DCLab at MediaEval 2015 Retrieving Diverse Social Images Task

Zsombor Paróczi
Budapest University of Technology and Economics
paroczi@tmit.bme.hu

Máté Kis-Király
Budapest University of Technology and Economics
kis.kiraly.mate@gmail.com

Bálint Fodor
Budapest University of Technology and Economics
balint.fodor@gmail.com

ABSTRACT

In this paper we present our contribution to the MediaEval 2015 Retrieving Diverse Social Images Task which requested participants to provide methods for refining Flickr image retrieval results thus to increase their relevance and diversification. Our approach is based on re-ranking the original result, using a precomputed distance matrix and a spectral clustering scheme. We use color related visual features, text and credibility descriptors to define similarity between images.

1. INTRODUCTION

When a potential tourist makes an image search for a place, she expects to get a diverse and relevant visual result as a summary of the different views of the location.

In the official challenge (Retrieving Diverse Social Images at MediaEval 2015) [2] a ranked list of location photos retrieved from Flickr is given, and the task is to refine the result by providing a set of images that are both relevant and provide a diversified summary. An extended explanation for the task objectives, provided dataset and evaluation descriptors can be found in the task description paper [2]. The diversity means that images can illustrate different views of the location at different times of the day/year and under different weather conditions, creative views, etc. The utility score of the refinement process can be measured using the precision and diversity metric [8].

Our team participated in previous challenges [7, 6], each year we experimented with a different approach. In 2013 we used diversification of initial results using clustering, but our solution was focused on diversification only [7]. In 2014 we tried to focus on relevance and diversity with the same importance as a new idea [6].

In the previous approaches to the task we treated our feature vectors (calculated values from metrics) as an N dimensional continuous space with Euclidean coordinates. In this year apporach we will define a set of hand crafted distance matrices with non-Euclidean coordinates, which can be used during the clustering.

2. RUNS

In this section we introduce the approaches used to generate the runs for each task.

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

2.1 Run1: Visual based re-ranking

In the first run participants could use only visual based descriptors or own descriptors calculated using only the images.

For the first run we use the following approach: step 1 - calculating $FACE$ descriptor for each image, step 2 - filter the images using $FACE$ and $CN[0]$ descriptors, step 3 - creating a distance matrix from color similarity, step 4 - doing spectral clustering using the distance matrix, step 5 - using the cluster information create the new result list.

Our main approach was using color based distances [1, 5] and filtering photos with faces [7, 6]. We used two of the descriptors provided by the organizers [2]:

CM Global Color Moments on HSV Color Space: represent the first three central moments of an image color distribution: mean, standard deviation and skewness

CN Global Color Naming Histogram: maps colors to 11 universal color names: "black", "blue", "brown", "grey", "green", "orange", "pink", "purple", "red", "white", and "yellow"

First we calculated a new descriptor for each image: the $FACE$ descriptor is the ratio of the calculated area occupied by the possible face regions on an image and whole image area [7]. Then we used the CN descriptor to filter out black color based images, since mostly dark images tend to have less colors and those are mainly shifted into the gray range rather than having bright colors.

In the reordering step we started from the original result. We did our initial filtering by putting images to the end of the result list where $FACE > 0$ or $CN[0] > 0.8$, the first value in CN corresponds to the color black.

After the preprocessing step we built the distance matrix F, between each A and B images the distance was calculated using the following equation:

$$F_{A,B} = \sum_{i=0}^{10} |CN_A[i] - CN_B[i]| + \sum_{i=0}^{10} |s_i * (CM_A[i] - CM_B[i])|$$

$$s_i = \begin{cases} 5, \text{ where } 0 \leqslant i < 3 \\ 1.5, \text{ where } 3 \leqslant i < 5 \\ 0.5, \text{ where } 5 \leqslant i < 9 \end{cases}$$

After the distance matrix was created we used unsupervised spectral clustering [3, 4] to create clusters from the first 150 images, the target cluster count was 10.

The final result was generated by picking the lowest ranking item from each cluster, appending those to the result

list, then repeating this until all the items are used. The same clustering and sorting method was used during run2 and run3.

2.2 Run2: Text based re-ranking

The second run was the text based re-ranking which is accomplished using the title, tags and description fields of each image.

For the second run we use the following approach: step 1 - filtering stop words and characters, step 2 - creating a distance matrix from text similarity, step 4 - doing spectral clustering using the distance matrix, step 5 - using the cluster information create the new result list.

As a preprocessing step we executed a stop word filtering. We also removed some special characters (namely: .,-:;0123456789()_@) and HTML specific character sets (&, " and everything between < and >), then we used the remaining text as the input for a simple TF-IDF calculation [9].

We calculated the distance between images (e.g. description fields) A and B in the following manner. We initialize distance $G_{A,B}$ to zero and compared A and B at the term level. All occurring t terms in document A compared with all terms in the document B and so on. If term t is contained by both documents, then $G_{A,B}$ will not be increased. If t contained by only one document, we take into consideration the document frequency (DF_t): if $DF_t < 5$, then it is a rare term and $G_{A,B}$ should be increased by 2; if $DF_t > DN/4$, then it is a common term and $G_{A,B}$ should be increased by 0.1 (where DN is the total number of documents). If the term is not common nor rare, then we added the DF_t/DN to the distance.

