Unsupervised Rank Fusion for Diverse Image Metasearch*

José Solenir L. Figuerêdo[1], **Rodrigo Tripodi Calumby**[1]

[1] University of Feira de Santana – Feira de Santana – BA – Brazil

jslfigueredo@ecomp.uefs.br, rtcalumby@uefs.br

Abstract. *For a given query and a set of image ranked lists retrieved from multiple search engines, the metasearch technique aims at combining these lists to build an unified ranking with improved relevance. Rank aggregation is an approach that has been widely used to support this task. This paper investigates the use of rank aggregation methods in the metasearch scenario for diverse image retrieval. Although metasearch systems are usually driven by the relevance of the final result, the impact on diversification has also been analyzed. The experimental findings suggest metasearch based on rank aggregation allows significant improvements, both in terms of relevance and diversity.*

1. Introduction

In recent years, advances in data capture and storage technologies allowed the production of large amounts of digital content, enabling advanced studies in many fields. These collections have been explored in several contexts such as healthcare, biodiversity, social networks, and digital libraries [Bahri et al. 2019]. However, given the user needs and the large amount of information available, effective techniques are demanded to explore these collections. Over time, several methods have been proposed with the goal of generating better results for many information retrieval tasks.

In order to maximize the quality of the search results, the scientific and industrial communities developed robust systems to exploit as much information as possible for determining the relevance of objects in the databases [Calumby et al. 2016]. This allowed the improvement of ranking algorithms and consequently to better meet users' expectations in their search routines. However, given the complexity of the task, different systems tend to give different answers to the same information need of a given user. In this sense, the results achieved by each system tend to be complementary. A proposed solution to this scenario, known as metasearch, is the combination of results obtained from multiple databases or different search systems. This integration can be performed in many ways and applying rank aggregation algorithms is a popular approach for such task [McDonald and Smeaton 2005, Farah and Vanderpooten 2007].

In a more specific scenario, users may not be able to properly express their information need, leading to poorly specified or ambiguous queries [Santos et al. 2015]. Moreover, given a retrieval model considering only relevance maximization, systems eventually produce result lists with objects that are considerably similar (redundant) or do not necessarily include the different aspects from the items that are available in the collection (low coverage). An approach used to tackle these problems is the promotion of diversity into the result set and several algorithms have been proposed to fulfill this

*This work was partially supported by PIBIC/CNPq (grant number 158204/2018-2).

goal [Calumby et al. 2017]. The purpose of such algorithms is to maximize the information gain within the retrieved items and attenuate redundancy whilst responding to different interpretations from the same query. Hence, it allows simultaneously considering the query intents from different users.

Given the aforementioned challenges, this paper describes and investigates the use of rank aggregation methods for metasearch in the diverse image retrieval scenario. Although metasearch systems are usually focused on the relevance of the final result, in this paper we analyze the rank aggregation of diversified rankings and the impact on the relevance and the diversity of the final ranking.

2. Related Work

Rank aggregation algorithms can be divided in: score-based or order-based. In the former, the aggregation function takes as input the ranking scores associated to each object in the original rankings. In the latter, only the ordering among of items is required to perform the aggregation. Many score-based methods have been proposed, e.g., CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ) [Muñoz et al. 2015]. In turn, Borda Count, Median Rank Aggregation, and Reciprocal Rank Fusion are popular order-based methods [Muñoz et al. 2015]. These algorithms have been applied in many contexts including metasearch. However, these methods do not consider diversification explicitly. Nevertheless, some studies have investigated the application of data fusion for diversification tasks. In particular, as stated in [Wu et al. 2019], data fusion is expected to ensure a broader coverage of different types of relevant documents fostering diverse promotion.

In [Liang et al. 2014] the diversification includes three steps. Initially, the aggregation relies on traditional methods: CombSUM and CombMNZ. Hence, an inference is made for latent subtopics. Finally, the result generated by the fusion and topic modeling steps is submitted to diversification. Alternatively, in [Xu and Wu 2017] rather than fusing results that are already diversified, an early fusion approach is applied. It consists in: i) Considering only the relevance, a set of results are generated with algorithms of typical searches; ii) The results are combined with a fusion algorithm such as CombMNZ; and iii) An explicit diversification method is applied, e.g., the xQuAD. The experimental findings indicated that the early fusion strategy was as effective as late fusion ones.

Previous works focused on the analysis of diversification through fusion methods in the context web page retrieval. The investigation of such methods in other multimedia scenarios (e.g., images or videos) are still incipient. Beyond it, the image retrieval imposes additional challenges due to the inherent characteristics of the tasks and the heterogeneity of data collections. This work experimentally investigates the effectiveness of several rank aggregation methods for metasearch in the context of diverse image retrieval.

