A Novel Graph-based Diversity-aware Rank Fusion Method Applied to Image Metasearch

José Solenir L. Figuerêdo[1], Ana Lúcia L. Marreiros Maia[1], Rodrigo T. Calumby[1]

[1]Department of Exact Sciences, University of Feira de Santana
Feira de Santana – BA – Brazil

jslfigueredo@ecomp.uefs.br, anamarreiros@gmail.com, rtcalumby@uefs.br

Abstract. *While search result diversification is used to handle ambiguous or underspecified queries, rank aggregation is a widely used approach in metasearch. However, current aggregation methods assume that the input rankings are built only according to the relevance of the items, disregarding the inter-relationship between images in each ranking. Hence, these methods tend to be inadequate for diversity-oriented retrieval. In this work, we introduce a diversity-aware rank fusion method that is validated in the context of diverse image metasearch. The experimental findings indicate that the proposed method significantly improves the overall diversity of metasearch results, in comparison to the state-of-the-art positional and score-based fusion methods.*

1. Introduction

In information retrieval tasks, given a user information need, various rankings can be defined for the same data collection, e.g., considering different search engine configurations, feature representations, and ranking criteria. Hence, rank fusion methods can successfully combine multiple rankings into a unified result. Beyond that, given alternative search systems may present different results for the same user information need, the metasearch technique combines numerous search systems to build a final aggregated ranking. Since those independent results tend to complement each other, metasearch is expected to generate final rankings with improved relevance [Aslam and Montague 2001].

Besides rank fusion, when dealing with complex queries, a technique called diversification is widely used to attenuate some ranking challenges [McDonald et al. 2022]. Specifically, diversification has been demonstrated beneficial to maximizing intent coverage for broad, ambiguous, or under-specified queries, enhancing content-based recommendation systems, handling the redundancy among the retrieved items (e.g., near-duplicate images/documents), and improving user-system information transfer in interactive retrieval sessions [Calumby et al. 2017]. In general, diversification aims at ensuring that at least some items (e.g., documents, images, products) related to different user intentions, interpretations or query aspects are placed at the top positions of the ranking [Yigit-Sert et al. 2020].

The problem addressed by ranking aggregation methods regards the combination of a set of candidate relevance-oriented rankings so that the final combination includes more relevant items than any individual candidate list [Dwork et al. 2001]. Those methods consider relevance as directly related to ranking positions, i.e., the higher the relevance of an item, the higher its ranking position. In contrast, in diversified results, ranking positions do not hold a strict direct relationship to the relevance of the items, i.e., an item in a subsequent lower position is not necessarily less relevant than the previous one, but may just contribute less to the ranking diversity (up to that position) considering the other items in higher positions.

Although fusion strategies have achieved significant gains in terms of relevance improvement, there are still some open challenges regarding diversified rankings. In general, the fusion methods proposed so far consider that the results to be merged were built only on the relevance of the objects, which is not always true. Therefore, by not considering the interrelationship between items, the diversity of the aggregated results can be under-optimized.

For the image metasearch task, some diversity improvements have been reported with the use of relevance and position-based rank aggregation methods [Figuerêdo and Calumby 2019]. Nevertheless, to the best of our knowledge, no previous work has explicitly integrated the concept of diversity and inter-image positional relationship into the rank aggregation procedure itself.

Given the aforementioned challenges, we developed a Graph-based Diversity-aware Rank Fusion method (GDRF) that explicitly considers the concept of diversity in the rank fusion process. For this, we propose a diversity-aware preference graph structure, that stores the positional preference relations between each pair of images in a ranking. The preference graphs generated for each input ranking are combined to produce a new diversity-oriented ranking score. The proposed method suggests a template for ranking diversity representation and is also completely unsupervised. A detailed description of the method is presented in section 3.

2. Related Works

In general, aggregation algorithms fall into two main categories: score-based and order-based. In the former, the fusion procedure takes as input the ranking scores associated with each object in the original rankings. In the latter, order-based algorithms consider only the position of the items in the ranking to perform the fusion process. Some of the most widespread score-based methods are: CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ). In turn, BordaCount, Median Rank Aggregation (MRA) and Reciprocal Rank Fusion (RRF) are popular order-based methods [Vargas Muñoz et al. 2015]. However, although these algorithms have been used in many applications, they do not consider diversification explicitly.

