Retrieving Diverse Social Images at MediaEval 2016: Challenge, Dataset and Evaluation

Bogdan Ionescu
LAPI, University Politehnica of Bucharest, Romania
bionescu@alpha.imag.pub.ro

Alexandru Lucian Gînscă
CEA, LIST, France
alexandru.ginsca@cea.fr

Maia Zaharieva*
University of Vienna & Vienna University of Technology, Austria
maia.zaharieva@univie.ac.at

Bogdan Boteanu
LAPI, University Politehnica of Bucharest, Romania
bboteanu@alpha.imag.pub.ro

Mihai Lupu
Vienna University of Technology, Austria
lupu@ifs.tuwien.ac.at

Henning Müller
HES-SO, University of Applied Sciences Western Switzerland
henning.mueller@hevs.ch

ABSTRACT

This paper provides an overview of the Retrieving Diverse Social Images task that is organized as part of the MediaEval 2016 Benchmarking Initiative for Multimedia Evaluation. The task addresses the problem of result diversification in the context of social photo retrieval where images, metadata, text information, user tagging profiles and content and text models are available for processing. We present the task challenges, the proposed data set and ground truth, the required participant runs and the evaluation metrics.

1. INTRODUCTION

An efficient image retrieval system should be able to present results that are both relevant and that are covering different aspects, i.e., *diversity*, of the query. By diversifying the pool of possible results, one can increase the likelihood of providing the user with the information needed. Relevance was more thoroughly studied in existing literature than diversification [1, 2, 3], especially within the text community. Even though a considerable amount of diversification literature exists [8, 9, 10], the topic remains important, especially in the emerging fields of social multimedia [4, 5, 6, 7, 11].

The 2016 Retrieving Diverse Social Images task is a followup of the 2015 edition [14, 13, 12, 15] and aims to foster new technology to improve both relevance and diversification of search results with explicit emphasis on the actual *social media context*. The task was designed to support evaluation of techniques emerging from a wide range of research fields, such as image retrieval (text, vision, multimedia communities), machine learning, relevance feedback and natural language processing, but not limited to these.

2. TASK DESCRIPTION

The task is built around the use case of a general ad-hoc image retrieval system, which provides the user with diverse representations of the queries (see for instance Google Image Search[1]). Participants are required, given a ranked list of query-related photos retrieved from Flickr[2], to refine the results by providing a set of images that are at the same time relevant to the query and to provide a diversified summary of it. Compared to the previous editions, this year's task includes complex and general-purpose multi-concept queries.

The requirements of the task are to refine these results by providing a ranked list of up to 50 photos that are both *relevant* and *diverse* representations of the query, according to the following definitions:

Relevance: a photo is considered to be relevant for the query if it is a common photo representation of the query topics (all at once). Bad quality photos (e.g., severely blurred, out of focus, etc.) are not considered relevant in this scenario;

Diversity: a set of photos is considered to be diverse if it depicts different visual characteristics of the query topics and subtopics with a certain degree of complementarity, i.e., most of the perceived visual information is different from one photo to another.

To carry out the refinement and diversification tasks, participants may use the social metadata associated with the images, the visual characteristics of the images, information related to user tagging credibility (an estimation of the global quality of tag-image content relationships for a user's contributions) or external resources (e.g., the Internet).

3. DATASET

The 2016 data consists of a development set (*devset*) containing 70 queries (20,757 Flickr photos — including 35 multi-topic queries related to events and states associated with locations from the 2015 dataset [14]), a user annotation credibility set (*credibilityset*) containing information for ca. 300 location-based queries and 685 users (different than the ones in *devset* and *testset* — updated version of the 2015 dataset [14]), a set providing semantic vectors for general English terms computed on top of the English Wikipedia[3] (*wikiset*), which could help the participants in developing advanced text models, and a test set (*testset*) containing 65

*This task is partly supported by the Vienna Science and Technology Fund (WWTF) through project ICT12-010.

Copyright is held by the author/owner(s).
MediaEval 2016 Workshop, October 20-21, 2016, Hilversum, Netherlands.

[1] https://images.google.com/.
[2] https://www.flickr.com/.
[3] https://en.wikipedia.org/.

queries (19,017 Flickr photos).

