Div150Multi: A Social Image Retrieval Result Diversification Dataset with Multi-topic Queries

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

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

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

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

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

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

ABSTRACT

In this paper we introduce a new dataset, Div150Multi, 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 trusted annotators. The data consists of around 300 complex queries represented via 86,769 Flickr photos, around 27M photo links for around 6,000 users, metadata, Wikipedia pages and content descriptors for text and visual modalities, including state of the art deep features. To facilitate distribution, only Creative Commons content allowing redistribution was included in the dataset. The proposed dataset was validated during the 2015 Retrieving Diverse Social Images Task at the MediaEval Benchmarking.

CCS Concepts

•Information systems → Information retrieval diversity; Test collections;

Keywords

social photo retrieval, search result diversification, multi-topic queries, user tagging credibility, MediaEval benchmark.

1. INTRODUCTION

The huge amount of social multimedia data now generated by users require adequate multimedia retrieval capabilities. While until recently research focused primarily on improving search result *relevance*, an effective retrieval system should also cover *diverse* aspects of a topic. The motivation behind this is the fact that most queries have a large spectrum of visual representations, e.g., different instances of the same concept, multiple contexts or temporal and placement variations. Increasing result diversification tends to improve the system's overall performance, e.g., by responding to the needs of different users and tackling queries with unclear information needs. Different approaches are being proposed to meet these novel challenges in image retrieval, examples include the integration of result categorization in Google Image Search[1] or Flickr's image search by *interestingness* feature[2]. However, despite current progress, developing image search diversification methods in a social setting and, more notably, assessing performance of such systems are still open research questions.

In this paper, we introduce a benchmarking dataset designed to support this emerging area of social image retrieval focusing on *diversification*. It 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. Its richness of social data (including user tagging credibility estimation) could be exploited in adjacent fields as well, e.g., the emerging field of multimedia Web data quality.

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 Div150Multi dataset. Section 4 presents the ground truth creation protocol. Section 5 introduces an evaluation framework and several baselines. Section 6 concludes the paper.

2. RELATED WORK

Most of the retrieval result diversification efforts have been carried out naturally in the setting of Web and text search [1, 22, 21]. Some public search result diversification evaluation tasks, such as the TREC Web Track Diversity task [4] and the NTCIR Intent/IMine task [10], have been organized to evaluate diversification approaches via public collections.

There are however, notable approaches that focus on multimedia and propose frameworks for evaluating diversity in the context of image retrieval [13, 3, 19, 6, 20]. A drawback is that the evaluation of these approaches tends to be carried out mostly on very particular and closed datasets,

*work supported by the CHIST-ERA FP7 MUCKE Multimodal User Credibility and Knowledge Extraction project.

MMSys'16, May 10-13, 2016, Klagenfurt, Austria

DOI: http://dx.doi.org/10.1145/2910017.2910620

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

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

which limits the reproducibility and comparison of the results between methods. For instance, Hoque et al [6] use annotators to rate the relevance of the top 60 search results from Google Image Search for 12 topics. For each topic, diversity is achieved by testing several levels of concept-based query expansion. The Flickr dataset proposed in [20] is collected based on a diverse set of popular tags and consists of 104,000 images and 83,999 unique tags. Each image is labeled with three relevance levels. Although the authors do not provide ground truth annotations for diversity, they propose a novel ranking evaluation measure (Average Diverse Precision Based) and diversity is obtained by optimizing this metric. Rudinac et al [13] introduce a collection of Flickr 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. Closer related to our data is the collection introduced by van Leuken et al [19], which uses a dataset 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. Apart from these user introduced data, there are however very few attempts to constitute a standardized evaluation framework for image search diversification. To the best of our knowledge, the most notable are the ImageCLEF benchmarking with the Photo Retrieval task [11] and the MediaEval benchmarking with the Retrieving Diverse Social Images task [8, 9].

