Div150Cred: A Social Image Retrieval Result Diversification with User Tagging Credibility Dataset

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

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

Henning Müller
HES-SO, Sierre, Switzerland
henning.mueller@hevs.ch

ABSTRACT

In this paper we introduce a new dataset and its evaluation tools, Div150Cred, that was designed to support shared evaluation of diversification techniques in different areas of social media photo retrieval and related areas. The dataset comes with associated relevance and diversity assessments performed by human annotators. The data consists of 300 landmark locations represented via 45,375 Flickr photos, 16M photo links for around 3,000 users, metadata, Wikipedia pages and content descriptors for text and visual modalities. To facilitate distribution, only Creative Commons content was included in the dataset. The proposed dataset was validated during the 2014 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*collection, dissemination.*

General Terms

Algorithms, Experimentation, Performance.

Keywords

social photo retrieval, search result diversification, user tagging credibility, MediaEval benchmark, Flickr retrieval.

1. INTRODUCTION

Media search is currently a hot topic especially in the context of the ever growing social media. Media platforms are one of the most prolific means for sharing and searching information over the Internet. Until recently, research focused mainly on how to improve the *relevance* of the search results. However, an efficient information retrieval system should be able to surface results that are both relevant and that are covering *different aspects* (i.e., *diversity*) of a query. A reason for this is the fact that most of the queries involve many declinations such as sub-topics, e.g., animals are of different species, points of interest can be photographed from different angles and so on. Therefore, to improve search efficiency, one should consider equally topic diversification in a retrieval scenario. An example is this year's Google Image Search[1], which started to integrate result categorization.

In this paper we introduce a new benchmarking dataset designed to support this emerging area of social image retrieval focusing on diversification. The dataset may be used to validate methods from different communities, e.g., information retrieval (text, vision and multimedia), social media and networking, re-ranking, relevance feedback, crowdsourcing, automatic geo-tagging, recommender systems.

The remainder of the paper is organized as follows: Section 2 presents a brief overview of the literature and situates our contribution. Section 3 describes the proposed dataset while Section 4 deals with the creation of ground truth. Section 5 discusses its validation during the 2014 MediaEval benchmark campaign while Section 6 concludes the paper.

2. RELATED WORK

One of the critical points of the diversification approaches [1, 2, 4] are the evaluation tools. In general, experimental validation is carried out on very particular and closed datasets while ground truth annotations are restraint, which limits the reproducibility of the results. There are however a few attempts to constitute a standardized evaluation framework such as the ImageCLEF benchmarking and in particular the 2009 Photo Retrieval task [3]. It proposes a dataset consisting of 498,920 news photographs (images and caption text) classified into sub-topics (e.g., location type for locations, animal type for photos of animals). Other existing datasets are determined for the experimentation of specific methods. Rudinac et al [1] introduce a collection of Flickr[2] images captured around 207 locations in Paris (with 100 images per location). Ground truth is determined automatically by exploiting the geographical coordinates accompanying the images. Taneva et al [2] address the diversifica-

MMSys'15, March 18 - 20 2015, Portland, OR, USA
Copyright 2015 ACM 978-1-4503-3351-1/15/03$15.00
http://dx.doi.org/10.1145/2713168.2713192.

[1] https://www.google.com/imghp.

[2] https://www.flickr.com/

tion problem in the context of populating a knowledge base, YAGO[3], containing about 2 million typed entities (e.g., people, buildings, mountains, lakes, etc). van Leuken et al [4] use a data set of 75 randomly selected queries from Flickr logs for which only the top 50 results are retained. Diversity annotation is provided by human assessors that grouped the data into clusters with similar appearance.

In this context, the added value of the proposed dataset is: (i) it focuses on improving the current technology by using state-of-the-art Flickr's relevance system as baseline; (ii) while smaller in size than the ImageCLEF collections [3], it contains images that are already associated with topics by Flickr thus pushing diversification into the foreground; (iii) unlike ImageCLEF, that worked with generic ad-hoc retrieval scenarios, a focused real-world usage scenario is set up, i.e., tourism, to disambiguate the diversification need; (iv) addresses the social dimension of the diversification problem that is reflected both in the nature of the data and in the methods devised to retrieve.