Based on the premise that the fusion process itself can ensure wide coverage of relevant items, some studies have been developed. Two of the main works were developed by [Liang et al. 2014, Xu and Wu 2017]. In the former, diversification is performed in three stages. Initially, the fusion is executed using the CombSUM and CombMNZ methods. Then an inference of latent subtopics is made. Finally, the result generated by the two previous steps is submitted to the diversification process. In the latter, instead of merging already diversified results, the authors chose a direct diversification approach. It also includes three stages: i) Generation of results using search algorithms based only on relevance; ii) Fusion of these results using any algorithm, such as CombMNZ; and iii) Application of an explicit diversification method.

While previous work has focused on the analysis of diversification through fusion methods in the context of web page retrieval, the investigation of such methods in other multimedia scenarios (e.g., image or video retrieval) is still incipient. Furthermore, the applied fusion methods do not consider that the retrieved results may have come from systems that already consider ranking diversification. This work aims at filling this gap by explicitly considering diversified rankings as input to a metasearch approach.

3. Proposed method and experimental setup

The GDRF has three main steps. The first step corresponds to the representation of input rankings as preference graphs, followed by the attribution of edge weights. The preference graph is a position-guided structure with a directed edge between every pair of nodes in the ranking. Each edge characterizes the preference relationship between the connected items. Therefore, rank diversity is captured as multiple preference links between images. Figure 1 illustrates this representation process. In our context, each node corresponds to an image present in the considered ranking. For example, if the ranking contains an image (Img1) in a higher position than another image (Img2), the graph will contain a preference edge directed from node Img2 to node Img1.

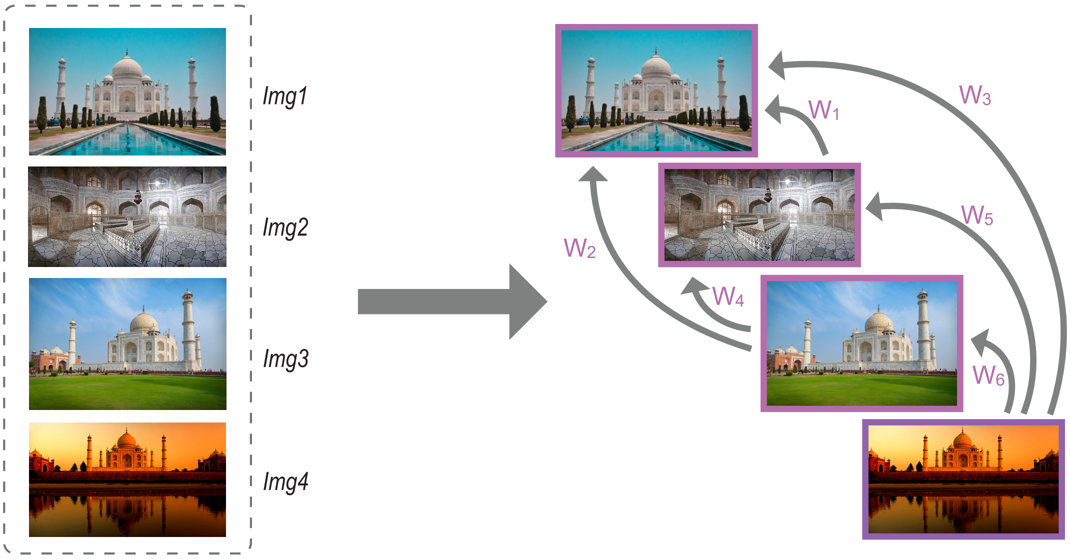

Figure 1. Converting a diversified input ranking to a preference graph.

