Rank Fusion and Multimodal Per-topic Adaptiveness for Diverse Image Retrieval

Rodrigo T. Calumby[1], Iago B. A. C. Araujo[1], Felipe S. Cordeiro[1], Fabiana C. Bertoni[1],
Sérgio Canuto[2], Fabiano Belém[2], Marcos A. Gonçalves[2],
Icaro Dourado[3], Javier A. V. Munoz[3], Lin Tzy Li[3], Ricardo da S. Torres[3]
[1]University of Feira de Santana, Brazil, [rtcalumby,ibacaraujo,fscordeiro,fabianabertoni]@ecomp.uefs.br
[2]Federal University of Minas Gerais, Brazil, [sergiocomputacao,famube,magoncalv]@gmail.com
[3]University of Campinas, Brazil, [icaro.dourado,javier.munoz,lintzyli,rtorres]@ic.unicamp.br

ABSTRACT

This paper presents the MultiBrasil team experience in the Retrieving Diverse Social Images Task at MediaEval 2017. The teams were required to develop a diversification approach for social image retrieval, enhanced with visual summarization. Our proposal for relevance improvement relies on text and credibility-based reranking and rank aggregation. For diversification, we use diversity-oriented reranking and also propose a clustering-based query-adaptive diversity promotion approach. We applied Genetic Programming and Genetic Algorithm-based approaches for combining textual, visual, and user credibility information.

1 INTRODUCTION

Beyond relevance, for many search tasks the coverage of different query aspects/intents in the retrieved set has great impact on fulfilling user needs [5, 11]. Promoting diversity has been shown to positively impact the user search experience specially for ambiguous, underspecified, and visual summarization queries [1–4].

However, tackling the balance between relevance and diversity is still a great challenge. The *Retrieving Diverse Social Images Task 2017* [15] models it into a general ad-hoc image retrieval challenge in which systems are supposed to handle complex and general-purpose multi-concept queries. This paper describes the MultiBrasil team proposals based on reranking and query-adaptive diversity promotion boosted by multimodal rank fusion.

2 PROPOSED APPROACH

The proposed approach consists in improving the original Flickr ranking for relevance-based filtering followed by a diversification step with diversity-oriented reranking or query-adaptive clustering-based summarization. By improving the original ranking and keeping only the top-ranked items, we intend to construct a more relevant subset, which may reduce data noise for the subsequent diversity promotion step. In turn, for diversification, we have evaluated two approaches: (i) relevance-diversity balancing via reranking; and (ii) representative image selection via query-adaptive clustering metric learning.

2.1 Relevance Enhancement and Filtering

For improving the original ranking of the images from each topic, we explored textual and credibility-based ranking. For text ranking, we used the original topic terms as query and for each flickr image obtained for that topic, the title, description, and tag data were concatenated before preprocessing (see Section 3.2). For credibility-based ranking, the user credibility scores were used as relevance of her uploaded images (see Section 3.3).

Additionally, for the aggregation of multiple rankings, we applied the Genetic Programming approach from GPAgg [13], which combines several well-known rank aggregation methods. This method was trained using the development data and integrated order-based (MRA [8], RRF [7], and BordaCount [14]) and score-based (CombMIN, CombMAX, CombSUM, ComMED, CombANZ [12], and RL-Sim [9]) rank fusion methods.

As a relevance-based filtering, from the final aggregated list, the top-ranked images were selected as input for the diversification step. We evaluated multiple cutoff points with best results achieved by using only the 200-top images (run 1) and the 50-top images (runs 2 to 5). Considering the diversification approaches, keeping more than 50 images degraded the final ranking by pushing more non-relevant images to be reranked and consequently also degrading diversity. Since we did not visually reranked the images for run 1, a deeper reranking had to be considered to allow better diversification, which in turn may negatively impact final relevance.

2.2 Diversification

For diversification, we tested two methods. First, a reranking method following the traditional Maximal Marginal Relevance [6] approach considering multiple features. For this method, the feature combination is performed by averaging the individual similarity scores. The relevance-diversity trade-off adjustment was selected based on the best results on the development set.