3. Evaluation Scenario and Experimental Setup

For the experimental evaluation, the collection provided by the *Information Fusion for Social Image Retrieval & Diversification Task* [Ramírez-de-la-Rosa et al. 2018] was used. It includes results from the many image search systems proposed and evaluated between 2013 and 2016 in the *MediaEval Retrieving Diverse Social Images tasks* [Ionescu et al. 2014, Ionescu et al. 2015, Ionescu et al. 2016a,

Ionescu et al. 2016b]. There are ranked results for numerous queries. Moreover, it includes relevant and diverse results with different levels of quality. The dataset is organized in development, validation and test sets (Table 1). The test set was not considered in the experimental evaluation, since ground-truth was not publicly available.

Table 1. Overview of the collection used in the experimental evaluation.

Dataset		# Queries	# Rankings	Topic Category
Devset	devset1	346	39	Single-topic
	devset2	60	56	Single-topic
Validset		139	60	Single/Multi-topic
Testset	seenIR	63	56	Single-topic
	unseenIR	64	29	Multi-topic

Many rank aggregation methods were considered. The score-based methods evaluated were CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ, CombMED, and Multiplication Scores (MScores) [Li et al. 2014]. In turn, the order-based methods Borda Count, Median Rank Aggregation (MRA), and Reciprocal Rank Fusion (RRF) were assessed.

Precision and Cluster-Recall [Zhai et al. 2003] measures were used for effectiveness assessment. Precision represents the quality of the ranking in terms of relevance and Cluster-Recall computes the percentage of conceptual clusters that were represented in the final diversified result. These metrics were computed based on the available ground-truth. For effectiveness analysis these measures were computed for up to the 50th position of the ranking. As the baseline, in addition to the one provided by [Ramírez-de-la-Rosa et al. 2018], we also considered the best input ranking systems.

For the comparison to the baseline, we computed the relative gain of the best aggregation method for each rank depth. The selection of the best systems to be used as baselines relied on the Precision (PR@20) e Cluster-Recall(CR@20). The cutoff at 20 simulates the content of a single page of a typical web image search engine and reflects user behavior, i.e., inspecting the first page of results [Ramírez-de-la-Rosa et al. 2018].

4. Results and Discussion

The comparative analysis against the baselines was performed based on the relative gains at multiple ranking depths. Initially, Table 2 shows the effectiveness of the fusion methods, including the baselines. The highest values are highlighted in boldface. These values are used for the comparison with the baselines. Considering the aggregation methods, RRF was the top performing one. It was not the most frequent top performer for Devset1. However, for the Deveset2 and the Validset it outperformed the other methods for most of the evaluation measures and ranking depths. For the Devset2, the RRF method, besides improving the diversity, did not negatively impact the relevance of the results, which is a desirable behaviour for diverse image retrieval systems.

Table 3 presents the comparison between the fusion methods (taking the highest values) and the baselines. It indicates the relative gains of the fusion methods over the baselines (positive gains are highlighted in bold). In Table 3, *ICPR* stands for the baseline provided with the collection as part ICPR challenge [Ramírez-de-la-Rosa et al. 2018]. In turn, *Best_P* and *Best_CR* represent the best input system considering Precision and Cluster-Recall as the selection criteria, respectively.

Table 2. Results for the rank aggregation methods and baselines. Top values are highlighted in boldface.