Each edge between a pair of images has a weight (W), as illustrated in Figure 1. The weight assignment can be performed with different strategies. In our method, the attribution of weights follows Eq. 1. Considering any pair of images *(x,y)*, μ corresponds to the average position occupied by them in the input rankings. The position is designated in descending order. For instance, if the ranking has 50 images, the image at the first position would have 50 as its position score, while the last image scores 1. With this assignment, the images that occupy the first positions are considered more relevant to the query than others. Considering a pair *(x,y)*, α would be the number of nodes (images) that are preferable to *y*, that is, they are above *y* in the ranking. In turn, β represents the number of nodes between the higher node *(y)* and the lower node *(x)*, which indicates how many pairs of images are preferable to the pair being evaluated *(x,y)*.

$$W = 1 - \frac{1}{1 + \frac{\mu}{\alpha+\beta}} \tag{1}$$

Equation 1 aims at capturing the diversity existing in the rankings. Therefore, it favors pairs of diverse images, considering that the base rankings, in addition to being generated considering the relevance, also used diversity as a simultaneous ranking criterion. In the ranking illustrated in Figure 1, assuming Img1 as a reference, Img2 is considered the most relevant to the query while is also more diverse than the others. Otherwise, it wouldn't be the second on the list. In turn, Img3, while possibly more relevant to the query than Img2, contributes less to diversity maximization than Img2. Therefore, the score that aims to capture the degree of diversity between images is greater for the pair Img2-Img1 than for Img3-Img1.

In the second step of the GDRF, an aggregated graph ($\boldsymbol{AG}$) is constructed considering the individual graphs formed in the previous step. The $\boldsymbol{AG}$ relies on the combination of the preference relations obtained from the individual graphs. The resulting graph contains as vertices all the images that appear in at least one of the input rankings. The combination of the weights of an edge *(x,y)* is calculated according to Equation 2.

$$AG_{xy} = \sum G_{xyk} \qquad (2)$$

In equation 2, the summation runs through all the individual graphs that provide preference relations for the *(x,y)* pair. $\boldsymbol{G_{xyk}}$ denotes the preference edge weight from *x* to *y* in the preference graph corresponding to input ranking $\boldsymbol{k}$. Then, step 3 begins, which corresponds to obtaining the final ranking. The induction of the final ranking is carried out from the combined preference relations stored in the $\boldsymbol{AG}$. For this, as different approaches could be followed, we report preliminary experiments with the best performance occurring by sequentially selecting the main nodes, i.e., the ones with the highest accumulated preference weights.

For the experimental evaluation of the GDRF, the collection provided by the Information Fusion for Social Image Retrieval & Diversification Task [Ramírez-de-la-Rosa et al. 2018] was used. This collection includes results from many image search systems proposed and evaluated between 2013 and 2016 in the MediaEval Retrieving Social Images tasks. There are ranked results for numerous queries. In addition, it includes relevant and diverse results with different levels of quality. The dataset is organized into development, validation, and test sets. In this work, we consider only the development set, given the unsupervised nature of the GDRF. Specifically, we pooled *devset1* (39 candidate rankings for 346 queries) and *devset2* (56 candidate rankings for 60 queries). Thus, all analyses were performed on this combined set.

Precision and Cluster-Recall measures were used for effectiveness assessment. Precision represents the quality of the ranking in terms of relevance. The Cluster-Recall measure computes the percentage of conceptual clusters that were represented in a diversified result. For effectiveness analysis, these measures were computed up to the 50th position of the ranking. As baselines, the following order-based fusion methods were used: Borda Count, Median Rank Aggregation (MRA), and Reciprocal Rank Fusion (RRF). For the strict comparison of the effectiveness results, the GDRF was compared to the baselines using Wilcoxon's Signed Rank Test in order to assess the statistical significance of the results.

4. Results and Discussions

Table 1 shows the effectiveness of the proposed method and baselines. The highest values are highlighted in boldface. Considering the relevance of the final rankings, the baseline RRF algorithm achieved the best Precision@N results. However, considering diversification as the main objective in this study, the Cluster-Recall measure plays an important role. The proposed method achieved numerically superior performance over all considered baselines, except for RRF for $N = 5$.

In Table 2 we present the results of Wilcoxon's Signed Rank Test. Green cells represent statistical superiority, white cells mean equivalence, while pink cells represent inferiority against the baseline. Regarding Precision@N, the GDRF was statistically inferior to RRF and MRA ($P@30$ and $P@50$). On the other hand, considering diversity (CR) the GDRF was statistically superior at multiple ranking levels. As the relevance-diversity trade-off is a central and long-lasting challenge in this task, the results reported here suggest that the GDRF is preferable to the baselines, for scenarios in which diversity maximization is a key factor, while further investigations should be performed on how to better optimize trade-off towards simultaneously better relevance results.

