DWS at MediaEval 2014 Retrieving Diverse Social Images: Approaches for Visual and Textual Runs

Lydia Weiland
Data and Web Science Research Group
University of Mannheim, Germany
lydia@informatik.uni-mannheim.de

Simone Paolo Ponzetto
Data and Web Science Research Group
University of Mannheim, Germany
simone@informatik.uni-mannheim.de

ABSTRACT

We present an overview of our two runs submitted to the MediaEval 2014 Retrieving Diverse Social Images task, each relying only on visual and textual features. Whilst the approach for textual features is based on a standard tf-idf bag-of-words approach, we focused for visual features on a more complex contribution for the task which consists of clustering the images and diversifying the result list based on visual features. At its heart, our method relies on using images collected for each location from Wikipedia. These images are then used as centroids of clusters, where the images collected from Flickr based on their similarity to the Wikipedia images are later grouped in. Both runs, namely using either visual or textual information only, achieve precision-oriented (i.e., more than twice higher precision than recall) results.

1. INTRODUCTION

This year's dataset for the MediaEval 2014 Retrieving Diverse Social Images task contains images and their textual descriptions of 30 locations for the development dataset and of 123 locations for the test dataset [3]. There are several visual and textual features for each image and their corresponding textual description. These features can be used in order to create a ranked result list of images. The characteristics of the images in the result list is defined by two requirements: The images have to be relevant with respect to the query and from these groups of relevant images, intragroup-wise the most diverse ones [1].

Filtering out images can be grouped into two main sub-tasks. The first task is to identify images which are indeed wrong, because the location is not shown. These images are hard to find with visual features only, because they have often similar visual characteristics, e.g., for the location `Obelisco` an image of the Berlin Victory Column would bring noise into the data (cf. Figure 1). The second task focuses instead on identifying images where the location is not the central aspect of the image (e.g., photographs showing a person in focus and just a tiny piece of the location in the background). These images are instead hard to find with textual features only, because they use the name of the location within at least one of the textual descriptors (title, description, or tag) and fit to the textual query for collecting the initial data (e.g., an image for location `Leaning Tower` showing a woman next to a green lawn taking a photograph with the caption "Mary taking a picture up the tower of Dave taking a picture down from the tower", cf. Figure 1). As a result of this, textual and visual features can be expected to be both beneficial to estimate the relevance of image results.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

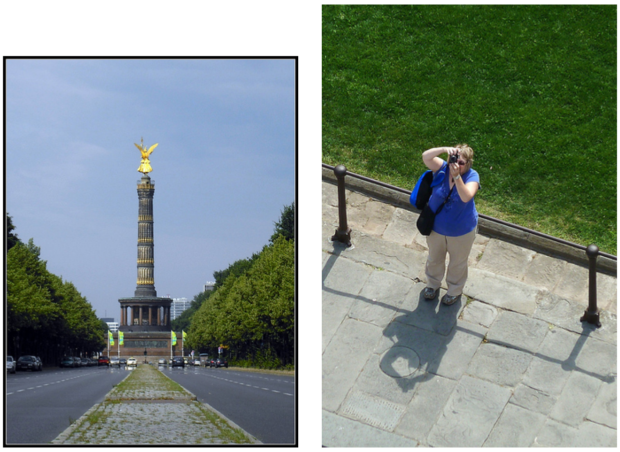

Figure 1: Left: Berlin's Victory Column (Source: `http://flic.kr/p/7R1SDi`). The image is an example of a noisy instance, since it refers to the location `Obelisco`. This kind of noisy images is hard to identify based on visual features only. The image on the right (Source: `http://flic.kr/p/c86JBf`), instead, is hard to locate with textual features only. Its description, in fact, is relevant to the query "`Leaning Tower`", but the image is not.

Diversification is defined for this task in terms of different visual compositions, e.g., images showing the location in daylight or by night, or from near or a bird's perspective. Even if the key goal is similar to last year's task [2], the number of images within the devset and testset changed and new features were added, e.g., user annotation credibility [3]. The official ranking metrics of this year reflect the two requirements for relevance and diversification, namely a balanced F-measure computed over the first 20 images.

2. METHODOLOGY

The methods we developed are based on the provided features and data. Besides the features of the images, additional external information for each location was provided. As an exception, the participants were allowed to use these additional information for the first runs, e.g., images from Wikipedia with their affiliated image features (Color Names, Histogram of Oriented Gradients (HOG), etc.). No other external information than the provided ones was allowed in the final runs. Wikipedia images were given without textual descriptions: accordingly we developed two different process chains for the two runs.

