MIS @ Retrieving Diverse Social Images Task 2015

Maia Zaharieva[1,2] and Lukas Diem[2]

[1]Interactive Media Systems Group, Vienna University of Technology, Austria
[2]Multimedia Information Systems Group, University of Vienna, Austria
maia.zaharieva@[tuwien|univie].ac.at, l.diem@univie.ac.at

ABSTRACT

In this paper, we describe our approach for the MediaEval 2015 Retrieving Diverse Social Images Task. The proposed approach exploits available user-generated textual descriptions and the visual content of the images in a combination with common, unsupervised clustering techniques in order to increase the diversification of retrieval results. Preliminary experiments indicate that the approach generalizes well for different datasets and achieves comparable results for single- and multi-topic queries.

1. INTRODUCTION

Manual assessment of the relevance of publicly available images to a particular query is not feasible due to the immense amount of data captured and shared daily on social media platforms. As a result, the automated optimization of image retrieval results gains constantly in importance. Next to relevance, the aspect of diversification of retrieval results plays a crucial role in order to reduce the redundancy in the retrieved images and, thus, to increase the efficiency in overviewing the underlying data. The MediaEval 2015 *Retrieving Diverse Social Images Task* [4] addresses these challenges in form of a tourist-oriented retrieval task, where the topics of interest represent sightseeing spots around the world. The aim of the task is to refine the set of images retrieved from Flickr while taking into account both the relevance and the diversity of the selected images.

Previous work in this context shows a broad range of possible approaches. The original Flickr ranking is commonly improved by a direct comparison with the corresponding Wikipedia images [5][8]. Other methods employ training by support vector machines (SVM) [6] or regression models [3]. The diversification of retrieval results is usually approached by means of conventional clustering algorithms, such as k-means [3][6], hierarchical clustering [1][2], and random forest [8] or by an ensemble of clustering approaches [5].

In this paper, we address relevance re-ranking by means of a similarity score to a reference set of images. This reference set is given by Wikipedia images (if available) or by the top ranked images provided by Flickr. To increase diversification, we employ a hierarchical clustering algorithm and compare the performance of recently-introduced powerful visual features with text-based approaches, which are well-established in the context of web mining and retrieval.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2. APPROACH

We employ a multi-stage workflow for the retrieval of diverse social images, which passes the following steps: 1) data preprocessing, 2) relevance reranking, and 3) image clustering and final image selection.

In the first step, *data preprocessing*, we filter potentially irrelevant images, i.e., images with humans as the main subjects and images that are captured far away from the topic of interest. We employ the OpenCV[1] face detector and remove images with faces of area exceeding 5% of the total image area. Additionally, if GPS data is available, we measure the distance between the topic of interest and the corresponding images and remove those with a Harvesine distance [7] greater than 100km. The reason for this strict threshold is the underlying tourist application scenario where the precision of location's specification ranges strongly from a particular spot (e.g., the *Tower Bridge* in London) to large-scale locations such as national parks or entire cities.

The aim of the second stage, *relevance reranking*, is to improve the original Flickr rating. Since the provided Wikipedia images are per definition representative [4], we measure the visual similarity between the images of a set and the associated Wikipedia images by means of the Euclidean distance between the corresponding adapted convolutional neural network (CNN) based descriptors. In case that there are no Wikipedia images provided for a given query, we consider the top 10 images from the original Flickr ranking as reference images. Following, all images are reranked according to the achieved similarity score.

In the third step, *image clustering*, we aim at finding groups of similar images which can be used to diversify the final image results. For the visual-based runs, preliminary experiments with the provided visual descriptors [4] and different clustering algorithms (k-means, k-medoids, XMeans, and agglomerative hierarchival clusteirng (AHC)) showed that the best performing method for the development data considers CNN as a visual feature and the AHC clustering method. The final selection of images from the clusters follows a Round-Robin approach. We start by selecting the image with the best relevance score from each cluster. These images, sorted in ascending order, constitute the m highest ranked results, where m is the number of detected clusters. The selected images are removed from their corresponding clusters and the selection process is repeated until the required number of retrieved results is achieved. We employ the Ward's aggregation method and limit the number of final clusters to 50 based on preliminary experiments.

