Supervised Learning Methods for Diversification of Image Search Results

Burak Goynuk and Ismail Sengor Altıngövd^(✉)

Middle East Technical University, Ankara, Turkey
{burak.goynuk, altıngövd}@ceng.metu.edu.tr

Abstract. We adopt a supervised learning framework, namely R-LTR [17], to diversify image search results, and extend it in various ways. Our experiments show that the adopted and proposed variants are superior to two well-known baselines, with relative gains up to 11.4%.

1 Introduction

Diversification of search results is a recent trend employed in various contexts (such as searching the web [11], social media [10], product reviews [8], structured databases [3], etc.), where the user query might be ambiguous/underspecified and/or user satisfaction can be increased by providing results related to the alternative aspects of a query. Image search is one such scenario that can benefit from diversification of results, as the diversification requirement is not only due to the different semantic intents of the queries, but may further stem from the visual properties of the images [9]. For instance, for the query “Hagia Sophia”, there may be different photos of this landmark taken in daytime or nighttime, summer or winter, etc., and hence, diversification of results is still required.

In this work, we first apply a recently introduced supervised method to diversify web search results, namely, relational learning to rank (R-LTR) [17], for the diversification of image search results. To this end, we adopt the latter approach to capture both textual and visual diversity of images separately, which is referred to as R-LTR_{IMG}. To learn the feature weights, we employ a neural network (NN) framework with back-propagation, which also enables us to explore more general models, i.e., beyond the linear scoring function of R-LTR [17]. In particular, we train a fully connected two-layer NN using the same set of textual and visual features, called as R-LTR_{IMG-NN}.

Our second contribution is based on the following observation: R-LTR learns a ranking function based on an iterative selection process, where the diversity of a given document is computed wrt. the previously selected documents, i.e., following the paradigm of the well-known Maximal Marginal Relevance (MMR) diversification [2]. We extend R-LTR with an alternative approach, inspired by the Maximum Marginal Contribution (MMC) idea of [12]. While diversifying a result set, the MMC approach takes into account an upperbound for the *future*

diversity contribution that can be provided by the document being scored (details provided in Sect. 2). As far as we know, the earlier approaches for supervised diversification (such as [4, 13, 16, 17]) essentially follow the MMR paradigm and hence, ours is the first attempt to *learn* an alternative ranking function.

Our experiments are conducted using the *Div150Cred* dataset employed in the 2014 Retrieving Diverse Social Images Task (of MediaEval Initiative) in a well-crafted framework. For the baseline strategies, MMR and MSD (described in Sec primer), we employed a dynamic feature weighting strategy for higher performance. For all the diversification methods, we used various pre-processing techniques, and employed a particular strategy based on representative images, to better capture the query-image relevance. We show that the adopted R-LTR_{IMG} and its proposed variants outperform MMR and MSD in diversification effectiveness. Furthermore, according to the results reported in the Diversity task of MediaEval (in 2014), our best-performing variant, R-LTR_{IMG-NN}, is superior to all but one of the methods explored in this evaluation campaign.

2 Background and Preliminaries

The diversification of web search results based on textual evidence is well-explored in the literature [11]. We focus on so-called implicit diversification approaches that solely rely on features extracted from the documents in the ranking to be diversified. In this section, we review representative implicit strategies that are employed for textual diversification of search results, namely, MMR [2], MSD [5], MMC [12] and Relational-LTR (R-LTR) [17]. The former two approaches, MMR and MSD, have been widely employed in the literature (e.g., [11, 14]), and hence, serve as the baselines in our setup. The last one, R-LTR, is a supervised strategy that we adopt and extend in this work.