Alternatively, our query-adaptive clustering method seeks to construct a more suitable clustering structure based on an evolutionary weight adjustment metric learning method guided by intrinsic clustering fitness evaluation. Here, we used a Genetic Algorithm (GA) for per-query feature weight learning. For combination compatibility, min-max normalization was applied for all distance matrices.

As an unsupervised optimization criteria, we evaluated the clustering quality of the discovered functions by clustering the 50-top images using agglomerative hierarchical clustering (average-linkage). For metric learning and final diversification, we used 25

MediaEval'17, 13-15 September 2017, Dublin, Ireland

Table 1: Runs Configurations.

Run	Reranking Function	Cutoff	Diversification Method	Diversification Features
1	-	200	MMR	ACC
2	BM25	50	MMR	BM25+Jaccard
3	BM25	50	MMR	BM25+Jaccard+Phog
4	GPAgg	50	MMR	BM25+Jaccard+Phog
5	BM25	50	GA	BM25+Jaccard+Phog

clusters. The clusters were ranked according to their sizes in descending order and intra-cluster sorting was applied using the images original ranking positions. The final ranking was constructed in a round-robin fashion from the final clusters.

3 FEATURES

3.1 Visual features

We evaluated only the provided visual descriptors for run 1. Additionally, the combination of each visual feature with the best text similarity measures was evaluated for the remaining runs. For the diversification step, the features were combined by averaging the respective similarity scores.

3.2 Text similarity

For text-only and multimodal reranking (runs 2 to 5), the text-based scores were computed as the similarity between the text vectors associated with the images and the original query terms. As text preprocessing, we applied stopwords removal[1] and stemming [10]. We evaluated several similarity scores: BM25, Cosine, Dice, Jaccard, and TF-IDF. These scores were also evaluated for the diversification procedure.

3.3 Credibility

The user credibility scores were individually used for ranking. In this step, we first evaluated the ranking quality of each score individually and finally aggregated the ranking for: bulkProportion, meanImageTagClarity, meanTagRank, meanTagsPerPhoto, meanTitleWordCounts, photoCount, uniqueTags and uploadFrequency. Additionally, we have also created a ranking considering a linear combination of such scores with the weights empirically adjusted, which here we name linearCred.

4 RUN CONFIGURATIONS

We submitted 5 runs. In Table 1, GPAgg (in run 4) is the GP-based rank aggregation of BM25, Jaccard, DICE, and linearCred rankings. For all runs, we have used only the data provided by the task, and the parameters and features used were chosen according to the best results yielded from development set.

5 RESULTS AND DISCUSSION

Tables 2 and 3 present the effectiveness results for the five runs, respectively, for development set and test set, considering all official measures.

Table 2: Development Set Results.

	Development Set				
Run	P@20	CR@20	F1@20	ERR-IA@20	α-NDCG@20
1	0.5832	0.4057	0.4595	0.5452	0.5020
2	0.6977	0.4896	0.5503	0.6344	0.5922
3	0.7073	0.4968	0.5587	0.6413	0.6008
4	0.6914	0.4957	0.5527	0.6308	0.5930
5	0.7009	0.4825	0.5447	0.6269	0.5868

Table 3: Test Set Results.

	Test Set				
Run	P@20	CR@20	F1@20	ERR-IA@20	α-NDCG@20
1	0.5976	0.5758	0.5657	0.5908	0.5618
2	0.7083	0.6524	0.6559	0.6692	0.6391
3	0.7208	0.6482	**0.6634**	0.6778	0.6461
4	0.7202	0.6498	0.6614	0.6806	0.6479
5	0.7173	0.6363	0.6512	0.6355	0.6202

Regarding the test set, the text-only run achieved superior effectiveness than the visual-only run, which is a direct consequence of the textual reranking and filtering of the input list. The visual-only run handled more non-relevant images, which impacted the final relevance and diversity.