This dataset is built on the base of the Div400 dataset [5]. Apart from providing the diversification of a significantly higher number of images per query and more diversified content descriptors, the main novelty of Div150Cred is in its stronger emphasis on the user/social context of the data. It proposes information about user image tagging credibility (as a dedicated dataset but also via specific descriptors) by providing metadata for 16M images and more than 3.000 users. Preliminary results [11] show that user credibility is a valuable lead for improving results' diversity.

3. DATASET DESCRIPTION

To disambiguate the diversification need, we have selected as use case for the proposed data, the search for images with tourist landmarks. As shown with our previous data [5], landmark locations are popular with social media platforms and also benefit from spatio-temporal and visual invariance which makes them suitable for benchmarking scenarios.

The dataset consists of information for 300 landmark locations in 35 different countries, natural or man-made, e.g., sites, museums, monuments, buildings, roads, bridges, houses, caves. Locations range from very famous, e.g., "Notre Dame de Paris" in France, to lesser known to the grand public, e.g., "Circo Massimo" in Italy. These locations were selected from Internet sources, the majority of them being listed with the World Heritage Site of the United Nations Educational, Scientific and Cultural Organization (UNESCO)[4]. Locations are unevenly distributed around the world and were selected based on the number of redistributable photos available on Flickr.

The dataset[5] consists of redistributable Creative Commons[6] Flickr and Wikipedia location data. For each location, the following information is provided: *location keyword* (unique textual identifier in the dataset), *location number* (unique numeric identifier), *GPS coordinates* (latitude and longitude in degrees) retrieved from GeoHack[7] via the location Wikipedia web page, a link to its *Wikipedia web page*, up to 5 *representative photos* from Wikipedia, a *ranked set of photos* retrieved from Flickr (up to 300), *metadata* from Flickr for all the retrieved photos and visual/text content descriptors.

3.1 Flickr data collection method

Apart from Wikipedia data, landmark information was collected from Flickr using the Flickr API[8] (under Java) and the *flickr.photos.search* function. Information was retrieved using the location name as query and ranked with Flickr's default relevance algorithm. Therefore, the dataset is built on top of the current state-of-the-art retrieval technology. Having Flickr results as baseline will encourage approaches that push the field forward.

For each location, we retain depending on their availability at most the first 300 photo results. All the retrieved photos are under Creative Commons licenses of type 1 to 7, which allow redistribution[8]. For each photo, the retrieved metadata consist of the *photo's id* and *title*, photo *description* as provided by author, *tags*, geotagging information (*latitude* and *longitude* in degrees), the *date* the photo was taken, photo *owner's name* (username) and *id* (userid), the *number of times* the photo has been displayed, the *url link* of the photo from Flickr[9], Creative Common *license type*, number of *posted comments* and the photo's *rank* within the Flickr results (a number from 1 to 300).

3.2 Visual and text descriptors

3.2.1 Visual descriptors

For each photo, we provide the following descriptors (for more details see [5]): *global color naming histogram* (code *CN* — 11 values): maps colors to 11 universal color names; *global Histogram of Oriented Gradients* (code *HOG* — 81 values): represents the HoG feature computed on 3 by 3 image regions; *global color moments* on HSV (Hue-Saturation-Value) color space (code *CM* — 9 values): represent the first three central moments of an image color distribution: mean, standard deviation and skewness; *global Locally Binary Patterns* on gray scale (code *LBP* — 16 values); *global Color Structure Descriptor* (code *CSD* — 64 values): represents the MPEG-7 Color Structure Descriptor computed on the HMMD (Hue-Min-Max-Difference) color space; *global statistics on gray level Run Length Matrix* (code *GLRLM* — 44 dimensions): provides 11 statistics computed on gray level run-length matrices for 4 directions; *local spatial pyramid representations* (code 3x3) of each of the previous descriptors (image is divided into 3 by 3 non-overlapping blocks and descriptors are computed on each patch — the global descriptor is obtained by the concatenation of all values).