[1] http://opencv.org

Table 1: Experimental results on the development dataset in terms of precision (P@20), cluster recall (CR@20), and F1-score (F1@20). Employed runs consider visual (V) and/or textual (T) information.

	Data preprocessing	Relevance reranking	Image clustering	P@20	CR@20	F1@20
–	Flickr baseline			0.812	0.343	0.471
T	GPS filter	–	–	0.820	0.350	0.478
V	Face filter	–	–	0.816	0.349	0.478
V,T	Face+GPS filter	–	–	0.825	0.355	0.485
T	–	–	TF-IDF	0.784	0.455	0.569
T	GPS filter	–	TF-IDF	0.799	0.462	**0.577**
T	–	–	LDA	0.798	0.420	0.542
T	GPS filter	–	LDA	0.815	0.429	0.553
V	–	CNN	–	0.936	0.282	0.420
V	–	–	CNN	0.783	0.437	0.553
V	–	CNN	CNN	0.831	0.454	0.578
V	Face filter	CNN	CNN	0.835	0.461	**0.584**
V,T	Face+GPS filter	–	TF-IDF	0.819	0.464	**0.584**
V,T	Face+GPS filter	CNN	TF-IDF	0.925	0.318	0.460
V,T	Face+GPS filter	–	LDA	0.830	0.437	0.564
V,T	Face+GPS filter	CNN	LDA	0.933	0.318	0.459
V,T	Face+GPS filter	CNN	CNN	0.849	0.468	**0.593**

For the text-based runs we consider two approaches. First, we perform topic modeling on the textual descriptions of each image (title and tags) using Latent Dirichlet Allocation (LDA) and the MALLET Toolbox[2] and extract T topics for the employed dataset. For each image, we estimate the likelihoods l_1 and l_2 of the first- and second-best matching topics. If the difference of the likelihoods is larger than a threshold τ ($l_2/l_1 < \tau$) the most likely topic (l_1) is assigned to the photo otherwise no topic is assigned. We set $T = 50$ and $\tau = 0.8$ for all experiments.

The second text-based approach considers the well-established term frequency-inverse document frequency (TF-IDF). We compute the TF-IDF vector for each image using the complete textual description (title, tags, and descriptions). The textual descriptions are first preprocessed to increase their expressiveness, i.e., we remove potential occurrences of the corresponding user name, web links, and stopwords and we additionally stem all remaining terms. Furthermore, we account for images with missing textual descriptions. In such a case, we search for timely closest image with a description which is either captured within a predefined radius (10 meter in our experiments) or by the same user within a predefined short time span (e.g., 5 minutes). In the following, we cluster the resulting TF-IDF vectors using again the AHC method, whereas the similarity between the TF-IDF vectors is measured using the cosine similarity. The selection of the final image set follows the Round-Robin approach as described for the visual-based approach.

3. EXPERIMENTAL RESULTS

Table 1 presents a selection of our preliminary experiments on the development dataset. The results show that the prerpocessing step (face and GPS filter) only marginally improves the performance for the top 20 retrieved images in comparison to the Flickr baseline results. Nevertheless, 95% of the rejected images are irrelevant with respect to the underlying search query. Experiments with the text-based runs show only minor differences in the performance of the TF-IDF and the LDA-based methods. While the achieved precision (P@20) is comparable to those of the Flickr baseline, the cluster recall (CR@20) improves notably, e.g. from 0.34 to 0.46 using the TF-IDF approach. For the visual-based runs, the consideration of the relevance reranking step using the CNN features demonstrates a significant increase in the relevance (P@20-score of 0.94). However, the drop in the clustering recall indicates an increase of redundancy in the retrieved images as a side-effect. Overall, the best-performing text-based and visual-based runs are comparable in terms of F1@20 with the computational costs for the text-based runs being significantly lower. The multimodal runs additionally slightly improve both the clustering recall and the F1-scores by approximately 1%. Surprisingly, the consideration of the reranking step in a combination with the text-based image clustering and selection cannot compensate for the drop in the clustering recall.

