(L_i, R) = \frac{1}{\sum_{\forall d_k \in I} (1/\tilde{p}(d_k, L_i))} \tag{7}$$

4 Experimental Setup

In order to evaluate the proposed approach, we used three different data sets from the CLEF 2008. In particular, we considered a total of 189,477 documents, 216 queries, and five different results lists per query. The following sections give further details about these data sets and the used evaluation measures.

4.1 Data Sets

We used three data sets from the CLEF-2008: one for evaluating Geographic IR [14], other for evaluating Image Retrieval [2], and another for evaluating Robust IR [1]. Table 2 describes some data about these collections.

Table 2. Data sets used in our experiments

Data set	Queries	Supported Queries	Number of Documents
GeoCLEF	25	24	169,477
ImageCLEF	39	39	20,000
RobustCLEF	160	153	169,477

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. In addition, we have only used the title and description parts of these queries[3]. Table 3 shows a query corresponding to the RobustCLEF collection.

Table 3. An example query from the RobustCLEF-2008 task

<title>Japanese Rice Imports </title>
<description>Find documents discussing reasons for and consequences of the first imported rice in Japan. </description>
<narrative>In 1994, Japan decided to open the national rice market for the first time to other countries. Relevant documents will comment on this question. The discussion can include the names of the countries from which the rice is imported, the types of rice, and the controversy that this decision prompted in Japan. </narrative>

4.2 Evaluation Measures

The evaluation of results was carried out using two measures that have demonstrated their pertinence to compare IR systems, namely, the Mean Average Precision (*MAP*) and the *R-prec*. The *MAP* is defined as the norm of the average precisions (*AveP*) obtained for each query. The *AveP* for a given query q is calculated as follows:

$$AveP = \frac{\sum_{r=1}^{m} P(r) \times rel(r)}{n} \quad (8)$$

[3] In CLEF, queries are commonly described by a title, a description, and a narrative.

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 q; n is the number of relevant documents for q that exist at the entire document collection; and m is the number of relevant documents retrieved for q. In all the experiments, we computed the *MAP* taking into account the first 1000 retrieved documents.

On the other hand, *R-prec* is defined as the precision reached after R documents have been retrieved, where R indicates the number of relevant documents for q that exist in the entire document collection.

4.3 IR Systems and Baseline Results

As we described in Section 3, the application of the proposed approach relies on the availability of several result lists. In particular, for the experiments, we considered five result lists per query. For the GeoCLEF data set, we used some IR systems developed in [21], 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 RobustCLEF data, we used five distinct retrieval strategies implemented in the Lemur IR toolkit[4]; these strategies considered different retrieval models (like the vector space model, and the probabilistic model) as well as different weighting schemes.

Tables 4-6 show the overall *MAP* and *R-prec* values for all result lists from each data set. Numbers in bold correspond to the best global individual system, that is, to the system with the highest *MAP* for all queries from the given data set. It is important to point out that these tables exclude details from the used IR systems since our relevance measures do not depend on any information about them.

Table 4. GeoCLEF collection**:** *MAP* and *R-prec* from input IR systems

IR system ID	MAP	Average R-prec
Geo_1	0.218	0.209
Geo_2	0.210	0.235
Geo_3	**0.263**	**0.254**
Geo_4	0.248	0.240
Geo_5	0.218	0.239

Table 5. ImageCLEF collection: *MAP* and *R-prec* from input IR systems

IR system ID	MAP	Average R-prec
Image_1	0.278	0.283
Image_2	0.255	0.259
Image_3	0.094	0.122
Image_4	**0.292**	**0.305**
Image_5	0.271	0.289

[4] www.lemurproject.org

Table 6. RobustCLEF collection: *MAP* and *R-prec* from input IR systems

IR system ID	MAP	Average R-prec
Robust_1	**0.359**	**0.346**
Robust_2	0.240	0.240
Robust_3	0.313	0.305
Robust_4	0.218	0.222
Robust_5	0.198	0.194

5 Results

The evaluation of the proposed approach consisted of two main experiments. The first focused on determining the effectiveness of the proposed relevance measures, whereas, the purpose of second was to compare our approach, i.e., the selection of the best retrieval result per query, against traditional data fusion methods. The following sections show the results from these experiments.

5.1 Experiment 1: Evaluating the Relevance Measures

In order to evaluate the effectiveness of proposed relevance measures to select the best retrieval result per query we proceeded as follows:

For each query, first, we retrieved five different result lists (refer to Section 4.3); then, we estimated the relevance of each list by means of a given relevance measure, and, finally, we selected the list with the greatest value as the final response. After this process, we count the number of times where the selected list has equal or higher performance (*MAP*) than the best individual system. Table 7 shows the results from this experiment indicating that, in the majority of the cases, all proposed measures achieved a useful selection.