3.2.2 Text models

Before computing the text models, data underwent the following pre-processing steps: tokenisation according to the Unicode standard annex UAX#29[10] as implemented by the Lucene StandardTokenizerFactory; lowercasing of all terms,

[3] http://datahub.io/dataset/yago/

[4] http://en.wikipedia.org/wiki/List_of_World_Heritage_Sites/

[5] to download the dataset see http://traces.cs.umass.edu/index.php/Mmsys/Mmsys

[6] http://creativecommons.org/

[7] http://tools.wmflabs.org/geohack/

[8] http://www.flickr.com/services/api/flickr.photos.licenses.getInfo.html/

[9] Please note that by the time you use the dataset some of the photos may not be available anymore at the same urls.

[10] http://unicode.org/reports/tr29/#Word_Boundaries

Table 1: Devset and testset image statistics.

devset			testset		
#locations	*#images*	*min-avg.-max #images/location*	*#locations*	*#images*	*min-avg.-max #images/location*
30	8,923	285 - 297 - 300	123	36,452	277 - 296 - 300

Table 2: Credibilityset statistics.

credibilityset			
#locations	*#image urls*	*#users*	average #images per user
300	3,651,303	685	5,330

removal of stop words, and finally minimal English stemming [6]. For each term, three values are provided: the term frequency (*TF*) — the number of times it appears in the document, the document frequency (*DF*) — the number of documents in which the term appears, and the TF-IDF, calculated simply as TF/DF.

Three sets of term weights are provided, by considering three different interpretations of document. The default interpretation is that a document is an *image*. In this case, TF is the number of occurrences of a term in the Flickr *description*, *title* or *tags* of an image, and the DF is the number of images that mention this term. The second interpretation is considering the document to be the *location*. We concatenate all titles, descriptions and tags of images assigned to a location and take the TF to be the number of occurrences of a term in this general description. The third interpretation for document is a *user*. As in the case of the location interpretation, we concatenate here all the titles, descriptions and tags of images published by a particular user. Therefore, DF here means the number of users mentioning a specific term in their text. The location- and user-based interpretations cannot be directly used to rank images, but are available to provide additional context to image ranking, or to estimate diversity.

3.2.3 User credibility information

Credibility information on user tagging attempts to give an automatic estimation of the global quality of tag-image content relationships for a user's contributions [11]. This information is in particular valuable for exploiting the social context of the data. It gives an indication about which users are most likely to share representative images in Flickr, according to the underlying use case of the data.

The following descriptors are extracted by visual or textual content mining:

- *visualScore*: descriptor obtained through visual mining using over 17,000 ImageNet[11] visual models obtained by learning a binary Support Vector Machine (SVM) per ImageNet concept. Visual models are built on top of overfeat, a powerful convolutional Neural Network feature[12]. At most 1,000 images are downloaded for each user in order to compute *visualScores*. For each Flickr tag identical to an ImageNet concept, we obtain a classification confidence score that can be seen as the likelihood of the concept being visually depicted in the image. The *visualScore* of a user is obtained by averaging individual tag scores. The intuition here is that the higher the predicted scores are the more relevant a user's images should be. Scores are normalized between 0 and 1, with higher scores corresponding to more credible users;
- *faceProportion*: descriptor obtained using the same set of images as for *visualScore*. The default face detector from OpenCV[13] is used here to detect faces. *faceProportion*, the percentage of images with faces out of the total of images tested for each user is computed. The intuition here is that the lower *faceProportion* is, the better the average relevance of a user's photos is. *faceProportion* is normalized between 0 and 1, with 0 standing for no face images;
- *tagSpecificity*: descriptor obtained by computing the average specificity of a user's tags. Tag specificity is calculated as the percentage of users having annotated with that tag in a large Flickr corpus (∼100 million image metadata from 120,000 users);
- *locationSimilarity*: descriptor obtained by computing the average similarity between a user's geotagged photos and a probabilistic model of a surrounding cell of approximately 1 km^2 geotagged images. These models were created using the model in [7]. The intuition here is that the higher the coherence between a user's tags and those provided by the community is, the more relevant her images are likely to be. *locationSimilarity* is not normalized and low values stand for the smallest similarity;
- *photoCount*: descriptor that accounts for the total number of images a user shared on Flickr. This descriptor has a maximum value of 10,000;
- *uniqueTags*: proportion of unique tags present in a user's vocabulary divided by the total number of tags of the user. *uniqueTags* ranges between 0 and 1;
- *uploadFrequency*: average time between two consecutive uploads in Flickr. This descriptor is not normalized;
- *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). Is normalized between 0 and 1.