Nevertheless, although in the best scenario, the same method should provide the best results for both objectives, for some applications diversity is of great importance. For example, in an e-commerce system, strategically, it may be better to present diverse results with different product models, different shapes, variety of colors, among other characteristics. In such a scenario, a user would be exposed to a wider set of options, even if a few cases of non-relevant items appear in the search result. In addition, by improving diversity, there is an indirect minimization of the redundancy of the results, which is important for a better user experience.

Table 1. Results for the GDRF and baselines. Top values are highlighted in boldface.

	Devset1 + Devset2											
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
Borda Count	0.5915	0.5908	0.5995	0.6052	0.6049	0.5978	0.1735	0.2859	0.4352	0.5552	0.6401	0.7065
MRA	0.8318	0.8271	0.8082	0.7927	0.7691	0.7418	0.2357	0.3890	0.5854	0.7030	0.7881	0.8413
RRF	**0.8567**	**0.8391**	**0.8154**	**0.7949**	**0.7716**	**0.7455**	**0.2618**	0.4115	0.5989	0.7161	0.7929	0.8416
GDRF	0.8308	0.8239	0.8056	0.7879	0.7666	0.7384	0.2580	**0.4167**	**0.6026**	**0.7258**	**0.7998**	**0.8550**

Table 2. Wilcoxon's Signed Rank Test. Green cells represent statistical superiority, white cells equivalence, while pink represent inferiority.

	Devset1 + Devset2											
Pair	P@5	P@10	P@20	P@30	P@40	P@50	CR@5	CR@10	CR@20	CR@30	CR@40	CR@50
GDRF vs Borda Count	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
GDRF vs MRA	⬜	⬜	⬜	🟥	⬜	🟥	🟩	🟩	🟩	🟩	⬜	🟩
GDRF vs RRF	🟥	🟥	🟥	🟥	🟥	🟥	⬜	⬜	⬜	🟩	🟩	🟩

5. Conclusion

This work introduces a novel graph-based diversity-aware rank fusion method validated in the context of metasearch. In terms of the relevance of the metasearch result, the proposed method achieved competitive results, but not enough to outperform the best baseline. On the other hand, the experimental findings indicate that the proposed method allowed superior results in terms of diversity at different ranking levels compared to the baselines. While alternatives should be investigated to more effectively balance the relevance-diversity trade-off, this proposal provides a significant contribution to the field by explicitly considering the diversity concept integrated into a rank aggregation strategy. Future work should investigate, e.g., specific weighting procedures for the input rankings, given that the metasearch is performed over systems with different quality. Additional, novel strategies for assigning weights to preference relationships and other ranking-to-graph and graph-to-ranking transformations could also be proposed.

Acknowledgement

This work was partially supported by UEFS AUXPPG 2023 and CAPES PROAP 2023 grants.

References

Aslam, J. A. and Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 276–284, New York, NY, USA. Association for Computing Machinery.

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

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, page 613–622, New York, NY, USA. ACM.

Figuerêdo, J. and Calumby, R. (2019). Unsupervised rank fusion for diverse image metasearch. In *Anais do XXXIV Simpósio Brasileiro de Banco de Dados*, pages 265–270, Porto Alegre, RS, Brasil. SBC.

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

McDonald, G., Macdonald, C., and Ounis, I. (2022). Search results diversification for effective fair ranking in academic search. *Information Retrieval Journal*, 25(1):1–26.

Ramírez-de-la-Rosa, G. et al. (2018). Overview of the multimedia information processing for personality & social networks analysis contest. In *ICPR'18, Beijing, China, August 20-24*, pages 127–139.

Vargas Muñoz, J. A., da Silva Torres, R., and Gonçalves, M. A. (2015). A soft computing approach for learning to aggregate rankings. page 83–92, New York, NY, USA. ACM.

Xu, C. and Wu, S. (2017). The early fusion strategy for search result diversification. ACM TUR-C '17, New York, NY, USA. ACM.

Yigit-Sert, S., Altingovde, I. S., Macdonald, C., Ounis, I., and Özgür Ulusoy (2020). Supervised approaches for explicit search result diversification. *Information Processing & Management*, 57(6):102356.