Maximal Marginal Relevance (MMR) [2]. Given a query q and an initial result set D , MMR constructs a diversified ranking S of size k (typically, $k < |D|$) as follows. At first, the document with the highest relevance score is inserted into S . Then, in each iteration, the document that maximizes Eq. 1 is added to S . The score of a document $d \in D$ is computed as a weighted sum of its relevance to q , denoted as $\text{rel}(q, d)$, and its average diversity from the documents that are already selected into the final result set S . Note that, the diversity part of MMR has different variants that employs minimum or maximum diversity wrt. the documents in S , and the version shown in Eq. 1 is based on [12]. λ is a trade-off parameter to balance the relevance and diversity in the final result set S .

$$MMR(d, q, S) = (1 - \lambda) * \text{rel}(q, d) + \frac{\lambda}{|S|} * \sum_{d_i \in S} \text{div}(d, d_i) \quad (1)$$

Maximum Marginal Contribution (MMC)[12]. This approach is very similar to MMR, but in addition to taking into account the documents already selected in to S , MMC also considers an upperbound on the future diversity, i.e., computed as the contribution of the most diverse l documents (remaining in

D/S) to the current document d . In Eq. 2, the first two components are exactly same as MMR, while the third component captures the highest possible diversity that can be obtained based on d , in case that it is chosen into S .

$$MMC(d, q, S) = (1 - \lambda) * \text{rel}(q, d) + \frac{\lambda}{|S|} * \left(\sum_{d_i \in S} \text{div}(d, d_i) + \sum_{\substack{l=1 \\ d_j \in D-S-d}}^{k-|S|-1} \text{div}(d, d_j) \right) \quad (2)$$

Max-Sum Dispersion (MSD) [5]. At each iteration, MSD selects a pair of documents that are most relevant to the query and most diverse from each other. In particular, for all pairs of documents $(d_i, d_j) \in D$, Eq. 3 is calculated, and the pairs with the highest scores are selected until k results are obtained.

$$MSD(d_i, d_j, q) = (1 - \lambda) * (\text{rel}(q, d_i) + \text{rel}(q, d_j)) + 2 * \lambda * \text{div}(d_i, d_j) \quad (3)$$

Relational-Learning to Rank (R-LTR) [17]. This is a supervised method that learns the weights for an MMR-like diversification approach using Stochastic Gradient Descent (SGD). Instead of relying on a single relevance and diversity score as in the aforementioned approaches, R-LTR computes multiple scores for each component and combines them using weight vectors, which are learnt over a training set.

$$\text{R-LTR}(d_i, R_i, S) = \omega_r * \mathbf{x}_i + \omega_d * h_S(R_i) \quad (4)$$

Following the notation in [17], in Eq. 4, $\mathbf{x}_i$ denotes a relevance feature vector, i.e., a vector of $\text{rel}(q, d)$ scores that are likely to be computed by different methods (e.g., tf-idf, BM25, etc.), while R_i is a matrix capturing the diversity scores of d_i to all other documents in D , again computed by various methods. Note that, R is a 3-way tensor that stores the relation, namely, diversity score, between each pair of documents in D computed using t different $\text{div}(d_i, d_j)$ methods (e.g., body text diversity, title text diversity, anchor text diversity, etc.).

In this case, the function $h_S(R_i)$ is used to compute the aggregated diversity of d_i from the documents that are already in S , for each of these t diversity methods. As in the case of MMR, for a given diversity computation method, the aggregated diversity score between d and S can be computed using average, min or max function, and $h_S(R_i)$ will return the so-called diversity feature vector of t scores computed by using the selected aggregation function.

The ground truth ranking for a query is constructed in a greedy way, i.e., by choosing the document that maximizes a diversification metric, say SubTopic-recall, at each step, which is based on the document’s relevance judgments (see [17] for the details). During training, in each iteration, the document chosen (i.e., maximizing Eq. 4) is compared to the document at the corresponding position in the ground truth so that the likelihood loss can be computed. Then,

model parameters (i.e., ω_r and ω_d) are updated using SGD until the loss converges to a pre-defined value.

We are aware of a previous work [4] that has also exploited R-LTR for image diversification in a similar setup, i.e., MediaEval evaluation campaign. Our work differs from the latter in three ways: First, we implement R-LTR using a neural network framework with back-propagation, which allows us to train more general models, namely, a two-layer neural network, and hence, to go beyond the linear scoring function of R-LTR. Second, we extend R-LTR and propose a new variant that learns the MMC ranking function instead of the MMR. Third, we compare R-LTR variants to two baseline approaches with carefully tuned parameters (to optimize their performance), while the previous work reports only the results of a direct application of R-LTR.