3.3 Dataset basic statistics

Development and test data. We provide a development set (code *devset*) containing Flickr and Wikipedia information (as described in the previous sections) for 30 locations. Its objective is to serve for the design and training of potential approaches; and a test set (code *testset*) that contains information for 123 locations and is intended for the actual benchmarking and validation of the methods. Some basic image statistics are presented in Table 1. In total, *devset* and *testset* account for 45,375 images.

User tagging credibility data. We provide a specially designed dataset (code *credibilityset*) that addresses the estimation of user tagging credibility. It provides Flickr photo information (the *date* the photo was taken, *tags*, *user's id* and photo *title*, the *number of times* the photo has been displayed, *url link* of the photo location, *GPS coordinates*) for about ca. 300 locations and 685 different users. Each user is assigned a manual credibility score which is determined

[11] `http://www.image-net.org/`

[12] `http://cilvr.nyu.edu/doku.php?id=code:start`

[13] `http://docs.opencv.org/trunk/doc/py_tutorials/py_objdetect/py_face_detection/py_face_detection.html`

as the average relevance score of all the user's photos (relevance annotations are determined as presented in Section 4). To obtain these scores, only 50,157 manual annotations are used (on average 73 photos per user). Apart from this information, each user is also provided with the estimated credibility descriptors introduced in Section 3.2.3. This dataset is intended for training and designing user credibility related descriptors. Some basic statistics are presented in Table 2. In total, it provides links to 3,651,303 images.

To give an indicator of the usefulness of this information, we present the Pearson correlation[14] of the proposed credibility descriptors with the manual relevance annotation scores (values range between -1 and 1 whereas 0 implies that there is no linear correlation between the two): *visualScore* 0.3663, *faceProportion* −0.2687, *tagSpecificity* −0.2883, *locationSimilarity* 0.1329, *photoCount* −0.2615, *uniqueTags* 0.2523, *uploadFrequency* 0.2563, *bulkProportion* −0.1284. A positive correlation is obtained for half of the eight descriptors (highest value is for *visualScore*).

Apart from the *credibilityset*, user tagging credibility information is provided also for the *devset* and *testset* via the credibility descriptors. In particular, *devset* contains information for 593 users and metadata for 3,348,465 images while *testset* contains 1,752 users and metadata for 8,968,713 images (see also Section 3.4).

3.4 Data format

Each dataset is stored in an individual folder (*devset*, *testset* and *credibilityset*). The following information is provided per dataset:

- **a topic xml file**: containing the list of the locations in the current dataset (e.g., *devset_topics.xml* accounts for *devset*). Each location is delimited by a <topic> </topic> statement and includes the location number and keyword identifier, the GPS coordinates and the url to the Wikipedia webpage of the location;
- **a name correspondence txt file**: containing the list of the location keyword identifiers within the dataset and their corresponding names used for querying data from Flickr (file *poiNameCorrespondences.txt*);
- **an img folder**: containing all the retrieved Flickr images for all the locations in the dataset, stored in individual folders named after each location keyword. Images are named after the Flickr photo ids. All images are stored in JPEG format and have a resolution of around 640 × 480 pixels;
- **an imgwiki folder**: containing Creative Commons location photos from Wikipidia (up to 5 photos per location). Each photo is named after the location keyword and has the owner's name specified in brackets, e.g, "agra_fort(Atmabhola) .jpg" is authored by Atmabhola;
- **a xml folder**: containing all the Flickr metadata stored in individual xml files. Each file is named according to the location keyword and is structured as following:

```
<photos monument="acropolis athens">
<photo date_taken="2013-06-04 02:45:20" description="View of Athens from the entrance of Acropolis" id="9067739127" latitude="37.970805" license="2" longitude="23.721167" nbComments="0" rank="1" tags="athens greece" title="Acropolis - Athens" url_b="http://farm8.static.flickr.com/7362/9067739127_edda2711ca_b.jpg" username="pfischermx" userid="56505984@N06" views="70"/>
... </photos>
```

The *monument* value is the location query name, then, each of the photos is delimited by a <photo /> statement. Each field is explained in Section 3.1;