Each query is provided with the following information: query text formulation (the actual query formulation used on Flickr to retrieve all the data), a ranked list of up to 300 photos in jpeg format retrieved from Flickr using Flickr's default "relevance" algorithm (all photos are Creative Commons licensed allowing redistribution[4]), an xml file containing metadata from Flickr for all the retrieved photos (e.g., photo title, photo description, photo id, tags, Creative Common license type, the url link of the photo location from Flickr, the photo owner's name, user id, the number of times the photo has been displayed, etc), and ground truth for both relevance and diversity.

Apart from the metadata, to facilitate participation from various communities, we also provide the following content descriptors:

- *convolutional neural network based descriptors* — *generic* CNN based on the reference convolutional neural network (CNN) model provided along with the Caffe framework[5] (this model is learned with the 1,000 ImageNet classes used during the ImageNet challenge); and an *adapted* CNN based on a CNN model obtained with an identical architecture to that of the Caffe reference model. Adaptation is done only for the 2015 location-based multi-topic queries (35 queries from the devset), i.e., the model is learned with 1,000 tourist points of interest classes of which the images were automatically collected from the Web [16]. For the other queries, the descriptor is computed as the generic one, because queries are diverse enough and do not require any adaptation;

- *text information* that consists as in the previous edition of term frequency information, document frequency information and their ratio, i.e., TF-IDF, which is computed on per image basis, per query basis and per user basis (see [17]);

- *user annotation credibility descriptors* that give an automatic estimation of the quality of the users' tag-image content relationships. These descriptors are extracted by visual or textual content mining: *visualScore* (measure of user image relevance), *faceProportion* (the percentage of images with faces), *tagSpecificity* (average specificity of a user's tags, where tag specificity is the percentage of users having annotated with that tag in a large Flickr corpus), *locationSimilarity* (average similarity between a user's geotagged photos and a probabilistic model of a surrounding cell), *photoCount* (total number of images a user shared), *uniqueTags* (proportion of unique tags), *uploadFrequency* (average time between two consecutive uploads), *bulkProportion* (the proportion of bulk taggings in a user's stream, i.e., of tag sets that appear identical for at least two distinct photos), *meanPhotoViews* (mean value of the number of times a user's image has been seen by other members of the community), *meanTitleWordCounts* (mean value of the number of words found in the titles associated with users' photos), *meanTagsPerPhoto* (mean value of the number of tags users put for their images), *meanTagRank* (mean rank of a user's tags in a list in which the tags are sorted in descending order according the the number of appearances in a large subsample of Flickr images), and *meanImageTagClarity* (adaptation of the Image Tag Clarity from [18] using as individual tag language model a tf/idf language model).

4. GROUND TRUTH

Both relevance and diversity annotations were carried out by expert annotators. For *relevance*, annotators were asked to label each photo (one at a time) as being relevant (value 1), non-relevant (0) or with "don't know" (-1). For *devset*, 9 annotators were involved, for *credibilityset* 9 and for *testset* 8. The data was partitioned among annotators such that in the end each image has been marked by 3 different annotators. The final relevance ground truth was determined after a lenient majority voting scheme. For *diversity*, only the photos that were judged as relevant in the previous step were considered. For each query, annotators were provided with a thumbnail list of all relevant photos. After getting familiar with their contents, they were asked to re-group the photos into clusters with similar visual appearance (up to 25). *Devset* and *testset* were annotated by 5 persons, each of them annotating distinct parts of the data (leading to only one annotation). An additional annotator acted as a master annotator and reviewed once more the final annotations.

5. RUN DESCRIPTION

Participants were allowed to submit up to 5 runs. The first 3 are *required runs*: **run1** — automated using visual information only; **run2** — automated using text information only; and **run3** — automated using text-visual fused without other resources than provided by the organizers. The last 2 runs are *general runs*: **run4** and **run5** — everything allowed, e.g., human-based or hybrid human-machine approaches, including using data from external sources (e.g., Internet). For generating *run1* to *run3* participants are allowed to use only information that can be extracted from the provided data (e.g., provided descriptors, descriptors of their own, etc).

6. EVALUATION

Performance is assessed for both diversity and relevance. The following metrics are computed: Cluster Recall at X (CR@X) — a measure that assesses how many different clusters from the ground truth are represented among the top X results (only relevant images are considered), Precision at X (P@X) — measures the number of relevant photos among the top X results and F1-measure at X (F1@X) — the harmonic mean of the previous two. Various cut off points are to be considered, i.e., X=5, 10, 20, 30, 40, 50. *Official ranking metric* is the F1@20 which gives equal importance to diversity (via CR@20) and relevance (via P@20). This metric simulates the content of a single page of a typical Web image search engine and reflects user behavior, i.e., inspecting the first page of results with priority.