Devset1												
Method	P@5	P@10	P@20	P@30	P@40	P@50	CR@5	CR@10	CR@20	CR@30	CR@40	CR@50
Baseline(ICPR)	0.7883	0.7558	0.7289	0.7194	0.7080	0.6877	0.2331	0.3649	0.5346	0.6558	0.7411	0.7988
Baseline(best_P)	0.8947	0.8936	0.8788	0.8569	0.8265	0.7891	0.2047	0.2963	0.4410	0.5476	0.6416	0.7122
Baseline(best_CR)	0.7409	0.7330	0.7487	0.7603	0.7145	0.5915	0.2614	0.4291	0.6314	0.7228	0.7473	0.7484
CombMAX	0.7602	0.7512	0.7512	0.7349	0.7242	0.7031	0.2531	0.4176	0.6069	0.7315	0.8158	0.8639
CombMIN	0.6544	0.6626	0.6744	0.6790	0.6736	0.6582	0.2049	0.3580	0.5455	0.6822	0.7647	0.8201
CombSUM	0.8287	0.8243	0.8034	0.7867	0.7626	0.7315	0.2694	0.4410	0.6255	**0.7437**	**0.8261**	0.8738
CombANZ	0.7573	0.7561	0.7469	0.7347	0.7205	0.6971	0.2500	0.4020	0.5905	0.7234	0.8073	0.8552
CombMED	0.8287	0.8243	0.8034	0.7867	0.7626	0.7315	0.2694	**0.4410**	0.6255	0.7437	0.8261	0.8738
CombMNZ	0.8456	0.8289	**0.8098**	**0.7888**	**0.7645**	**0.7344**	0.2753	0.4370	**0.6258**	0.7412	0.8238	0.8708
MScores	0.8240	0.8228	0.8044	0.7870	0.7624	0.7313	0.2672	0.4383	0.6235	0.7413	0.8235	**0.8745**
Borda Count	0.6129	0.6143	0.6213	0.6256	0.6246	0.6150	0.1884	0.3090	0.4642	0.5883	0.6780	0.7437
RRF	**0.8491**	**0.8316**	0.8092	0.7874	0.7616	0.7314	**0.2781**	0.4327	0.6235	0.7421	0.8203	0.8655
MRA	0.7614	0.7567	0.7361	0.7262	0.7092	0.6872	0.2544	0.4157	0.6168	0.7351	0.8186	0.8690
Devset2												
Baseline(ICPR)	0.8100	0.8067	0.8058	0.8056	0.8025	0.7917	0.1316	0.2135	0.3435	0.4517	0.5364	0.5977
Baseline(best_P)	0.8933	0.8933	0.8550	0.8167	0.8021	0.7850	0.1461	0.2377	0.3809	0.4750	0.5531	0.6210
Baseline(best_CR)	0.8867	0.8600	0.8492	0.8189	0.8017	0.7933	0.1682	0.3003	0.4697	0.5594	0.6274	0.6788
CombMAX	0.7467	0.7417	0.7558	0.7639	0.7333	0.7257	0.1378	0.2534	0.4209	0.5375	0.6186	0.6725
CombMIN	0.4667	0.5000	0.5025	0.5189	0.5312	0.5463	0.0878	0.1663	0.2804	0.3914	0.4790	0.5565
CombSUM	0.8667	0.8550	0.8267	0.8194	0.8075	0.7980	0.1650	0.2861	0.4430	0.5579	0.6390	0.7046
CombANZ	0.4733	0.5433	0.5808	0.5944	0.6096	0.6147	0.0917	0.1882	0.3374	0.4480	0.5427	0.6045
CombMED	0.8667	0.8550	0.8267	0.8194	0.8075	0.7980	0.1650	0.2861	0.4430	0.5579	0.6390	0.7046
CombMNZ	0.8933	0.8650	0.8417	0.8361	**0.8292**	0.8123	0.1685	0.2820	0.4525	**0.5692**	**0.6417**	0.7041
MScores	0.8733	0.8517	0.8308	0.8250	0.8075	0.8013	0.1665	0.2797	0.4433	0.5616	0.6362	0.7018
Borda Count	0.4700	0.4567	0.4750	0.4889	0.4925	0.4993	0.0885	0.1541	0.2694	0.3665	0.4238	0.4948
RRF	**0.8967**	**0.8817**	**0.8508**	**0.8372**	**0.8292**	**0.8260**	**0.1692**	**0.2906**	**0.4586**	0.5676	0.6367	**0.7052**
MRA	0.6633	0.6733	0.6775	0.6733	0.6725	0.6797	0.1287	0.2370	0.4066	0.5185	0.6141	0.6831
Validset												
Baseline(ICPR)	0.7281	0.7086	0.7000	0.6952	0.6838	0.6776	0.1489	0.2402	0.3684	0.4616	0.5284	0.5851
Baseline(best_P)	0.8101	0.8108	0.7906	0.7736	0.7646	0.7534	0.1904	0.2908	0.4051	0.5005	0.5703	0.6246
Baseline(best_CR)	0.7755	0.7633	0.7309	0.7002	0.6899	0.6790	0.1935	0.3163	0.4963	0.6112	0.6933	0.7514
CombMAX	0.7439	0.7209	0.7065	0.7041	0.7007	0.7014	0.1658	0.2704	0.4207	0.5230	0.6084	0.6716
CombMIN	0.5597	0.5813	0.5968	0.5986	0.6031	0.6029	0.1188	0.2019	0.3282	0.4152	0.4876	0.5440
CombSUM	0.7626	0.7554	0.7320	0.7271	0.7214	0.7173	0.1757	0.2812	0.4256	0.5404	0.6246	0.6875
CombANZ	0.6230	0.6201	0.6399	0.6441	0.6550	0.6544	0.1390	0.2274	0.3755	0.4639	0.5436	0.5956
CombMED	0.7626	0.7554	0.7320	0.7271	0.7214	0.7173	0.1757	0.2812	0.4256	0.5404	0.6246	0.6875
CombMNZ	0.7683	0.7662	0.7471	0.7331	0.7243	0.7219	0.1772	0.2893	0.4344	0.5494	0.6326	0.6951
MScores	0.7612	0.7597	0.7367	0.7254	0.7212	0.7168	0.1760	0.2825	0.4298	0.5391	0.6254	0.6861
Borda Count	0.5669	0.5727	0.5770	0.5878	0.5946	0.5994	0.1208	0.2049	0.3161	0.3972	0.4643	0.5191
RRF	**0.7813**	**0.7698**	**0.7536**	**0.7405**	**0.7318**	**0.7281**	**0.1776**	**0.3008**	**0.4570**	**0.5596**	**0.6388**	**0.7025**
MRA	0.6619	0.6590	0.6680	0.6743	0.6761	0.6753	0.1546	0.2641	0.4085	0.5310	0.6119	0.6744