The proposed Div150Multi dataset falls in line with the latest trends in multimedia evaluation, namely the Yahoo! Flickr Creative Commons dataset (YFCC100M) [18], which consists of 100 million Flickr user-uploaded images and videos along with their corresponding metadata. Although this initiative reinforces the use of social Web multimedia data for research purposes, it does not provide manual relevance and diversity assessments. Div150Multi uses similar Flickr real-world data but accompanied by expert annotations. It builds upon the groundwork laid out by the Div150Cred [8] and Div400 [9] datasets and brings the following new contributions to the state of the art: (i) it introduces a novel challenge by addressing the issue of *multi-topic queries* in the diversification context, i.e., queries related to events and states associated with locations; (ii) it provides a strong baseline by building on top of the Flickr's state-of-the-art relevance system thus pushing forward the advances in the field; (iii) it proposes a focused real-world usage scenario, i.e., tourism, which disambiguates the diversification need; (iv) it provides a strong focus on the the social dimension of the diversification problem by directly incorporating large and diverse user data including information about user tagging credibility; (v) it addresses different retrieval communities by incorporating, pre-computed, visual, social and text models and descriptors, including state of the art deep features.

3. DATASET DESCRIPTION

To disambiguate the diversification need, we have selected as use case for the proposed data, image searches related to tourist landmarks and, most notably, location specific events, location aspects or general activities.

The data consists of a development set (*devset*) containing 153 location queries (45,375 Flickr photos — the complete Div150Cred dataset [8]), a user annotation credibility set (*credibilityset*) containing information for approximately 300 locations and 685 users 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). One-concept location related queries refer typically to natural or man-made landmarks (e.g., sites, museums, monuments, buildings, caves), while multi-concept location related queries refer to events and locations (e.g., "Oktoberfest in Munich", "Bucharest in winter"). Location and event queries were selected based on the number of Creative Commons photos available on Flickr.

The dataset[3] consists of redistributable Creative Commons[4] Flickr and Wikipedia location data. For each single-topic query, the following information is provided: *query keyword* (unique textual identifier in the dataset), *query number* (unique numeric identifier), *GPS coordinates* (latitude and longitude in degrees) retrieved from GeoHack[5] 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 and models. The same data are available for the multi-topic queries, with the exception of GPS coordinates and Wikipedia photos (which were not always available due to the nature of these queries).

3.1 Flickr data collection method

Apart from Wikipedia data, query information was collected from Flickr using the Flickr API[6] (under Python) and the *flickr.photos.search* function. Information was retrieved using the query text formulation and ranked with Flickr's default relevance algorithm. Therefore, the dataset is built on top of the current state-of-the-art retrieval technology.

For each query, 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[6]. 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[7], Creative Common *license type*, number of *posted comments* and the photo's *rank* within the Flickr results (a generated number from 1 to 300).

3.2 Visual and text descriptors

To facilitate exploitation of these data by various communities, data are accompanied by pre-computed descriptors.

3.2.1 General purpose visual descriptors

For each photo, we provide several general purpose visual descriptors. These are the same descriptors we provided with the previous editions of this dataset [9], namely: Color Naming Histogram (code CN - 11 values), Histogram of Ori-

[3]to download the dataset see http://imag.pub.ro/~bionescu/index_files/Page13288.htm

[4]http://creativecommons.org/

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

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

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

Table 1: Basic statistics for development (*devset*), evaluation (*testset*) and user tagging credibility (*credibilityset*) data.

		#topics	*#images*	*min-average-max #images per query*
devset	single-topic	153	45,375	281 - 297 - 300
testset	single-topic	69	20,700	300 - 300 - 300
	multi-topic	70	20,694	176 - 296 - 300
		#topics	*#image URLs*	*#users*	*average #images per user*
credibilityset	single-topic	300	3,651,303	685	5,330

ented Gradients (code HOG - 81 values), Color Moments in the HSV Color Space (code CM - 9 values), Local Binary Patterns (code LBP - 16 values), Color Structure Descriptor (code CSD - 64 values), statistics on gray level Run Length Matrix (code GLRLM - 44 values) and their pyramid representations (code GLRLM 3x3 - 396 values).

3.2.2 CNN based descriptors

New for this dataset, we provide state of the art convolutional neural network based descriptors that were specifically tuned for the diversification task, namely [15]:
• *CNN generic* (code cnn_gen - 4,096 values): descriptor based on the reference convolutional (CNN) neural network model provided along with the Caffe framework. This model is learned 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* (code cnn_ad - 4,096 values): descriptor 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 of which the images were automatically collected from the Web. Similar to CNN generic, the descriptors are extracted from the last fully connected layer of the network (i.e., *fc7*).

3.2.3 Text models

We provide standard term frequency document frequency representations of the social information, i.e., 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. These are the same descriptors as in Div150Cred [8].