- **a gt folder**: containing all the dataset ground truth files (details are presented in Section 4). Relevance ground truth is stored in the *rGT* subfolder and diversity ground truth in the *dGT* subfolder. Please note that relevance ground truth is not provided for *credibilityset* in the recorded form, but only through the manual annotation scores;
- **a descvis folder**: containing all the visual descriptors. The *img* subfolder contains the descriptors for the Flickr images as individual csv (comma-separated values) files on a per location and descriptor type basis. Each file is named after the location keyword followed by the descriptor code, e.g., "acropolis_athens CM3x3.csv" refers to the global Color Moments (CM) computed on the 3x3 spatial pyramid for the location acropolis_athens (see Section 3.2.1). Within each file, each photo descriptor is provided on an individual line (ending with carriage return). The first value is the unique Flickr photo id followed by the descriptor values separated by commas. The *imgwiki* subfolder contains the descriptors for the Wikipedia images as individual location csv files using the same convention as for the Flickr images. Different from the previous case, within the files, the first value is now the Wikipedia photo file name;
- **a desctxt folder**: containing all the text descriptors that are provided on a per dataset basis. For each dataset or combination (denoted *all*), the text descriptors are computed on: a per image basis (file id *textTermsPerImage*), a per location basis (file id *textTermsPerPOI*) and a per user basis, respectively (file id *textTermsPerUser*). The descriptors for the combination of *devset* and *testset* are provided in the *testset* folder. In each file, each line represents an entity with its associated terms and their weights: the first token is the id of the entity followed by a list of 4-tuples: "term" TF DF TF-IDF, where "term" is a term which appeared in the text data (see also Section 3.2.2). The term lists provided and described above were generated using Solr 4.7.1[15]. The dataset contains also information for getting your own Solr server running, containing all the data necessary for retrieving images. After installing Solr you need to replace the folder inside the examples folder with the one provided for the dataset. The provided data contains also a data folder that contains all the data provided but in a format ingestible by Solr and that can be used with the *post2solr.sh* script to generate new indexes with different pre-processing steps or similarity functions. We also provide the Bash scripts that we have been used to generate the text descriptors;
- **a desccred folder**: containing all the credibility descriptors computed on a per dataset and per user basis. Each user information is stored in a separate XML file named according to the unique Flickr user id, e.g.,:

```
<metadata user="21953562@N07">
<credibilityDescriptors>
<visualScore>0.791442635512724</visualScore> ...
</credibilityDescriptors>
<photos>
<photo date_taken="2013-08-19 14:11:49" id="9659825826" latitude="42.36115" longitude="-71.03523" tags="boston nhl ..." url_b="http://farm8.static.flickr.com/7408/9659825826_55cb51182d_b.jpg" userid="21953562@N07" views="533" /> ...
```

[14] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

[15] http://lucene.apache.org/

```
</photos>
</metadata>
```

User annotation credibility descriptors are separated by <credibilityDescriptors> </credibilityDescriptors> statements. In addition to these, as for the *credibilityset*, each user is provided with Flickr metadata for a relevant number of images (separated by <photos> </photos> statements and then by <photo /> statements).

4. DATASET ANNOTATION

Images are annotated for their *relevance* and *diversity.* As presented in Section 2, the dataset is built around a tourism use case where a person tries to find more visual information about a place she might visit. Therefore, the annotations were adapted to this scenario. Annotations were performed by experts (trusted annotators) who have advanced knowledge of the location characteristics, mainly learned from Internet sources. To facilitate the process, dedicated visual software tools were employed. During the annotation, the following definitions of relevance and diversity were adopted:

- **relevance**: a photo is considered to be relevant for the location 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 in this scenario;
- **diversity**: a set of photos is considered to be diverse if it depicts different visual characteristics of the target location (see the examples above) with a certain degree of complementarity, i.e., most of the perceived visual information is different from one photo to another.

Definitions were determined and validated in the community via the feedback gathered from 53 respondents of the 2013 and 2014 MediaEval benchmarking surveys[16].

4.1 Task design

Relevance annotation task. For each location, the annotators were provided with one photo at a time. A reference photo of the location (e.g., a Wikipedia photo) was also displayed during the process. Annotators were asked to classify the photos as being relevant (score 1), non-relevant (0) or with "don't know" answer (-1). The definition of relevance was displayed to the annotators during the entire process. The annotation process was not time restricted. Annotators were recommended to consult any additional written or visual information source (e.g., from Internet) in case they were unsure about the annotation.