Table 3. Relative gains (%) of top performing methods against the baselines.

Devset1												
	P@5	P@10	P@20	P@30	P@40	P@50	CR@5	CR@10	CR@20	CR@30	CR@40	CR@50
Gain Over ICPR	**7.71**	**10.03**	**11.10**	**9.65**	**7.98**	**6.79**	**19.31**	**20.86**	**17.06**	**13.40**	**11.47**	**9.48**
Gain Over Best_P	-5.10	-6.94	-7.85	-7.95	-7.50	-6.93	**35.86**	**48.84**	**41.90**	**35.81**	**28.76**	**22.79**
Gain Over Best_CR	**14.60**	**13.45**	**8.16**	**3.75**	**7.00**	**24.16**	**6.39**	**2.77**	-0.89	**2.89**	**10.54**	**16.85**
Devset2												
Gain Over ICPR	**10.70**	**9.30**	**5.58**	**3.92**	**3.33**	**4.33**	**28.57**	**36.11**	**33.51**	**26.01**	**19.63**	**17.99**
Gain Over Best_P	**0.38**	-1.30	-0.49	**2.51**	**3.38**	**5.22**	**15.81**	**22.25**	**20.40**	**19.83**	**16.02**	**13.56**
Gain Over Best_CR	**1.13**	**2.52**	**0.19**	**2.23**	**3.43**	**4.12**	**0.59**	-3.23	-2.36	**1.75**	**2.28**	**3.89**
Validset												
Gain Over ICPR	**7.31**	**8.64**	**7.66**	**6.52**	**7.02**	**7.45**	**19.27**	**25.23**	**24.05**	**21.23**	**20.89**	**20.06**
Gain Over Best_P	-3.56	-5.06	-4.68	-4.28	-4.29	-3.36	-6.72	**3.44**	**12.81**	**11.81**	**12.01**	**12.47**
Gain Over Best_CR	**0.75**	**0.85**	**3.11**	**5.76**	**6.07**	**7.23**	-8.22	-4.90	-7.92	-8.44	-7.86	-6.51

Considering the Devset1, there were relative gains for most of the ranking depths considered. Beyond it, gains over *ICPR* occurred at all ranking depths. For the *Best_P*, there was no gain on Precision. However, for this system the gains in terms of diversity were quite expressive, with gains above 20% for all observed depths. Moreover, for the *Best_CR*, there were gains in both relevance and diversity. Overall, the highest gains were achieved in the top positions of the ranking, except for the *Best_CR*, with the fusion methods achieving higher gains at deeper levels.

Regarding the Devset2, there were positive gains against all baselines for most of the considered depths. It indicates that when querying using the metasearch approach, the user obtained more relevant and diverse results. Similar to Devset1, taking *ICPR* and *Best_CR* baselines, the highest gains were achieved at the beginning and at the end of the rankings, respectively. In turn, the highest gains over *Best_P* occurred at the end of the ranking for Precision and at the beginning of the ranking for Cluster-Recall. Notice, specially considering the *ICPR* and *Best_P* baselines, that the fusion methods achieved higher gains in terms of diversity and maintained acceptable relevance.

