Retrieving Diverse Social Images at MediaEval 2014: Challenge, Dataset and Evaluation

Bogdan Ionescu
LAPI, University Politehnica of Bucharest, Romania
bionescu@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

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

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 2014 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 *diverse* aspects of a query (e.g., sub-topics). Relevance was more thoroughly studied in existing literature than diversification and even though a considerable amount of diversification literature exists, the topic remains an important one, especially in social media. The 2014 Retrieving Diverse Social Images task is a followup of last year's edition [1][2][3] and aims to foster new technology for improving both relevance and diversification of search results with explicit emphasis on the actual *social media context*. It creates an evaluation framework specifically designed to encourage the emerging of new diversification solutions from areas such as information retrieval (text, vision and multimedia), re-ranking, relevance feedback, crowdsourcing.

2. TASK DESCRIPTION

The task is build around a tourist use case where a person tries to find more information about a place she is potentially visiting. The person has only a vague idea about the location, knowing the name of the place. She uses the name to learn additional facts about the place from the Internet, for instance by visiting a Wikipedia page, e.g., getting a photo, the geographical position of the place and basic descriptions. Before deciding whether this location suits her needs, the person is interested in getting a more complete and diversified visual description of the place.

In this task, participants receive a list of photos for a certain location retrieved from Flickr and ranked with Flickr's default "relevance" algorithm. These results are typically noisy and redundant. The requirements of the task are to refine these results by providing a set of images that are in the same time *relevant* and provide a *diversified* summary (up to 50 images), according to the following definitions:

Relevance: a photo is considered to be relevant if it is a common photo representation of the location, e.g., different views at different times of the day/year and under different weather conditions, inside views, close-ups on architectural details, drawings, sketches, creative views, etc, which contain partially or entirely the target location. Bad 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;

Diversity: a set of photos is considered to be diverse if it depicts different visual characteristics of the target location, as stated by the relevance definition above, with a certain degree of complementarity, i.e., most of the perceived visual information is different from one photo to another.

The refinement and diversification process will be based on the social metadata associated with the images and/or on the visual characteristics of the images. New for this year, we provide information about *user annotation credibility*. Credibility is determined as an automatic estimation of the quality (correctness) of a particular user's tags. Participants are allowed to exploit this credibility estimation or to compute their own approach, in addition to classical retrieval techniques.

3. DATASET

The 2014 data set is constructed around the 2013 data [1] [2] and consists of ca. 300 locations (e.g., monuments, cathedrals, bridges, sites, etc) spread over 35 countries around the world. Data is divided into a development set, *devset*, containing 30 locations — intended for designing the approaches; a test set, *testset*, containing 123 locations — to be used for the official evaluation; as well as an additional *credibilityset*, ca. 300 locations and 685 users (chose to be different from the ones in devset and testset), used to train the credibility descriptors. All the data was retrieved from Flickr using the name of the location as query.

Each location contains: the name of the location, its GPS coordinates, a link to a Wikipedia webpage, up to 5 representative photos from Wikipedia, a ranked list of up to 300 photos retrieved from Flickr using Flickr's default "relevance" algorithm[1] (*devset* provides 8,923 images and *testset*

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

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

36,452) 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, the dataset contains also content descriptors (visual, text and credibility based). *Visual descriptors* include the same general purpose descriptors (e.g., color, texture and feature information) as in 2013 [3]. *Text information* consists this year 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. Text descriptors are computed on per dataset basis and also per image basis, per location basis and per user basis, respectively. *User annotation credibility descriptors* provide an automatic estimation of the quality of tag-image content relationships. This information gives an indication about which users are most likely to share relevant images in Flickr according to the underlying task scenario. The following descriptors are provided: 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) and 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).

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 year's task and Internet sources). Specifically designed visual tools were employed to facilitate the annotation process. Annotation was not time restricted.

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). To help with their decisions, annotators were able to consult any additional information source during the evaluation (e.g., representative photos, Internet, etc). For *devset*, 3 annotators were involved while *testset* and *credibilityset* used 11 and 9 annotators, respectively, that 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 similar visual appearance clusters (up to 25) and tag these clusters with appropriate keywords that justify their choices. *Devset* was annotated by 2 persons and *testset* by 3. Each person annotated 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 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 in priority.

7. CONCLUSIONS

The 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 benefits of employing automatically estimated user annotation credibility information to the diversification task. Details on the methods and results of each individual participant team can be found in the working note papers of the MediaEval 2014 workshop proceedings.

Acknowledgments

This task is supported by the following projects: MUCKE[2], CUbRIK[3] and PROMISE[4].

8. REFERENCES

[1] 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.

[2] B. Ionescu, A. Popescu, H. Müller, M. Menéndez, A.-L. Radu, "Benchmarking Result Diversification in Social Image Retrieval", IEEE ICIP, France, 2014.

[3] B. Ionescu, M. Menéndez, H. Müller, A. Popescu, "Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation", CEUR-WS, Vol. 1043, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_3.pdf Spain, 2013.

[2] http://ifs.tuwien.ac.at/~mucke/

[3] http://www.cubrikproject.eu/

[4] http://www.promise-noe.eu/