3 Image Diversification Framework: R-LTR_{IMG}

In this paper, as in the Diversity task of MediaEval [6], we assume that a textual query is submitted to an image search engine, where each image is associated with textual metadata, and an initial result list D is retrieved. The goal is to obtain a ranking S that is both relevant to the query and including diverse images. The diversity of two images, again denoted as d_i and d_j , can be computed using textual and/or visual features. Therefore, we first adopt the R-LTR scoring function to separately capture these different types of diversity scores, as follows:

$$\text{R-LTR}_{\text{IMG}}(d_i, RT_i, RV_i, S) = \omega_r * \mathbf{x}_i + \omega_{\text{textDiv}} * h_S(RT_i) + \omega_{\text{visDiv}} * h_S(RV_i) \quad (5)$$

In Eq. 5, the 3-way tensors RT and RV store the pairwise image diversity scores based on the textual and visual diversity, respectively. This adopted version, referred to as R-LTR_{IMG}, also allows using different aggregation (h_S) functions for different types of diversity scores.

As a further extension, instead of considering an MMR-style approach in R-LTR_{IMG}, which only takes into account the diversity wrt. the images that are already in S , we apply the philosophy of aforementioned MMC approach. More specifically, for a given image d_i to be scored, we also compute the upperbound of the diversity that can be brought to S afterwards, i.e., if d_i is selected. To the best of our knowledge, earlier works on supervised approaches for implicit diversification are based on MMR, and ours is the first attempt to *learn* a framework that considers both the images selected into S and those *to be inserted* into S .

In Eq. 6, the last two components address the textual and visual diversity of d_i with respect to l images that are most dissimilar to it in the set of *remaining* images $U = D - S - d_i$. We refer to this version as R-LTR_{IMG-MMC}.

$$\begin{aligned} \text{R-LTR}_{\text{IMG-MMC}}(d_i, RT_i, RV_i, S, U) = & \omega_r * \mathbf{x}_i + \omega_{\text{textDiv}} * h_S(RT_i) + \omega_{\text{visDiv}} * h_S(RV_i) \\ & + \omega_{\text{textDivNext}} * h_U(RT_i, l) + \omega_{\text{visDivNext}} * h_U(RV_i, l) \end{aligned} \quad (6)$$

Finally, we implement these variants using PyTorch’s neural network framework with back-propagation. This choice enables us to train more general models that can go beyond the linear scoring function of R-LTR. In particular, we implement a scoring function (taking the same input as Eq. 5) using a fully connected two-layer neural network and refer to this version as R-LTR_{IMG-NN}. The source codes are available at <https://github.com/burakgoynuk/rltr-img>.

4 Evaluation Setup and Results

Dataset. We employed *Div150Cred* dataset that has been validated in the 2014 Retrieving Diverse Social Images Task (of MediaEval Benchmarking Initiative) [6, 7]. This dataset includes 45,375 images of around 150 landmark locations (e.g., Hagia Sophia) shared in Flickr. Each such location is considered as a query, for which the dataset provides the location’s GPS coordinates, the link to its Wikipedia web page, some descriptive photos from Wikipedia and a ranked set of around 300 images retrieved from Flickr (each associated with various textual and visual features, described later). For the queries and retrieved images, relevance and diversity judgments (by human annotators) are also made available.

We use the default training and test splits as provided in the aforementioned evaluation campaign. Specifically, 30 locations (together with the retrieved images, all metadata and relevance judgments) are used for training, and 123 locations are held as the test set. Note that, while testing, we diversify the top-100 images retrieved for a query, as earlier works imply that going deeper in the ranking increases the likelihood of irrelevant results (e.g., [1]).

Following the common practice (e.g., [1]), we pre-processed all the images in training and test sets to reduce the noise. We first removed the images that may include people, using the Open-CV’s built-in face detection algorithms. Secondly, the GPS coordinate of the queried location is compared to that of an image in the result set, and it is discarded if the distance is greater than 10 km.

Computing the Relevance Scores. To compute the $rel(q, d)$, we employ a strategy based on the *representative image* idea that is widely employed in the literature (e.g., [9]). This strategy aims to go beyond the textual relevance and make use of the visual features, even when the query is expressed textually. To train a model, we consider the top-ranked answer from Flickr as the representative image, i.e., the ground truth. Then, a neural network is trained with three basic features for each (query, image) pair, as follows: BM25 score (between query and image’s textual metadata), the GPS distance of the query and image, and visual similarity between the Wikipedia image of the location and image in the result set. For all diversification methods employed here, we first automatically identify the representative image using the trained model, and obtain the $rel(q, d)$ scores based on visual features, described next.