Diversity annotation task. Diversity is annotated only for the photos that were judged as relevant in the previous relevance step. For each location, annotators were provided with a thumbnail list of all the relevant photos. The first step required annotators to get familiar with the photos by analyzing them for about 5 minutes. Next, annotators were required to re-group the photos in clusters based on visual similarity. The number of clusters was limited to maximum 25. Full size versions of the photos were available by clicking on the photos. The definition of diversity was displayed to the annotators during the entire process. For each of the clusters, annotators provided also some keywords reflecting their judgments in choosing these particular clusters. The diversity annotation was not time restricted.

Table 3: Annotation statistics.

relevance	**devset**	**testset**	**credibilityset**
avg. Kappa	0.85	0.755	0.751
% relev./"don't know"	70/0.03	67.4/0.01	68.6/0.01
diversity	**devset**	**testset**	**credibilityset**
avg.clusters/location	23.17	22.58	-
avg.img./cluster	8.89	8.82	-

4.2 Annotation statistics

The relevance ground truth for *devset* was collected from 3 expert annotators who annotated the entire dataset. The diversity annotation was collected from 2 experts that annotated distinct parts of the data set. In this case, a third annotator acted as a master annotator and reviewed once more the annotations. For *testset*, we employed 11 expert annotators who annotated different parts of the dataset leading in the end to 3 different annotations. The diversity ground truth was collected from 3 expert annotators who annotated distinct parts of the data set. As for *devset*, a master annotator reviewed the annotations. To compute the manual relevance annotation scores of the *credibilityset*, 9 experts annotated different parts of a sub-set of 50,157 images. In the end, there were 3 distinct annotations for each photo. Annotators were both females and males with ages ranging from 23 to 35. Final relevance ground truth was determined after a lenient majority voting scheme (equal numbers of 1 and 0 lead to a 1 decision, -1 are disregarded if not in majority).

For measuring the agreement among pairs of annotators, we computed the Kappa statistics that measure the level of agreement discarding agreement given by chance. Kappa values range from 1 to -1, where values from 0 to 1 indicate agreement above chance, values equal to 0 indicate equal to chance, and values from 0 to -1 indicate agreement worse than chance. In general, Kappa values above 0.6 are considered adequate and above 0.8 are considered almost perfect [8]. The annotation statistics are summarized in Table 3. We achieve a good agreement between annotators, average Kappa being above 0.7. In the same time, the amount of "do not know" labels after majority voting is negligible, e.g., less than 0.01% for *testset.* For the diversity annotation, the average number of clusters per location and the average number of images per cluster are consistent for both *devset* and *testset* being situated around 23 and 9, respectively,

4.3 Annotation data format

Ground truth is provided on a per dataset and location basis (see the folder structure in Section 3.4). We provide individual txt files for each location. Files are named according to the location keyword identifier followed by the ground truth code: *rGT* for relevance, *dGT* for diversity and *dclusterGT* for the cluster tags, e.g., "atomium dGT.txt" refers to the diversity ground truth for the location Atomium. For the *rGT* files, each file contains photo ground truth on individual lines. The first value is the unique photo id from Flickr followed by the ground truth value (1, 0 or -1) sep-

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

Table 4: Best performance/baseline at MediaEval 2014.

team & approach	CR@20	P@20	F1@20
PRa-MM, hierarch. cluster. with re-ranking [10]	46.92%	85.12%	59.71%
Flickr initial results	34.27%	80.65%	46.99%

arated by comma. The *dGT* files are structured similarly to *rGT* but having after the comma the cluster id number to which the photo was assigned (a number from 1 to 25). The *dclusterGT* files, complement the *dGT* by providing the cluster tag information. Each line contains the cluster id followed by the provided user tag separated by a comma.