3.2.4 User annotation 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 [5]. 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.

Apart from the descriptors introduced with the previous edition of the dataset [8], i.e., visualScore, faceProportion, tagSpecificity, locationSimilarity, photoCount, uniqueTags, uploadFrequency, and bulkProportion, the following new descriptors were computed by visual or textual content mining:
• *meanPhotoViews*: the mean value of the number of times a user's image has been seen by other members of the community (not normalized);
• *meanTitleWordCounts*: the mean value of the number of words found in the titles associated with users' photos (not normalized);
• *meanTagsPerPhoto*: the mean value of the number of tags users add to their images (not normalized);
• *meanTagRank*: the 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 (not normalized). We eliminate bulk tagging and obtain a set of 20,737,794 unique tag lists out of which we extract the tag frequency statistics. To extract this descriptor we take into consideration only the tags that appear in the top 100,000 most frequent tags;
• *meanImageTagClarity*: this descriptor is based on an adaptation of the Image Tag Clarity score described in [17]. The clarity score for a tag is the KL-divergence between the tag language model and the collection language model. We use the same collection of 20,737,794 unique tag lists to extract the language models. The collection language model is estimated by the relative tag frequency in the entire collection. For a tag, its clarity score is an indicator on the diversity of contexts the tag is used. A low clarity score suggests a tag is generally used together with the same tags. *meanImageTagClarity* is the mean value of the clarity score of a user's tags (not normalized).

3.3 Dataset basic statistics

The data are structured into a development set (*devset*) containing Flickr and Wikipedia information (as described in the previous sections) for 153 single-topic queries. Its objective is to serve for the design and training of potential approaches; and a test set (*testset*) that contains 69 single-topic and 70 multi-topic queries and is intended for the actual benchmarking and validation of the methods. In total, *devset* and *testset* account for 86,769 images. Some basic image statistics are presented in Table 1.

In addition, we provide a specially designed dataset (*credibilityset*) that addresses the estimation of user tagging credibility (actualization to the new data of the same dataset from [8]). This dataset is intended for training and designing user tagging credibility related descriptors and contains information for around 685 users (different than the ones in *devset* and *testset*). Each user is assigned a manual credibility score which is determined as the average relevance score of all the user's photos (relevance annotations are determined as presented in Section 4).

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 more than 2,000 users and metadata for ca. 12 million images while *testset* contains more than 4.000 users and metadata for 15 million images (see also Section 3.4).

3.4 Data format

Each dataset is stored in an individual folder (*devset*, *testset* with a folder for each type of query, i.e., *one-topic* and *multi-topic*; and *credibilityset*) providing the following information:

• **a topic xml file**: containing the list of the queries in the current dataset (e.g., *devset_topics.xml* accounts for *devset*). Each query is delimited by a <topic> </topic> statement and includes the query number and keyword identifier, the GPS coordinates and the url to the Wikipedia webpage of the query (if available);
• **a name correspondence txt file**: containing the list of the query keyword identifiers within the dataset and their corresponding text used for querying data from Flickr (file *poiNameCorrespondences.txt*);
• **an img folder**: containing all the retrieved Flickr images for all the queries in the dataset, stored in individual folders named after each query keyword. Images are named after the Flickr photo ids. All images are stored in JPEG format and have a resolution of 640 × 480 pixels;
• **an imgwiki folder**: containing Creative Commons photos from Wikipidia (up to 5 photos per query), only for the single-topic queries. 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 query 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 query name, then, each of the photos is delimited by a <photo /> statement. Each field was 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 general purpose visual descriptors (Section 3.2.1). The *img* subfolder contains the descriptors for the Flickr images as individual csv (comma-separated values) files on a per query and descriptor type basis. Each file is named after the query 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. 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 descCNN folder**: containing all the CNN based descriptors (Section 3.2.2). Using the same format as for the general purpose visual descriptors above, the *img* subfolder contains the descriptors for the Flickr images and the *imgwiki* subfolder for the Wikipedia images;
• **a desctxt folder**: containing all the text descriptors (Section 3.2.3) that are provided on a per dataset and subdataset basis. For each dataset, the text descriptors are computed on: per image basis (file id *textTermsPerImage*), per query basis (file id *textTermsPerPOI*) and per user basis, respectively (file id *textTermsPerUser*). 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 that appeared in the text data. The term lists provided and described above were generated using Solr 4.10.3[8]. The dataset contains also information for getting a personal Solr server running, containing all the data necessary for retrieving images. After installing Solr the folder inside the examples folder needs to be replaces with 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 (Section 3.2.4) 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*" /> ...
</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 3, the dataset is built around a tourism use case, therefore, the annotations were adapted to this scenario. Annotations were performed by experts (trusted annotators) who have advanced knowledge of the query 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[9]:
• **relevance**: a photo is considered to be relevant for the query if it is a common photo representation of all query concepts at once. This includes sub-locations (e.g., subsuming indoor/outdoor, close up, far back view), temporal information (e.g., historical shots, times of day, typical events), typical actors/objects (e.g., people who frequent the location, animals, vehicles), genesis information (e.g., images