Using the three text descriptors we created a weighted sum for the field distances, where the empirically determined weights are as follows: title=1, tags=2, description=0.5. From these $G_{A,B}$ values we created the G distance matrix.

2.3 Run3: Multimodal re-ranking

In the third run both visual and textual descriptors could be used to create the results.

For the third run we use the following approach: step 1 - creating the distance matrix F (see Section 2.1), step 2 - creating the distance matrix G (see Section 2.2), step 2 - creating a new distance matrix from combining F and G, step 4 - doing spectral clustering using the distance matrix, step 5 - using the cluster information create the new result list.

We used our visual distance matrix F and text distance matrix G and created a new aggregate matrix H. This matrix is simply the sum of the corresponding values from both F and G matrix. We tried different kind of weighting methods, but the pure matrices supplied the best results on the development set.

2.4 Run4: Credibility based re-ranking

In the fourth run participants were provided with credibility descriptors [2].

Using the original result we filtered the images by users who had $faceProportion$ more than 1.3 to create the same effect as we did with the $FACE$ descriptor.

run name	P@20	CR@20	F1@20
Run1 single	.7022	.3702	.4751
Run1 multi	.7164	.3857	.4813
Run2 single	.6435	.3494	.4379
Run2 multi	.7021	.3813	.4748
Run3 single	.6732	.3563	.4554
Run3 multi	.6993	.3683	.4651
Run4 single	.7014	.3589	.4651
Run4 multi	.7150	.3498	.4479

Table 1: Official results on the test data.

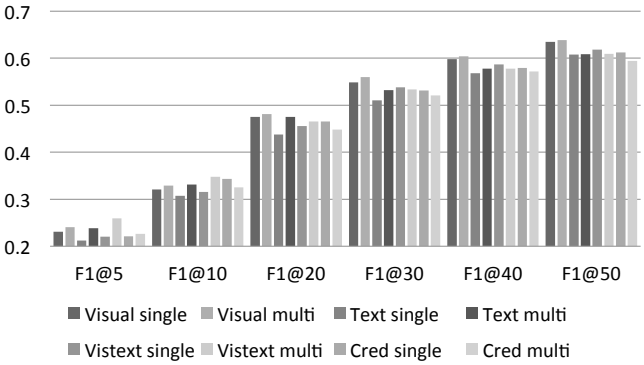

Figure 1: Official runs F1-score metric for various cutoff points (results of test data).

With the purpose of increasing the diversity we used the *locationSimilarity* descriptor, if this value exceeds the threshold of 3.0 we excluded the image. Despite our simple approach we had great results on the development set.

3. RESULTS

The 2015 dataset contained 153 location queries (45,375 Flickr photos) as the development set, we used this to develop our approach, all methods and thresholds were calculated using the whole development set.

The test set containing 139 queries: 69 one-concept location queries (20,700 Flickr photos) and 70 multi-concept queries related to events and states associated with locations (20,694 Flickr photos). Single-topic queries are basic formulations such as the name of a location, multi-concept queries are more complex, they are related to events and states associated with locations (like 'sunset in the city').

Our results can be seen in Table 3. and the F1 metrics can be seen in Figure 1, we listed the single and multi-concept based results separately.

4. CONCLUSION AND FUTURE WORK

As one can see the visual information based results are the best among all the runs. In the development set we experienced that the textual information for many images are missing or do not describe the content very well. It is not uncommon that an author gives the same textual information to all of the images in a topic.

The credibility based descriptors are proved to be much more useful than we initially thought, in the future we should focus on those to improve textual and visual descriptor based results.

5. REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60, May 2008.

[2] G. A. B. B. P. A. L. M. M. H. Ionescu, B. Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, CEUR WS.org*, 2015.

[3] X. Ma, W. Wan, and L. Jiao. Spectral clustering ensemble for image segmentation. In *Proceedings of the First ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, GEC '09, pages 415–420, New York, NY, USA, 2009. ACM.

[4] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856. MIT Press, 2001.

[5] M. L. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: Overview of the imageclefphoto task 2009. In *Proceedings of the 10th International Conference on Cross-language Evaluation Forum: Multimedia Experiments*, CLEF'09, pages 45–59, Berlin, Heidelberg, 2010. Springer-Verlag.

[6] Z. Paróczi, B. Fodor, and G. Szucs. Dclab at mediaeval2014 search and hyperlinking task. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, October 16-17, CEUR-WS. org, ISSN 1613-0073*, 2014.

[7] G. Szűcs, Z. Paróczi, and D. Vincz. Bmemtm at mediaeval 2013 retrieving diverse social images task: Analysis of text and visual information. In *Working Notes Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, October 18-19, CEUR-WS. org, ISSN 1613-0073*, 2013.

[8] B. Taneva, M. Kacimi, and G. Weikum. Gathering and ranking photos of named entities with high precision, high recall, and diversity. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 431–440, New York, NY, USA, 2010. ACM.

[9] J.-B. Yeh and C.-H. Wu. Video news retrieval incorporating relevant terms based on distribution of document frequency. In *Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '08, pages 583–592, Berlin, Heidelberg, 2008. Springer-Verlag.