Table 2: Official runs configurations.

Run	Data preprocessing	Relevance reranking	Image clustering
1 (V)	Face filter	CNN	CNN
2 (T)	GPS filter	–	TF-IDF
3 (V,T)	Face+GPS filter	CNN	CNN
5 (V,T)	Face+GPS filter	–	TF-IDF

Table 3: *MediaEval 2015 Benchmark* results. Bold values indicate best values in terms of F1-score for the different types of test data.

	single-topic			multi-topic			overall		
	P@20	CR@20	F1@20	P@20	CR@20	F1@20	P@20	CR@20	F1@20
1	0.779	0.450	0.552	0.763	0.434	0.534	0.771	0.442	0.543
2	0.690	0.407	0.501	0.697	0.434	0.519	0.694	0.421	0.511
3	0.794	0.455	**0.560**	0.764	0.435	**0.535**	0.778	0.445	**0.547**
5	0.688	0.409	0.501	0.699	0.429	0.517	0.694	0.419	0.509

Following our preliminary experiments we submitted four runs corresponding to the best configuration for the respective modality (see Table 2). Table 3 summarizes the results of the official runs on the test dataset. In opposite to the development data, which contains the retrieval results of single-topic queries only, the test data differentiates between single- (e.g., *Niagara Falls*) and multi-topics queries (e.g., *Academy awards in Hollywood*). Overall, there is no significant difference in the performance for the two subsets. While the (predominantly) visual-driven runs (runs 1 and 3) show a slight decrease in the clustering recall for the multi-topic queries, the text-driven runs (runs 2 and 4) indicate the opposing trend. Furthermore, in contrast to the results on the development data, the test runs show notable difference between the performance of the text- and the visual-based runs. This reveals the better generalization ability of the visual-based runs to different datasets. Overall, the best performance in terms of F1-score of 0.55 is achieved by the visual-based run which additionally considers the face and GPS filters to reject irrelevant images (run 3).

4. CONCLUSION

In this paper we investigated both text- and visual-driven approaches for the diversification of Flickr image retrieval results. The achieved performances indicate that the visual-based approach copes well with different data and varying query types. Overall, the relevance ranking shows promising results in terms of precision. However, the diversification increases only slowly by means of clustering recall. Our future work will exploit the potential of combining features of different modalities in the clustering process, e.g. by means of a late fusion approach.

Acknowledgment

This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010.

[2] http://mallet.cs.umass.edu.

5. REFERENCES

[1] A. Castellanos, A. Garcia-Serrano, and J. Cigarran. UNED @ retrieving diverse social images task. In *MediaEval Benchmark Workshop*, 2014.

[2] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. D. Natale. Retrieval of diverse images by pre-filtering and hierarchical clustering. In *MediaEval Benchmark Workshop*, 2014.

[3] A. L. Ginsca, A. Popescu, and N. Rekabsaz. CEA LIST's participation at the MediaEval 2014 retrieving diverse social images task. In *MediaEval Benchmark Workshop*, 2014.

[4] B. Ionescu, A. L. Gînscâ, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving diverse social images at MediaEval 2015: Challenge, dataset and evaluation. In *MediaEval Benchmark Workshop*, 2015.

[5] J. R. M. Palotti, N. Rekabsaz, M. Lupu, and A. Hanbury. TUW @ retrieving diverse social images task 2014. In *MediaEval Benchmark Workshop*, 2014.

[6] M. I. Sarac and P. Duygulu. Bilkent-RETINA at retrieving diverse social images task of MediaEval 2014. In *MediaEval Benchmark Workshop*, 2014.

[7] R. W. Sinnott. Virtues of the haversine. *Sky and Telescope*, 68(2):159, 1984.

[8] C. Spampinato and S. Palazzo. PeRCeiVe@UNICT at MediaEval 2014 diverse images: Random forests for diversity-based clustering. In *MediaEval Benchmark Workshop*, 2014.