7. CONCLUSIONS

The 2016 Retrieving Diverse Social Images task provides participants with a comparative and collaborative evaluation framework for social image retrieval techniques with explicit focus on *result diversification*. This year in particular, the task explores the diversification in the context of a challenging, ad-hoc image retrieval system, which should be able to tackle complex and general-purpose multi-concept queries. Details on the methods and results of each individual participant team can be found in the working note papers of the MediaEval 2016 workshop proceedings.

[4] `http://creativecommons.org/`.

[5] `http://caffe.berkeleyvision.org/`.

8. REFERENCES

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-based Image Retrieval at the End of the Early Years", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), pp. 1349 - 1380, 2000.

[2] R. Datta, D. Joshi, J. Li, J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys, 40(2), pp. 1-60, 2008.

[3] R. Priyatharshini, S. Chitrakala, "Association Based Image Retrieval: A Survey", Mobile Communication and Power Engineering, Springer Communications in Computer and Information Science, 296, pp. 17-26, 2013.

[4] R.H. van Leuken, L. Garcia, X. Olivares, R. van Zwol, "Visual Diversification of Image Search Results", ACM World Wide Web, pp. 341-350, 2009.

[5] M.L. Paramita, M. Sanderson, P. Clough, "Diversity in Photo Retrieval: Overview of the ImageCLEF Photo Task 2009", ImageCLEF 2009.

[6] B. Taneva, M. Kacimi, G. Weikum, "Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity", ACM Web Search and Data Mining, pp. 431-440, 2010.

[7] S. Rudinac, A. Hanjalic, M.A. Larson, "Generating Visual Summaries of Geographic Areas Using Community-Contributed Images", IEEE Transactions on Multimedia, 15(4), pp. 921-932, 2013.

[8] R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, "Diversifying Search Results", ACM International Conference on Web Search and Data Mining, pp. 5-14, 2009.

[9] Y. Zhu, Y. Lan, J. Guo, X. Cheng, S. Niu, "Learning for Search Result Diversification", ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 293-302, 2014.

[10] H.-T. Yu, F. Ren, "Search Result Diversification via Filling up Multiple Knapsacks", ACM International Conference on Conference on Information and Knowledge Management, pp. 609-618, 2014.

[11] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, F.G.B. De Natale, "A Hybrid Approach for Retrieving Diverse Social Images of Landmarks", IEEE International Conference on Multimedia and Expo, pp. 1-6, 2015.

[12] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, "Div400: A Social Image Retrieval Result Diversification Dataset", ACM MMSys, Singapore, 2014.

[13] B. Ionescu, A. Popescu, M. Lupu, A.L. Gînscă, B. Boteanu, H. Müller, "Div150Cred: A Social Image Retrieval Result Diversification with User Tagging Credibility Dataset", ACM MMSys, Portland, Oregon, USA, 2015.

[14] B. Ionescu, A.L. Gînscă, B. Boteanu, M. Lupu, A. Popescu, H. Müller, "Div150Multi: A Social Image Retrieval Result Diversification Dataset with Multi-topic Queries", ACM MMSys, Klagenfurt, Austria, 2016.

[15] B. Ionescu, A. Popescu, A.-L. Radu, H. Müller, "Result Diversification in Social Image Retrieval: A Benchmarking Framework", Multimedia Tools and Applications, 2014.

[16] E. Spyromitros-Xioufis, S. Papadopoulos, A. Gînscă, A. Popescu, I. Kompatsiaris, I. Vlahavas, "Improving Diversity in Image Search via Supervised Relevance Scoring", ACM International Conference on Multimedia Retrieval, ACM, Shanghai, China, 2015.

[17] B. Ionescu, A. Popescu, M. Lupu, A.L. Gînscă, H. Müller, "Retrieving Diverse Social Images at MediaEval 2014: Challenge, Dataset and Evaluation", CEUR-WS, Vol. 1263, http://ceur-ws.org/Vol-1263/mediaeval2014_submission_1.pdf, Spain, 2014.

[18] A. Sun, S.S. Bhowmick, "Image Tag Clarity: in Search of Visual-Representative Tags for Social Images", SIGMM workshop on Social media, 2009.