In general, all multimodal runs (3, 4, and 5) achieved similar effectiveness, with run 3 being slightly superior. Nevertheless, considering F1@20 for the test set, although run 3 outperformed run 5, in a per-query analysis, we noticed that run 5 outperforms for ~40% of the topics. Moreover, the absolute difference between runs 3 and 5 was 0.0714 in terms of F1@20. Furthermore, even though runs 3 and 4 rely on the same diversification method, the average F1@20 difference was 0.0571, with run 4 outperforming for roughly 43% of the topics. Hence, we highlight the opportunity for further improvement with per-query adaptiveness, for instance, by selecting the most suitable diversification model or even dynamically combining them.

6 CONCLUSIONS

In our experiments, we have combined traditional reranking and clustering-based diversification methods along with ranking fusion and per-query adaptive feature fusion for clustering. Even though traditional methods slightly outperformed our more complex proposals, we found the results to be satisfactory. In this case, the small training corpus is considered a challenging factor for the learning strategies. Moreover, our results have shown the importance of improving the original ranking for allowing better results, both in terms of relevance and diversity. We have also shown that properly selecting the most suitable diversification approach or integrating alternative methods may lead to further improvements.

ACKNOWLEDGMENTS

The authors thank FAPESP, FAPEMIG, CAPES, CNPq, and FAPESB (grant #4098/2016) for partially funding this work.

[1] http://code.google.com/p/stop-words/ (version 2011.11.21 – As of Aug. 2017).

REFERENCES

[1] Fabiano M. Belém, Carolina S. Batista, Rodrygo L. T. Santos, Jussara M. Almeida, and Marcos A. Gonçalves. 2016. Beyond Relevance: Explicitly Promoting Novelty and Diversity in Tag Recommendation. *ACM Transactions on Intelligent Systems and Technology* 7, 3 (2016), 26:1–26:34.

[2] Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. 2015. Multimedia Summarization for Social Events in Microblog Stream. *IEEE Transactions on Multimedia* 17, 2 (2015), 216–228.

[3] R. T. Calumby, R. da S. Torres, and M. A. Gonçalves. 2014. Diversity-driven Learning for Multimodal Image Retrieval with Relevance Feedback. In *Proceedings of the 21st IEEE International Conference on Image Processing*. 2197–2201.

[4] R. T. Calumby, M. A. Gonçalves, and R. da S. Torres. 2016. On Interactive Learning-to-rank for IR: Overview, recent advances, challenges, and directions. *Neurocomputing* 208 (2016), 3–24.

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

[6] J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Anual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–336.

[7] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 758–759.

[8] Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Efficient Similarity Search and Classification via Rank Aggregation. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*. 301–312.

[9] Daniel Carlos Guimarães Pedronette and R. da S. Torres. 2011. Image Re-ranking and Rank Aggregation Based on Similarity of Ranked Lists. In *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns - Volume Part I*. 369–376.

[10] M. F. Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.

[11] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search Result Diversification. *Foundations and Trends in Information Retrieval* 9, 1 (2015), 1–90.

[12] Joseph A. Shaw, Edward A. Fox, Joseph A. Shaw, and Edward A. Fox. 1994. Combination of Multiple Searches. In *The Second Text REtrieval Conference (TREC-2)*. 243–252.

[13] Javier A. Vargas, R. da S. Torres, and Marcos A. Gonçalves. 2015. A Soft Computing Approach for Learning to Aggregate Rankings. In *Proceedings of the 24th ACM International Conference on Conference on Information and Knowledge Management*.

[14] H. P. Young. 1974. An axiomatization of Borda's rule. *Journal of Economic Theory* 9, 1 (1974), 43–52.

[15] Maia Zaharieva, Bogdan Ionescu, Alexandru Lucian Gînscă, Rodrygo L.T. Santos, and Henning Müller. 2017. Retrieving Diverse Social Images at MediaEval 2017: Challenges, Dataset and Evaluation. In *Proceedings of the MediaEval 2017 Workshop*. Dublin, Ireland, Sept. 13-15.