ACM SIGMM Records

<-- Back to table of contents for Preview of ACM SIGMM Records, Issue 3, 2017

Diversity and Credibility for Social Images and Image Retrieval

Authors:
Bogdan Ionescu - University Politehnica of Bucharest, Romania
Mihai Lupu - Vienna University of Technology, Austria
Maia Rohm - Vienna University of Technology, Austria
Alexandru Lucian Gînscă - CEA LIST, France
Henning Müller, University of Applied Sciences Western Switzerland in Sierre, Switzerland.
Editors: *Martha Larson and Bart Thomee*

Social media has established itself as an inextricable component of today's society. Images make up a large proportion of items shared on social media [1]. The popularity of social image sharing has contributed to the popularity of the Retrieving Diverse Social Images task at the MediaEval Benchmarking Initiative for Multimedia Evaluationa [2]. Since its introduction in 2013, the task has attracted a large participation and has published a set of datasets of outstanding value to the multimedia research community.

The task, and the datasets it has released, target a novel facet of multimedia retrieval, namely the search result diversification of social images. The task is defined as follows: Given a large number of images, retrieved by a social media image search engine, find those that are not only relevant to the query, but also provide a diverse view of the topic/topics behind the query (see an example in Figure 1). The features and methods needed to address the task successfully are complex and span different research areas (image processing, text processing, machine learning). For this reason, when creating the collections used in the Retrieving Diverse Social Images Tasks, we also created a set of baseline features. The features are released with the datasets. In this way, task participants who have expertise in one particular research area may focus on that area and still participate in the full evaluation.

Important links

Home

SIGMM main page

ISSN 1947-4598

Search Records by keyword

Search OK

Figure 1: Example of retrieval and diversification results for query “Pingxi Sky Lantern Festival” (results are truncated to the first 14 images for better visualization): (top images) Flickr initial retrieval results; (bottom images) diversification achieved with the approach from the TUW team (best approach at MediaEval 2015).

The collections

Before describing the individual collections, it needs to be noted that all data consist of redistributable Creative Commons Flickr and Wikipedia content and are freely available for download (follow the instructions here [3]). Although the task ran also in 2017, we focus in the following on the datasets already released, namely: Div400, Div150Cred, Div150Multi and Div150Adhoc (corresponding to the 2013-2016 evaluation campaigns). Each of the four datasets available so far covers different aspects of the diversification challenge, either from the perspective of the task/use-case addressed, or from the data that can be used to address the task. Table 1 gives an overview of the four datasets that we describe in more detail over the next four subsections. Each of the datasets is divided into a development set and a test set. Although the division of development and test data is arbitrary, for comparability of results and full reproducibility, users of the collections are advised to maintain the separation when performing their experiments.

Table 1: Dataset statistics (devset – development data, testset – testing data, credibilityset – data for estimating user tagging credibility, single (s) – single topic queries, multi (m) – multi-topic queries, ++ – enhanced/updated content, POI – location point of interest, events – events and states associated with locations, general – general purpose ad-hoc topics).

	Div400 (2013)		Div150Cred (2014)			Div150Multi (2015)				Div150Adhoc (2016)		
	devset	*testset*	*devset*	*testset*	*credibilityset*	*devset*	$testset_s$	$testset_m$	*credibilityset*	*devset*	*testset*	*credibilityset*
data source	2013	2013	2013++	2013++	2014	2014all	2015	2015	2014++	2015/2016	2016	2015++
#queries	50	346	30	123	300 POIs, 685 users, ~3.6M + ~12.3M (via *devset* & *testset*) image urls & metadata	153	69	70	300 POIs, 685 Flickr users, ~3.6M + ~27.1M (via *devset* & *testset*) image urls & metadata	70	64	300 POIs, 685 Flickr users, ~3.6M image urls
content type	POI single	POI single	POI single	POI single		POI single	POI single	events multi		events/general multi	general multi	
#Wiki.img./query	1	1	1-5	1-5		1-5	1-5	-		-	-	
#images	5,118	38,300	8,923	36,452		45,375	20,700	20,694		20,757	18,717	
min #img./query	30	30	285	277		281	300	176		176	141	
avg. #img./query	102.4	110.7	297	296		297	300	296		297	292	
max #img./query	150	150	300	300		300	300	300		300	300	
descriptors	visual TF-IDF		visual, TF-IDF credibility			visual, TF-IDF credibility, CNN				TF-IDF & semantic vectors credibility, CNN		

Div400

In 2013, the task started with a narrowly defined use-case scenario, where a tourist, upon deciding to visit a particular location, reads the corresponding Wikipedia page and desires to see a diverse set of images from that location. Queries here might be *"Big Ben in London" or "Palazzo delle Albere in Italy"*. For each such query, we know the GPS coordinates, the name, and the Wikipedia page, including an example image of the destination. As a search pool, we consider the top 150 photos obtained from Flickr using the name as a search query. These photos come with some metadata (photo ID, title, description, tags, geotagging information, date when the photo was taken, owner's name, number of times the photo has been displayed, URL in Flickr, license type, number of comments on the photo) [4].