[8]http://lucene.apache.org/
[9]validation via feedback gathered from more than 80 respondents of the 2013-2015 MediaEval benchmarking surveys, http://www.multimediaeval.org/.

showing how something got the way it is), and image style information (e.g., drawings, 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, genesis information, style information, etc with a certain degree of complementarity, i.e., most of the perceived visual information is different from one photo to another.

4.1 Task design

Relevance annotation task. For each query, the annotators were provided with one photo at a time. A reference photo of the query (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 the 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 query, 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 a maximum of 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.

4.2 Annotation statistics

For *relevance* and *devset*, 11 annotators were involved, for *credibilityset* 9 and for *testset* single-topic 7 and multi-topic 5. Each annotator annotated different parts of the data leading in the end to 3 different annotations for each photo. The 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 *diversity*, only the photos that were judged as relevant in the previous step were considered. *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.

For measuring the agreement among pairs of annotators, we computed the Kappa statistics that measure the level of agreement discarding agreement 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 [12]. The annotation statistics are summarized in Table 2. 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 20 and 10, respectively. One can notice however that multi-topic queries tend to have less diversity than the single-topic ones.

Table 2: Annotation statistics.

relevance	**devset**	**test-single**	**test-multi**
avg. Kappa	0.773	0.797	0.693
% relevant	0.679	0.63	0.69
diversity	**devset**	**test-single**	**test-multi**
avg. clusters/query	22.9	20.9	17.2
avg. img./query	8.9	9	12.6

4.3 Annotation data format

Ground truth is provided on a per dataset/sub-dataset and query basis (see the folder structure in Section 3.4). We provide individual txt files for each query. Files are named according to the query 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 query 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) separated by a 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 2015 VALIDATION

The proposed dataset was validated during the 2015 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative for Multimedia Evaluation[9]. 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 [7]).

In total, 24 teams from 18 countries registered to the task and 14 submitted a total of 59 runs. The tested approaches included the use of classification/clustering and fusion, reranking, optimization-based and relevance feedback including machine-human approaches. Various combination of information sources have been explored (visual — 15 runs, text — 14, credibility information — 7, multimodal — 22, human based — 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 3 presents the best *testset* average results for the official metric, i.e.,

Table 3: Best performances/baseline at MediaEval 2015.

test dataset	team	CR@20	P@20	F1@20
single-topic	USEMP [16]	50.44%	83.33%	61.77%
	Flickr	36.81%	68.77%	46.76%
multi-topic	PRaMM [2]	47.53%	76.07%	56.7%
	Flickr	36.87%	71.21%	46.67%
all	TUW [14]	49.63%	73.09%	57.27%
	Flickr	36.84%	70%	46.72%

F1@20, together with the Flickr initial results (for more information on the submitted methods and runs, see[10]). We notice here that Flickr search results have almost identical scores for the F1@20 metric when comparing single-topic and multi-topic queries, which shows the consistency of the system, regardless of the formulation of the query. However, as expected, developing further diversification for multi-topic queries is more challenging. We can see a 5.07% drop in performance when passing from the single-topic to the multi-topic setting. We also observe a consistent improvement for the best run over the Flick baseline regardless the test collection split: 32.1% relative improvement for *single-topic*, 21.49% relative improvement for *multi-topic* and 22.58% relative improvement for the complete *testset*. Although the top ranked approach was different for each testset setting, all of them used multi-modal approaches, i.e., USEMP run3 [16] — use of supervised Maximal Marginal Relevance with CNNs, Bag-of-Words and VLAD features; PRaMM run5 [2] — use of pre-filtering of outliers, clustering with BIRCH algorithm and summarization via agglomerative hierarchical clustering. Features used are TF-IDF representations and dense SIFT and HoGs; TUW run3 [14] — use of word embeddings with Word2Vec and learning a best clustering algorithm using the development data. Visual descriptors include histograms and CNNs. To help reproducing the exact evaluation conditions of the task, the dataset is provided also with the official evaluation tool ("div_eval.jar" — developed under Java), a sample run file ("Flickr_initial_2015_testset_all.txt") and a readme file describing all the task details and the run format ("Div150Multi_readme.txt").