Table 7. Effectiveness of the proposed relevance measures (the baseline corresponds to the best global individual result from Tables 4-6)

Relevance measure	GeoCLEF		ImageCLEF		RobustCLEF	
	< baseline	>= baseline	< baseline	>= baseline	< baseline	>= baseline
Q_1	8	16	20	19	33	120
Q_2	7	17	14	25	25	128
Q_3	7	17	2	37	24	129
Q_4	6	18	1	38	23	130
Q_5	11	13	5	35	23	130

Additionally, and with the aim of having a global evaluation of the usefulness of proposed approach, we computed the *MAP* and *R-prec* values obtained by the application of proposed measures. Table 8 shows these values as well as the results corresponding to the average and best-individual system performances (refer to Tables 4-6). Results in bold indicate the cases where our approach could improve the performance from the best global individual system.

Table 8. Results of the automatic selection of the best retrieval result per query

Relevance measure	GeoCLEF MAP	GeoCLEF Average R-prec	ImageCLEF MAP	ImageCLEF Average R-prec	RobustCLEF MAP	RobustCLEF Average R-prec
Q_1	**0.267**	**0.259**	0.264	0.279	0.317	0.307
Q_2	0.259	**0.285**	0.259	0.278	0.338	0.329
Q_3	0.248	**0.278**	**0.294**	**0.307**	0.338	0.330
Q_4	0.259	**0.288**	**0.299**	**0.309**	0.338	0.328
Q_5	0.219	0.244	**0.294**	0.304	0.339	0.330
Average	0.231	0.236	0.238	0.251	0.265	0.261
Best	0.263	0.254	0.292	0.305	0.359	0.346

Results from Table 8 are encouraging since they indicate that all proposed relevance measures could outperform the average results of the individual systems from all data sets. This is an important fact since it means that, in a real scenario, where there is not a priori information about the available IR systems, our approach is able to improve the results from a random selection of the retrieval system.

From a different perspective, results from Table 8 are not conclusive since they indicate that the proposed approach could only slightly outperform the results from the best individual IR system in two out of three collections. In particular, the improvement in *MAP* was as higher as 1.3% for the GeoCLEF collection and 2.2% for the ImageCLEF data set, whereas, the improvement in *R-prec* was as higher as 13.3% and 1.4% respectively.

Trying to understand the modest performance of our approach, we achieved a detailed analysis of the set of input result lists, and concluded that the proposed measures were seriously affected by the small number of relevant documents per query that exist in the reference collections; in average, 26 for GeoCLEF, 62 for ImageCLEF and 28 for RobustCLEF.

5.2 Experiment 2: Selecting the Best Retrieval Result vs. Data Fusion

As we previously mentioned, data fusion is the traditional approach for improving the retrieval performance by taking advantage from the strengths of different IR systems. The most commonly used methods of data fusion are the following:

Round Robin. This strategy takes one document in turn from each individual list and alternates them in order to construct the final merged output.

Raw Score Value (RSV). This strategy sorts all documents by their original score, computed independently from each IR system.

CombMNZ. In this strategy, the result scores from each IR system are initially (min-max) normalized. Afterward, the scores of documents occurring in various collections are summed and then multiplied by the number of result lists in which it occurs. For more details refer to Lee et al. (1997).

Table 9 shows the results achieved by these methods as well as the results from the proposed approach using the fourth relevance measure (Q_4), which turned out to be the best performing measure according to results from Table 8. The comparison of these results indicate that our approach is considerably superior to Round Robin and

RSV, and, on the other hand, that it is a competitive alternative to the CombMNZ method, which it is commonly defined as one of the most robust data fusion techniques (Lee, 1997). In this table, numbers in bold indicate the cases where our approach outperformed the results from all data fusion methods.

Table 9. Our approach vs. data fusion methods

Method	GeoCLEF		ImageCLEF		RobustCLEF	
	MAP	Average R-prec	MAP	Average R-prec	MAP	Average R-prec
Our approach (using Q_4)	**0.259**	**0.288**	0.299	**0.309**	0.338	0.328
Round Robin	0.026	0.011	0.058	0.024	0.026	0.020
RSV	0.180	0.197	0.251	0.270	0.231	0.236
CombMNZ	0.244	0.247	0.302	0.304	0.341	0.329

Results from Table 9 suggest that there is not a significant gain to consider using our method instead of CombMNZ. However, a detailed analysis showed us that for the cases where CombMNZ could not outperform the best global individual result (which turned out to be 17/24 queries from GeoCLEF, 24/39 from ImageCLEF and 102/153 from RobustCLEF), our method achieved better results. In particular, Table 10 shows the results from this analysis that indicate that, for these subsets of queries, our approach considerably improved the results from CombMNZ by 17.2%, 16.1% and 14.3% for GeoCLEF, ImageCLEF and RoubustCLEF respectively.

Table 10. Detailed analysis of our approach and CombMNZ

Method	GeoCLEF (24 q.)		ImageCLEF (39 q.)		RobustCLEF (153 q.)	
	Won queries	MAP	Won queries	MAP	Won queries	MAP
Our approach (using Q_4)	**12** (17)	**0.2369**	**17** (24)	**0.3561**	**73** (102)	**0.3651**
CombMNZ	5 (17)	0.2021	7 (24)	0.3066	29 (102)	0.3192

6 Conclusions and Future Work

This paper described an approach for selecting the best retrieval result from a given set of result lists generated by different IR systems. The approach relies on the estimation of the relative relevance of each result list. In particular, we proposed five heuristic measures to evaluate this relevance by taking into account information about the redundancy and ranking of documents from each result list.