In addition to providing the raw data, the collection also contains visual and text features of the data, such that researchers who are only interested in one of the two, can use the other without investing additional time in generating a baseline set of features.

As *visual descriptors*, for each of the images in the collection, we provide:

Global color naming histogram
Global histogram of oriented gradients
Global color moments on HSV
Global Locally Binary Patterns on gray scale
Global Color Structure Descriptor
Global statistics on gray level Run Length Matrix (Short Run Emphasis, Long Run Emphasis, Gray-Level Non-uniformity, Run Length Non-uniformity, Run Percentage, Low Gray-Level Run Emphasis, High Gray-Level Run Emphasis, Short Run Low Gray-Level Emphasis, Short Run High Gray-Level Emphasis, Long Run Low Gray-Level Emphasis, Long Run High Gray-Level Emphasis)
Local spatial pyramid representations (3×3) of each of the previous descriptors

As *textual descriptors* we provide the classic Term Frequency ($TF_{t,d}$ – the number of occurrences of term *t* in document *d*) and Document Frequency (DF_t – the number of documents containing term *t*). Note that the datasets are not limited to a single notion of *document*. The most direct definition of a "document" is an image that can be either retrieved or not retrieved. However, it is easily conceivable that the relative frequency of a term in the set of images corresponding to one topic, or the set of images

corresponding to one user might also be of interest in ranking the importance of a result to a query. Therefore, the collection also contains statistics that take a document to be a topic, as well as a user. All these are provided both as CSV files, as well as Lucene Index files. The former can be used as part of a custom weighting scheme, while the latter can be deployed directly in a Lucene/Solr search engine to obtain results based on the text without further effort.

Div150Cred

The tourism use case also underlies Div150Cred, but a component addressing the concept of user tagging credibility is added. The idea here is that not all users tag their photos in a manner that is useful for retrieval and, for this reason, it makes sense to consider, in addition to the visual and text descriptors also used in Div400, another feature set – a user credibility feature. Each of the 153 topics (30 in the development set and 123 in the test set) comes therefore, in addition to the visual and text features of each image, with a value indicating the credibility of the user. This value is estimated automatically based on a set of features, so in addition to the retrieval development and test sets, DIV150Cred also contains a *credibility set*, used by us to generate the credibility of each user, and which can be used by any interested researcher to generate better credibility estimators.

The credibility set contains images for approximately 300 locations from 685 users (a total of 3.6 million images). For each user there is a manually assigned credibility score as well as an automatically estimated one, based on the following features:

Visual score – learned predictor of a user's consistent and relevant tagging behavior
Face proportion
Tag specificity
Location similarity
Photo count
Unique tags
Upload frequency
Bulk proportion

For each of these, the intuition behind it and the actual calculation is detailed in the collection report [5].

Div150Multi

Div150Multi adds another twist to the task of the search engine and its tourism use-case. Now, the topics are not simply points of interest, but rather a combination of a main concept and a qualifier, namely multi-topic queries about location specific events, location aspects or general activities (e.g., "*Oktoberfest in Munich*", "*Bucharest in winter*"). In terms of features however, the collection builds on the existing ones used in Div400 and Div150Cred, but adds to the pool of resources the researchers have at their disposal. In terms of credibility, in addition to the 8 features listed above, we now also

have:

Mean Photo Views
Mean Title Word Counts
Mean Tags per Photo
Mean Image Tag Clarity

Again, for details on the intuition and formulas behind these, the collection report [6] is the reference material.

A new set of descriptors has been now made available, based on convolutional neural networks.

CNN generic: a descriptor based on the reference convolutional (CNN) neural network model provided along with the Caffe framework [7]. This model is trained with the 1,000 ImageNet classes used during the ImageNet challenge. The descriptors are extracted from the last fully connected layer of the network (named fc7).
CNN adapted: These features were also computed using the Caffe framework, with the reference model architecture but using images of 1,000 landmarks instead of ImageNet classes. We collected approximately 1,200 Web images for each landmark and fed them directly to Caffe for training [8]. Similar to CNN generic, the descriptors were extracted from the last fully connected layer of the network (i.e., fc7).