The gains in the Validset were not expressive, except over the *ICPR* baseline. For the other baselines the gains achieved were unidirectional, that is, for *Best_P* there is no gain in Precision, but only in Cluster-Recall. For the *Best_CR* there was no gain considering Cluster-Recall, but only in Precision. This may be a consequence of the characteristics of the Validset, which unlike the Devsets (with only single-topic queries) also contains multi-topic queries. Hence, it demands further investigations of this challenge and the development of suitable rank fusion methods.

5. Conclusions

This paper described and investigated the use of rank aggregation methods in the metasearch scenario, considering both the impact on relevance and diversification. Our experimental results suggest that fusion methods tend to allow better search results than independent systems. In addition, it was observed that the greatest gains were in terms of diversity, although there were gains in terms of relevance as well. Our findings validated the idea that metasearch systems may allow improvements in the diversity of the results.

It was also found that some fusion methods were flexible enough to improve one objective while maintaining a competitive performance for the other. For some cases, in addition to achieving high gains in terms of diversity, the fusion methods also maintained acceptable results in relevance. However, we noticed that for the multi-topic queries the metasearch did not outperform the best individual system. This highlights the demand of new investigations for the development of novel rank fusion methods able to enhance both relevance and diversity for the case of multi-topic queries.

References

Bahri, S., Zoghlami, N., Abed, M., and Tavares, J. M. R. S. (2019). Big data for healthcare: A survey. *IEEE Access*, 7:7397–7408.

Calumby, R. T., Gonçalves, M. A., and da Silva Torres, R. (2016). On interactive learning-to-rank for IR: overview, recent advances, challenges, and directions. *Neurocomputing*, 208:3–24.

Calumby, R. T., Gonçalves, M. A., and da Silva Torres, R. (2017). Diversity-based interactive learning meets multimodality. *Neurocomputing*, 259:159–175.

Farah, M. and Vanderpooten, D. (2007). An outranking approach for rank aggregation in information retrieval. In *SIGIR'07, Amsterdam, The Netherlands, July 23-27*, pages 591–598.

Ionescu, B., Gînsca, A., Boteanu, B., Lupu, M., Popescu, A., and Müller, H. (2016a). Div150multi: a social image retrieval result diversification dataset with multi-topic queries. In *MMSys'16, Klagenfurt, Austria, May 10-13*, pages 46:1–46:6.

Ionescu, B., Gînsca, A., Zaharieva, M., Boteanu, B., Lupu, M., and Müller, H. (2016b). Retrieving diverse social images at mediaeval 2016: Challenge, dataset and evaluation. In *MediaEval'16 Workshop, Hilversum, The Netherlands, October 20-21*.

Ionescu, B., Popescu, A., Lupu, M., Gînsca, A., Boteanu, B., and Müller, H. (2015). Div150cred: A social image retrieval result diversification with user tagging credibility dataset. In *MMSys'15, Portland, USA, March 18-20*, pages 207–212.

Ionescu, B., Radu, A., Menéndez, M., Müller, H., Popescu, A., and Loni, B. (2014). Div400: a social image retrieval result diversification dataset. In *MMSys'14, Singapore, Mar 19-21*, pages 29–34.

Li, L. T., Pedronette, D. C. G., Almeida, J., Penatti, O. A. B., Calumby, R. T., and Torres, R. d. S. (2014). A rank aggregation framework for video multimodal geocoding. *Multimedia Tools and Applications*, 73(3).

Liang, S., Ren, Z., and de Rijke, M. (2014). Fusion helps diversification. In *SIGIR'14, NY, USA*, pages 303–312. ACM.

McDonald, K. and Smeaton, A. F. (2005). A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *CIVR'05, Singapore, July 20-22, Proceedings*, pages 61–70.

Muñoz, J. A. V., da Silva Torres, R., and Gonçalves, M. A. (2015). A soft computing approach for learning to aggregate rankings. In *CIKM'15, Melbourne, Australia, October 19 - 23*, pages 83–92.

Ramírez-de-la-Rosa, G., Villatoro, E., Ionescu, B., Escalante, H. J., Escalera, S., Larson, M., Müller, H., and Guyon, I. (2018). Overview of the multimedia information processing for personality & social networks analysis contest. In *ICPR'18, Beijing, China, August 20-24, Revised Selected Papers*, pages 127–139.

Santos, R. L. T., MacDonald, C., and Ounis, I. (2015). Search result diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90.

Wu, S., Huang, C., Li, L., and Crestani, F. (2019). Fusion-based methods for result diversification in web search. *Information Fusion*, 45:16–26.

Xu, C. and Wu, S. (2017). The early fusion strategy for search result diversification. In *ACM TUR-C'17, New York, USA*, pages 47:1–47:6.

Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *ACM SIGIR, Toronto, Canada*, pages 10–17.