Table 1: Results of the two submitted runs on devset with the three metrics for relevance (Precision - P@X), diversity (Cluster Recall - CR@X), and the harmonic mean of both (F1-Measure - F1@X).

Name of run	P@20	CR@20	**F1@20**
Run 1: Visual	0.735	0.3499	**0.4652**
Run 2: Textual	0.7167	0.314	**0.4279**

Table 2: Results of the two submitted runs on testset with the three metrics for relevance (Precision - P@X), diversity (Cluster Recall - CR@X), and the harmonic mean of both (F1-Measure - F1@X).

Name of run	P@20	CR@20	**F1@20**
Run 1: Visual	0.7524	0.3405	**0.4600**
Run 2: Textual	0.6789	0.3022	**0.4104**

Run 1: Visual Information Only. We start with the assumption that images from Wikipedia are showing the location from different perspectives, i.e., they provide good examples of diverse images. We then developed the following pipeline to filter, cluster, and diversify the images. First, each image from Wikipedia is taken as centroid of a cluster. Images crawled from Flickr are then analyzed for their distance to one of the clusters. The candidate image is then grouped into one of the clusters with the lowest distance (we do not use soft clusters to avoid duplicates in the ranked result list). Distances are calculated using the Euclidean distance of the HOG values for each patch (an image has 9 patches). The method uses HOG features as they have shown to outperform other features [4]. We then use inverse distances as measures for the relevance ranking. Sorting the images in descending order of their best similarity value results in a list of ranked images (we store the top 50 as final output). Filtering images before clustering using face detection algorithms could potentially lead to slightly better results: however, it has also shown a negative impact with respect to the CR values of some of the locations of the devset. Thus, we decided not to use a filtering method for our final run.

Run 2: Textual Information Only. For each image the tf-idf weights are given with respect to different reference data (image, location, and user). We decided in our run to use only the location-related tf-idf weights. Storing these values in a vector for each image allows for calculating the cosine similarity between two images. For each location the cosine similarities are calculated for each pair of images. We remove pairs with maximum similarity, i.e., 1, based on two assumptions: i) either we have the very same image, or ii) two images have exactly the same textual description – both cases, in fact, cannot positively impact result diversification. We finally return the top 50 images with highest cosine similarity as output ranked result list.

3. RESULTS

Table 1 and 2 show the overall results of Precision, Cluster Recall, and F1-Measure on the devset and testset for the two runs. In general, we achieve for Run1 better results for all metrics of precision, recall and F1-measure. Results on the devset vs. testset are similar and show only marginal difference for both runs. The difference between the Precision and Recall values demonstrate the suitability of the methods for detecting similarities. The low CR and F1 values of Run 2 seem to indicate that, indeed, the filtered equal textual descriptions do not automatically indicate that the same perspective and image composition is given, and thus, that the image needs to be deleted from the result list. To reduce the difference between those two types of measures an additional processing step, which addresses the diversification part of the task, seems to be required.

4. CONCLUSIONS

Both our runs show weaknesses on the diversification side (CR@20), whilst the precision (P@20) have shown to be more than twice better. The results indicate that the methodologies are basically able to spot relevant data in a coarse way, whereas fine-grained diversification and ranking still need to be improved. In the future, we plan to collect Wikipedia articles for each location in order to build a topic model, as opposed to merely using tf-idf weights of the Flickr data only. From the visual processing perspective, we plan to improve on result diversification by either using other clustering approaches (e.g., allowing Flickr images to build an own cluster) or by crawling more than five Wikipedia images representing the location and at the same time being diverse to the Wikipedia images, which are already collected. Related work in the field of ranking and retrieving multimedia data have shown that the combination of visual and textual features in a multimodal model outperforms single modality models. Thus, we also plan to conduct experiments where both types of features are used jointly within one single model [5].

5. REFERENCES

[1] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 39:1–39:8, 2009.

[2] B. Ionescu, M. Menéndez, H. Müller, and A. Popescu. Retrieving diverse social images at MediaEval 2013: Objectives, dataset and evaluation. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013.*, 2013.

[3] B. Ionescu, A. Popescu, M. Lupu, A. L. Ginsca, and H. Müller. Retrieving diverse social images at MediaEval 2014: Challenge, dataset and evaluation. In *MediaEval 2014 Workshop, October 16-17, Barcelona, Spain, 2014.*, 2014.

[4] A. Popescu. CEA LIST's participation at MediaEval 2013 retrieving diverse social images task, 2013.

[5] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proc. of MM '10*, pages 251–260, 2010.