Div150AdHoc

For this dataset, the definition of relevance was expanded from previous years, with the introduction of even more challenging multi-topic queries unrelated to POIs. These queries address the diversification problem for a general ad-hoc image retrieval system, where general-purpose multi-topic queries are used for retrieving the images (e.g., “animals at Zoo”, “flying planes on blue sky”, “hotel corridor”). The Div150Adhoc collection includes most of the previously described credibility descriptors, but drops *faceProportion* and *location-Similarity*, as they were no longer relevant for the new retrieval scenario. Also, the *visualScore* descriptor was updated in order to keep up with the latest advancements on CNN descriptors. Consequently, when training individual visual models, the Overfeat visual descriptor is replaced by the representation produced by the last fully connected layer of the network [9]. Full details are available in the collection report [10].

Ground-truth and state-of-the-art

Each of the above collections comes with an associated ground-truth, created by human assessors. As the focus is on both relevance and diversity, the ground truth and the metrics used reflect it: Precision at cutoff (primarily P@20) is used for relevance, and Cluster Recall at cutoff (primarily CR@20) is used for diversity.

Figure 2 shows an overview of the results obtained by participants in the evaluation campaigns over the period 2013-2016, and serves as a baseline for future experiments on these collections. Results presented here are on the test set alone. The reader may find more information about the methods in the MediaEval proceedings, which are listed on the Retrieving Diverse Social Images yearly task pages on the MediaEval website (http://multimediaeval.org/).

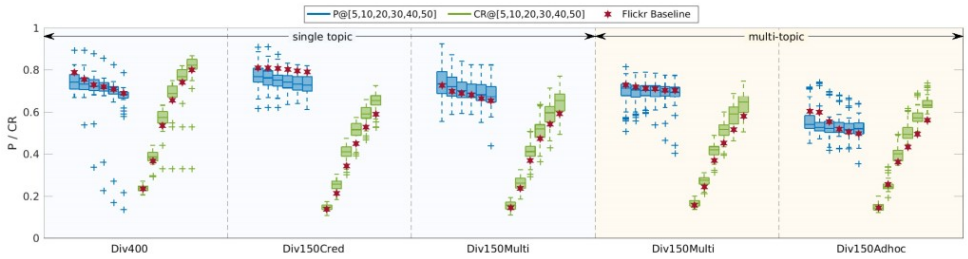

Figure 2. Evolution of the diversification performance (boxplots — the interquartile range (IQR), i.e. where the 50% of the values are; the line within the box = median; the tails = 1.5*IQR; the points outside (+) = outliers) for the different datasets in terms of precision (P) and cluster recall (CR) at different cut-off values. Flickr baseline represents the initial Flickr retrieval result for the corresponding dataset.

Conclusions

The Retrieving Diverse Social Image task datasets, as their name indicates, address the problem of retrieving images taking into account both the need to diversify the results presented to the user, as well as the potential lack of credibility of the users in their tagging behavior. They are based on already state-of-the-art retrieval technology (i.e., the Flickr retrieval system), which makes it possible to focus on the challenge of image diversification. Moreover, the data sets are not limited to images, but rather also include rich social information. The credibility component, represented by the credibility subsets of the last three collections, is unique to this set of benchmark datasets.

Acknowledgments

The Retrieving Diverse Social Image task datasets were made possible by the effort of a large team of people over an extended period of time. The contributions of the authors were essential. Further, we would like to acknowledge the multiple team members who have contributed to annotating the images and making the MediaEval Task possible. Please see the yearly Retrieving Diverse Social Images task pages on the MediaEval website (http://multimediaeval.org/).

Contact

Should you have any inquires or questions about the datasets, don't hesitate to contact us via email at: bionescu at imag dot pub dot ro.

References

[1] http://contentmarketinginstitute.com/2015/11/visual-content-strategy/ (last visited 2017-11-29).

[2] http://www.multimediaeval.org/

[3] http://imag.pub.ro/~bionescu/index_files/Page6657.htm

[4] http://imag.pub.ro/~bionescu/index_files/Page13170.htm

[5] http://imag.pub.ro/~bionescu/index_files/Page13170.htm

[6] http://imag.pub.ro/~bionescu/index_files/Page13288.htm

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding” in ACM International Conference on Multimedia, 2014, pp. 675–678.

[8] E. Spyromitros-Xioufis, S. Papadopoulos, A. L. Ginsca, A. Popescu, Y. Kompatsiaris, and I. Vlahavas, “Improving diversity in image search via supervised relevance scoring” in ACM International Conference on Multimedia Retrieval, 2015, pp. 323–330.

[9] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets” arXiv preprint arXiv:1405.3531, 2014.

[10] http://imag.pub.ro/~bionescu/index_files/Page15577.htm.

<-- Back to table of contents for Preview of ACM SIGMM Records, Issue 3, 2017