Computing the Diversity Scores. We use both textual and visual features to compute various types of diversity scores between a pair of images, $div(d_i, d_j)$.

Using the textual features (i.e., image metadata), we compute two types of diversity scores, namely, the tf-idf weighted cosine similarity and Jaccard Coefficient, both of which are in the range $[0, 1]$. In terms of visual features, we consider four types of descriptors per image, namely, Global Histogram of Oriented Gradients, Global Color Structure Descriptor, Global Color Naming Histogram, and Global Color Moments on HSV Color Space (the other descriptors in the dataset are found to be unhelpful in our preliminary experiments on the training set and hence, discarded.). Thus, we compute four different types of diversity score between images, corresponding to each feature type. In particular, the diversity score based on the former two feature types are computed using Euclidean distance, while Cosine distance is found to work better for the latter two feature types. Note that, these scores (actually, their difference from 1) are also used to compute the relevance between a representative image and result image.

Baseline Diversification Methods. As the baselines, we employ MMR and MSD. For both methods, we again compute relevance scores based on the representative images, and diversity scores using the features described before. These scores are weighted using the dynamic feature weighting approach of [9] shown in Eq. 7. In particular, for each textual or visual feature $f \in F$, this approach weights the diversity score by θ_f , which denotes the variance of all diversity scores wrt. f . The trade-off parameter λ is learned over the training set.

$$\text{div}(d_i, d_j) = \frac{1}{|F|} * \sum_{f \in F} \left(\frac{1}{\theta_f^2} * \text{div}_f(d_i, d_j) \right) \quad (7)$$

Parameters for R-LTR Variants. We implement all R-LTR variants using PyTorch’s neural network framework with back-propagation and negative log likelihood loss (as in [13]). We set the number of epochs as 300 and learning rate 0.00001 based on the experiments on training data. For R-LTR_{IMG-MMC}, we set the parameter l in Eq. 6 as 1, i.e., we consider the diversity impact of the farthest image to the current one as an upperbound. For R-LTR_{IMG-NN}, the hidden layer has 3 nodes each with a sigmoid activation function, and a single output node combines the results of the hidden nodes. The ground truth ranking is obtained

Table 1. Diversification performance of baseline and proposed approaches. The symbol (*) denotes stat. significance wrt. MMR using paired t-test (at 0.05 confidence level).

Diversification method	α -nDCG@20	ST-recall@20
Flickr (original ranking)	0.573	0.342
MSD	0.617	0.369
MMR	0.654	0.413
R-LTR _{IMG-MMC}	0.667	0.425
R-LTR _{IMG}	0.691*	0.455*
R-LTR _{IMG-NN}	0.695*	0.460*

by greedily selecting the image that maximizes the ST-recall metric (see [17] for details), which was among the official metrics for MediaEval task.

Results. We report the SubTopic-recall and α -nDCG scores at cut-off value of 20. Table 1 compares the diversification performance of Flickr’s original ranking, MMR and MSD to the proposed R-LTR variants. We see that, as expected, non-diversified Flickr ranking has the lowest performance, and among the two traditional baselines, MMR is better. The proposed R-LTR_{IMG-MMC} approach outperforms MMR, with the relative gains of 2% and 3% in terms of α -nDCG and ST-recall metrics, respectively. However, it is still inferior to R-LTR_{IMG} and its variant, R-LTR_{IMG-NN}. Specifically, R-LTR_{IMG} provides a relative improvement of 10.2% (5.7%) over the best baseline, MMR, in terms of the ST-recall (α -nDCG) metrics, respectively. R-LTR_{IMG-NN} is the overall winner with a relative gain of 11.4% (1.1%) over MMR (R-LTR_{IMG}) in terms of ST-recall, respectively.