The evaluation results allow us to establish the following conclusions:

- The relevance measures considering the chorus and skimming effects tend to be more robust than the measure based only in the chorus effect. In particular, the fourth relevance measure, which includes a smoothing factor, achieved the best results.

- Our approach could only slightly improve the results from the best IR system in two out of three collections. We attribute this unexpected behavior to the small number of relevant documents per query that exist in the reference collections. Somehow, this fact indicates that, for some collections and/or queries, the redundancy and ranking of the items are not as determinant as we initially supposed.
- Our approach could significantly improve the average results of the individual systems from all data sets. From an application perspective, this is an important result, since it indicates that our approach is considerably better than a random selection of the retrieval system.
- Our approach is a competitive alternative to the traditional data fusion approach. It could improve the results from Round Robin and RSV, and achieved similar results than CombMNZ. However, a detailed analysis considering only the subset of queries where CombMNZ could not outperform the best global individual results, allowed us to conclude that our approach is less sensitive to the presence of poor quality results, and, therefore, that it may be considered a more robust strategy than CombMNZ.

As future work we plan to apply the proposed heuristic relevance measures to the problems of: (*i*) selecting the result lists to be include into the fusion process, and (*ii*) choosing the most appropriate fusion method for each particular situation.

Acknowledgments. This work was done under partial support of CONACYT (Project Grant CB-2007-01-83459 and scholarship 165499). We would also like to thank the CLEF organizing committee as well as to the EFE agency 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, Reading (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 (1994)
5. Belkin, N.J., Kantor, P., Fox, E.A., Shaw, J.A.: Combining the evidence of multiple query representations for information retrieval. Information Processing and Management 31(3), 431–448 (1995)
6. Chen, Y., Shahabi, C., Burns, G.A.P.C.: Two-Phase Decision Fusion Based on User Preference. In: The Hawaii International Conference on Computer Sciences, Honolulu, Hawaii (January 2004)
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. Gopalan, N.P., Batri, K.: Adaptive Selection of Top-m Retrieval Strategies for Data Fusion in Information Retrieval. International Journal of Soft Computing 2(1), 11–16 (2007)
9. Hsu, D.F., Taksa, I.: Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval. Information Retrieval 8(3), 449–480 (2005)
10. Hubert, G., Mothe, J.: Relevance Feedback as an Indicator to Select the Best Search Engine - Evaluation on TREC Data. In: Proceedings of the Ninth International Conference on Enterprise Information Systems, ICEIS (2007)
11. Kludas, J., Bruno, E., Marchand-Maillet, S.: Information Fusion in Multimedia Information Retrieval. In: Proceedings of the 5th International Workshop Adaptive Multimedia Retrieval, AMR, pp. 147–159 (2007)
12. Kompaoré, D., Mothe, J., Baccini, A., Dejean, S.: Query clustering to decide the best system to use. In: Proceedings of the RIAO 2007. 8th International Conference (2007)
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 (1997)
14. 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)
15. Martínez-Santiago, F., Ureña-López, L.A., Martín-Valdivia, M.: A merging strategy proposal: The 2-step retrieval status value method. Information Retrieval 9(1), 71–93 (2006)
16. Ng, K.B., Kantor, P.B.: Predicting the effectiveness of naive data fusion on the basis of system characteristics. Journal of American Society for Information Science 51, 1177–1189 (2000)
17. Nuray, R., Can, F.: Automatic ranking of information retrieval systems using data fusion. Information Processing and Management 42(3), 595–614 (2006)
18. Peters, C.: What happened in CLEF 2008 Introduction to the Working Notes. In: Working Notes of the Cross Language Evaluation Forum, CLEF (2008)
19. Shaw, J.A., Fox, E.A.: Combination of Multiple Searches. In: Proceedings of The Second Text REtrieval Conference, TREC, vol. 2 (1994)
20. Spoerri, A.: Using the structure of overlap between search results to rank retrieval systems without relevance judgments. Information Processing and Management 43(4), 1059–1070 (2007)
21. 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)
22. Vogt, C.C., Cottrell, G.W.: Fusion Via a Linear Combination of Scores. Information Retrieval (1), 151–173 (1999)
23. Vorhees, E.M.: Overview of TREC 2007. In: Proceedings of the sixteenth Text Retrieval Conference, TREC (2007)
24. Wu, S., Crestani, F.: Methods for Ranking Information Retrieval Systems Without Relevance Judgments. In: Proceedings of the 2003 ACM Symposium on Applied Computing, pp. 811–816 (2003)
25. Wu, S., McClean, S.: Performance prediction of data fusion for information retrieval. Information Processing and Management 42(4), 899–915 (2006)