6. CONCLUSIONS

We introduced the Div150Multi dataset, a rich dataset containing real-world Flickr image search results for over 220 landmark related queries and 70 multi-topic queries related to events and states associated with locations, together with ground truth annotations. The dataset is specifically designed for benchmarking social image search results diversification techniques and was successfully validated during the 2015 Retrieving Diverse Social Images Task at the MediaEval Benchmarking. The dataset is publicly available.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14. ACM, 2009.

[2] D.-T. Dang-Nguyen, G. Boato, F. G. Natale, L. Piras, G. Giacinco, F. Tuveri, and M. Angioni. Multimodal-based diversified summarization in social image retrieval. 2015.

[3] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. G. De Natale. A hybrid approach for retrieving diverse social images of landmarks. In *IEEE ICME*, pages 1–6, 2015.

[10] http://ceur-ws.org/Vol-1436/

[4] A. Dean-Hall, C. L. Clarke, J. Kamps, P. Thomas, N. Simone, and E. Voorhees. Overview of the trec 2013 contextual suggestion track. Technical report, DTIC Document, 2013.

[5] A. L. Ginsca, A. Popescu, B. Ionescu, A. Armagan, and I. Kanellos. Toward an estimation of user tagging credibility for social image retrieval. In *ACMMM*, pages 1021–1024. ACM, 2014.

[6] E. Hoque, O. Hoeber, and M. Gong. Balancing the trade-offs between diversity and precision for web image search using concept-based query expansion. *Journal of Emerging Technologies in Web Intelligence*, 4(1):26–34, 2012.

[7] B. Ionescu, A. L. Gînsca, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *MediaEval, vol. 1436*, pages 14–15, 2015.

[8] B. Ionescu, A. Popescu, M. Lupu, A.-L. Gînsca, B. Boteanu, and H. Müller. Div150cred: A social image retrieval result diversification with user tagging credibility dataset. In *MMSys*, pages 207–212, 2015.

[9] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, and B. Loni. Div400: a social image retrieval result diversification dataset. In *MMSys*, pages 29–34. ACM, 2014.

[10] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. P. Kato, H. Ohshima, and K. Zhou. Overview of the ntcir-11 imine task. In *NTCIR*, 2014.

[11] M. L. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: overview of the imageclefphoto task 2009. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 45–59. Springer, 2009.

[12] J. J. Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005.

[13] S. Rudinac, A. Hanjalic, and M. Larson. Generating visual summaries of geographic areas using community-contributed images. *Multimedia, IEEE Transactions on*, 15(4):921–932, 2013.

[14] S. Sabetghadam, J. Palotti, N. Rekabsaz, M. Lupu, and A. Hanbury. Tuw@ mediaeval 2015 retrieving diverse social images task.

[15] 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 *ICMR*, pages 323–330. ACM, 2015.

[16] E. Spyromitros-Xioufis, A. Popescu, S. Papadopoulos, and I. Kompatsiaris. Usemp: Finding diverse images at mediaeval 2015. 2015.

[17] A. Sun and S. S. Bhowmick. Image tag clarity: in search of visual-representative tags for social images. In *Proceedings of the first SIGMM workshop on Social media*, pages 19–26. ACM, 2009.

[18] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[19] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *WWW*, pages 341–350. ACM, 2009.

[20] K. Yang, M. Wang, X.-S. Hua, and H.-J. Zhang. Tag-based social image search: Toward relevant and diverse results. In *Social Media Modeling and Computing*, pages 25–45. Springer, 2011.

[21] H.-T. Yu and F. Ren. Search result diversification via filling up multiple knapsacks. In *CIKM*, pages 609–618. ACM, 2014.

[22] Y. Zhu, Y. Lan, J. Guo, X. Cheng, and S. Niu. Learning for search result diversification. In *SIGIR*, pages 293–302. ACM, 2014.