Retrieving Diverse Social Images at MediaEval 2015: 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

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

Adrian Popescu
CEA, LIST, France
adrian.popescu@cea.fr

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

Henning Müller
HES-SO, Sierre, 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 2015 Benchmarking Initiative for Multimedia Evaluation. The task addresses the problem of result diversification in the context of social photo retrieval. 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., diverse, of the query. Relevance was more thoroughly studied in existing literature than diversification [1, 2, 3] and even though a considerable amount of diversification literature exists, the topic remains important, especially in social multimedia [4, 5, 6, 7].

The 2015 Retrieving Diverse Social Images task is a followup of last years' editions [9, 8, 10] and aims to foster new technology for improving both relevance and diversification of search results with explicit emphasis on the actual *social media context*. Researchers will find this task interesting if they work in either machine-based or human-based media analysis, including areas such as: image retrieval (text, vision, multimedia communities), re-ranking, machine learning, relevance feedback, natural language processing, crowdsourcing and automatic geo-tagging.

2. TASK DESCRIPTION

The task is built around a tourist use case where a person tries to find more information about a place she is potentially visiting. Before deciding whether this location suits her needs, the person is interested in getting a more complete and diversified visual description of the place.

Participants are required to develop algorithms to automatically refine a list of images that has been returned by Flickr in response to a query. Proposed queries include either single-topic formulations such as the name of a location as well as multi-concept queries related to events and states associated with locations. 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 if it is a common photo representation of all query concepts at once. This includes sub-locations (e.g., subsuming indoor/outdoor, close up), temporal information (e.g., historical shots, times of day), typical actors/objects (e.g., people who frequent the location, vehicles), genesis information (e.g., images showing how something got the way it is), and image style information (e.g., creative views). Low quality photos (e.g., severely blurred, out of focus, etc) as well as photos with people as the main subject (e.g., a big picture of me in front of the monument) 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 target concepts, e.g., sub-locations, temporal information, typical actors/objects, genesis and style information, etc 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 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., Internet).

3. DATASET

The 2015 data consists of a development set (*devset*) containing 153 location queries (45,375 Flickr photos — the 2014 dataset [9]), a user annotation credibility set (*credibilityset*) containing information for ca. 300 locations and 685 users (other that the ones in *devset* and *testset*) and a test set (*testset*) 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).

Each query is provided with the following information: query text formulation (used to retrieve the data), GPS coordinates (latitude and longitude in degrees - only for one-topic location queries), a link to a Wikipedia webpage (only when available), up to 5 representative photos from Wikipedia (only for one-topic location queries), a ranked list of up to 300 photos retrieved from Flickr using Flickr's

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

default "relevance" algorithm[1], and 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, number of posted comments, 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).

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

- *general purpose visual descriptors* (e.g., color, texture and feature information) identical to the ones in the previous editions [10];
- *convolutional neural network based descriptors - generic* based on the reference convolutional neural network (CNN) model provided along with the Caffe framework (this model is learned with the 1,000 ImageNet classes used during the ImageNet challenge) and *adapted* CNN based on a CNN model obtained with an identical architecture to that of the Caffe reference model (this model is learned with 1,000 tourist points of interest classes whose images were automatically collected from the Web) [11];
- *text information* which consists as in the previous edition of term frequency information (the number of occurrences of the term in the entity's text fields), document frequency information (the number of entities which have this term in their text fields) and their ratio, i.e., TF-IDF [12];
- *user annotation credibility descriptors* that give an automatic estimation of the quality of 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 which 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 [13] 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 with advanced knowledge of the location characteristics (mainly learned from last years' tasks and Internet sources). 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*, 11 annotators were involved, for *credibilityset* 9 and for *testset* one-topic 7 and multi-topic 5. Each annotator annotated different parts of the data leading in the end to 3 different annotations. Final 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 location, 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 3 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.

[1]all the photos are under Creative Commons licenses that allow redistribution, see `http://creativecommons.org/`.

5. RUN DESCRIPTION

Participants are 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** - automated using user annotation credibility descriptors (either the ones provided by organizers or computed by the participants) 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 *run4* participants are allowed to use only information that can be extracted from the provided data (e.g., provided descriptors, descriptors of their own, etc). This includes also the Wikipedia webpages of the locations (provided via their links).

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 2015 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 also the diversification of 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 2015 workshop proceedings.

Acknowledgments

This task is supported by the CHIST-ERA FP7 MUCKE - Multimodal User Credibility and Knowledge Extraction project (`http://ifs.tuwien.ac.at/~mucke/`).

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 Trans. 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 AgeŤ, ACM Comput. Surv., 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 Trans. on Multimedia, 15(4), pp. 921-932, 2013.

[8] 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.

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

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

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

[12] 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.

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