5. MEDIAEVAL 2014 VALIDATION

The proposed dataset was validated during the 2014 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative for Multimedia Evaluation[16]. The task challenged participants to design either machine, human or hybrid approaches for refining Flickr results in view of providing a ranked list of up to 50 photos that are considered to be both a relevant and a diverse representation of the queries (for more details about the task see [9]). In total, 20 teams from 15 countries registered to the task and 14 submitted a total of 54 runs. The tested approaches included the use of clustering, re-ranking, optimization-based and relevance feedback including machine-human. Various combination of information sources have been explored (visual - 17 runs, text - 13, credibility information - 7, multimodal - 16, human - 1). System performance is assessed in terms of 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), precision at X (P@X — measures the number of relevant photos among the top X results) and their harmonic mean, i.e., F1-measure@X (X∈{5, 10, 20, 30, 40, 50}).

To provide a baseline for this dataset, Table 4 presents the best overall average result for the official metrics F1-measure@20 (all task results are available here[16]). Highest performance for a cutoff at 20 images was achieved when considering also the user credibility information with a hierarchical clustering and re-ranking approach [10]. Compared to the initial Flickr ranking results, the improvement of diversity is over 12%. This strengthens our hypothesis [11] that user credibility-based information is a promising approach for improving the diversification of search results in the social media context.

The following information will help reproducing the exact evaluation conditions of the task. Participant runs were processed in the form of trec topic files[17], each line containing the following information separated by whitespaces: *qid iter docno rank sim run_id*, where *qid* is the unique query id (see the topic files, Section 3.4), *iter* gets disregarded (e.g., 0), *docno* is the unique Flickr photo id, *rank* is the new photo rank in the refined list (an integer ranging from 0 — highest rank — up to 49), *sim* is a similarity score and *run_id* is the run label. A sample run file is provided in the root folder of the dataset ("me14div_example_run.txt") together with the official scoring tool, "div_eval.jar". To run the script, use the following syntax (make sure you have Java installed on your machine): *java* **-jar** *div_eval.jar* **-r** <runfilepath> **-rgt** <rGT_path> **-dgt** <dGT _path> **-t** <topic_filepath> **-o** <output_dir> [optional: **-f** <filename>]; where <runfilepath> is the file path of the run file, <rGT_path> is the path to the relevance ground truth, <dGT_path> is the path to the diversity ground truth, <topic_filepath> is the file path to the topic xml file.

6. CONCLUSIONS

We introduced a new dataset and its evaluation tools, Div150Cred, that contains image, metadata, descriptor (visual and text including user tagging credibility estimations) and ground truth information for 300 landmark locations. The dataset is designed for benchmarking social image search results diversification techniques. It was validated during the 2014 Retrieving Diverse Social Images Task at the MediaEval Benchmarking on more than 50 runs. Future extensions of this data will mainly target diversifying the provided use case while exploiting further user context.

7. ACKNOWLEDGMENTS

This data set was supported by MUCKE FP7 CHIST-ERA, CUbRIK FP7 n287704 and PROMISE FP7 n258191. We also acknowledge the MediaEval Benchmarking Initiative for Multimedia Evaluation[16] for their support.

8. REFERENCES

[1] 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.

[2] 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.

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

[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] 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.

[6] D. Harman, "How effective is suffixing", JASIS, 42, 1991.

[7] A. Popescu, "CEA LIST's Participation at MediaEval 2013 Placing Task", Working Notes of MediaEval 2013, CEUR-WS, Vol. 1043, Barcelona, Spain.

[8] J.J. Randolph, "Free-Marginal Multirater Kappa (multirater κfree): an Alternative to Fleiss Fixed-Marginal Multirater Kappa", Joensuu Learning and Instruction Symposium, 2005.

[9] B. Ionescu, A. Popescu, M. Lupu, A.L. Gînscă, H. Müller, "Retrieving Diverse Social Images at MediaEval 2014: Challenge, Dataset and Evaluation", Working Notes of MediaEval, Barcelona, Spain, CEUR-WS, Vol. 1263, 2014.

[10] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, F. De Natale, "Retrieval of Diverse Images by Pre-filtering and Hierarchical Clustering", Working Notes of MediaEval 2014, Barcelona, Spain, CEUR-WS, Vol. 1263, 2014.

[11] A.L. Gînscă, A. Popescu, B. Ionescu, A. Armagan, I. Kanellos, "Toward Estimating User Tagging Credibility for Social Image Retrieval", ACM Multimedia, Orlando, Florida, USA, 2014.

[17] http://trec.nist.gov/trec_eval/