Comparison to Diversity 2014 Task Results at MediaEval. Among 14 participants of the Diversity task, 10 of them have submitted a run employing both textual and visual features, as we do here. Their median score for ST-recall is 0.4191 and 9 out of these 10 runs report a score less than 0.45, i.e., inferior to R-LTR_{IMG} and R-LTR_{IMG-NN}. The run outperforming the R-LTR variants achieves a score of 0.473 [15], but they exploit additional features that are not provided in the dataset. Indeed, most submitted runs derive new features and/or employ different pre-processing techniques; so there is still a need for evaluating these methods within a common framework, which is left as a future work.

Concluding Summary. We adopted and extended a supervised learning framework to diversify image search results and showed that the proposed variants are superior to two well-known baselines. As our future work, we aim to train our models by directly optimizing the evaluation metrics, as suggested in [16].

Acknowledgements. This work is partially funded by The Scientific and Technological Research Council of Turkey (TÜBİTAK) grant 117E861 & TÜBA GEBİP award.

References

1. Boteanu, B., Mironică, I., Ionescu, B.: Pseudo-relevance feedback diversification of social image retrieval results. *Multimed. Appl.* **76**(9), 11889–11916 (2016). <https://doi.org/10.1007/s11042-016-3678-6>
2. Carbonell, J.G., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of SIGIR*, pp. 335–336 (1998)
3. Demidova, E., Fankhauser, P., Zhou, X., Nejdl, W.: DivQ: diversification for keyword search over structured databases. In: *Proceedings of SIGIR*, pp. 331–338 (2010)
4. Dudy, S., Bedrick, S.: OHSU @ mediaeval 2015: adapting textual techniques to multimedia search. In: *Working Notes of the MediaEval 2015 Workshop* (2015)

5. Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. In: Proceedings of WWW, pp. 381–390 (2009)
6. Ionescu, B., Gînsca, A.L., Boteanu, B., Popescu, A., Lupu, M., Müller, H.: Retrieving diverse social images at mediaeval 2014: challenge, dataset and evaluation. *MediaEval* **1263** (2014)
7. Ionescu, B., Popescu, A., Lupu, M., Gînsca, A.L., Boteanu, B., Müller, H.: Div150cred: a social image retrieval result diversification with user tagging credibility dataset. In: Proceedings of the 6th ACM Multimedia Systems Conference, pp. 207–212. ACM (2015)
8. Krestel, R., Dokoochaki, N.: Diversifying customer review rankings. *Neural Netw.* **66**, 36–45 (2015)
9. van Leuken, R.H., Pueyo, L.G., Olivares, X., van Zwol, R.: Visual diversification of image search results. In: Proceedings of WWW, pp. 341–350 (2009)
10. Onal, K.D., Altıngöve, I.S., Karagoz, P.: Utilizing word embeddings for result diversification in tweet search. In: Proceedings of AIRS, pp. 366–378 (2015)
11. Santos, R.L.T., Macdonald, C., Ounis, I.: Search result diversification. *Found. Trends Inf. Retr.* **9**(1), 1–90 (2015)
12. Vieira, M.R., et al.: On query result diversification. In: Proceedings of ICDE, pp. 1163–1174 (2011)
13. Xia, L., Xu, J., Lan, Y., Guo, J., Cheng, X.: Modeling document novelty with neural tensor network for search result diversification. In: Proc. of SIGIR. pp. 395–404 (2016)
14. Xioufis, E.S., Papadopoulos, S., Gînsca, A., Popescu, A., Kompatsiaris, Y., Vlahavas, I.P.: Improving diversity in image search via supervised relevance scoring. In: Proceedings of International Conference on Multimedia Retrieval, pp. 323–330 (2015)
15. Xioufis, E.S., Papadopoulos, S., Kompatsiaris, Y., Vlahavas, I.P.: Socialsensor: finding diverse images at mediaeval 2014. In: Working Notes of the MediaEval 2014 Workshop (2014)
16. Xu, J., Xia, L., Lan, Y., Guo, J., Cheng, X.: Directly optimize diversity evaluation measures: a new approach to search result diversification. *ACM TIST* **8**(3), 41:1–41:26 (2017)
17. Zhu, Y., Lan, Y., Guo, J., Cheng, X., Niu, S.: Learning for search result diversification. In: Proceedings of SIGIR, pp. 293–302 (2014)