SubDiv17: A Dataset for Investigating Subjectivity in the Visual Diversification of Image Search Results

Maia Rohm
TU Wien, Austria
maia.rohm@tuwien.ac.at

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

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

Rodrygo L.T. Santos
Universidade Federal de Minas Gerais, Brazil
rodrygo@dcc.ufmg.br

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

ABSTRACT

In this paper, we present a new dataset that facilitates the comparison of approaches aiming at the diversification of image search results. The dataset was explicitly designed for general-purpose, multi-topic queries and provides multiple ground truth annotations to allow for the exploration of the subjectivity aspect in the general task of diversification. The dataset provides images and their metadata retrieved from Flickr for around 200 complex queries. Additionally, to encourage experimentations (and cooperations) from different communities such as information and multimedia retrieval, a broad range of pre-computed descriptors is provided. The proposed dataset was successfully validated during the MediaEval 2017 Retrieving Diverse Social Images task using 29 submitted runs.

CCS CONCEPTS

• **Information systems → Test collections**; *Information retrieval diversity*;

KEYWORDS

Benchmark dataset, search result diversification, image retrieval, annotation subjectivity, MediaEval, Flickr

ACM Reference Format:
Maia Rohm, Bogdan Ionescu, Alexandru Lucian Gînscă, Rodrygo L.T. Santos, and Henning Müller. 2018. SubDiv17: A Dataset for Investigating Subjectivity in the Visual Diversification of Image Search Results. In *MMSys'18: 9th ACM Multimedia Systems Conference, June 12–15, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3204949.3208122

1 INTRODUCTION

The task of search result diversification aims at the creation of a broad representation of a data set retrieved in result to a given user query. As such, the task is a way to address multi-faceted or ambiguous user information needs. Previously, search result diversification was predominantly explored in the context of text information retrieval [30, 36]. Common applications range from web search [24, 33, 34, 37] and personalization [18] to microblogging [17, 19, 26], news [9], and product review summarization [20]. In general, the specific subject to diversification can take very different shapes such as potential subtopics, opinion types, genre, etc. Currently, user intent is commonly defined as different query aspects [16]. However, in many real-world applications, the underlying query aspects (subtopics) are not known in advance but defined by the data itself.

With the increasing number of publicly available media, more and more research focuses on the diversification of image data. Existing, publicly available datasets commonly target a very tailored application scenario, e.g. tourist-oriented [13–15]. In contrast, in this paper we present a dataset addressing several crucial aspects of the image search result diversification task. First, the dataset is build on top of general-purpose, multi-topic queries. While this aspect increases the real-world applicability of potential approaches evaluated on the provided data, it notably increases the challenge of a) assessing relevance in general (expert knowledge of the subject of the query is often required) and b) finding media relevant to a complex query corresponding to a combination of multiple terms. Second, we target the common use case where the underlying user intent is not known. Therefore, the diversification task is only data-driven. Eventually, different people tend to have varying views on a given dataset and to consider different aspects and data characteristics when assessing items as being similar. On the one side, this subjectivity plays a crucial role in the annotation process (or the creation of the ground truth) of the provided dataset. On the other side, it also reflects potential differences in the information needs of the end user. Therefore, we provide multiple annotations for image search result diversification to allow for an in-depth analysis of subjective assessments in the context of image analysis.

This paper is organized as follows. Section 2 outlines the data collection procedure. Section 3 provides in-depth insights into the annotation process and the resulting ground truth. Section 4 gives a brief overview of pre-computed descriptors provided with the dataset. Section 5 outlines the dataset validation in the context of the MediaEval Benchmark 2017. Section 6 concludes the paper.

2 DATA COLLECTION

The collected dataset is built around the use case of a general ad-hoc image retrieval system that provides the user with visually diversified representations of query results. To ensure a broad query

MMSys'18, June 12–15, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5192-8/18/06…$15.00
https://doi.org/10.1145/3204949.3208122

Table 1: General data statistics.

set	#queries	origin	#unique users	#unique images
devset	110	User Study	6,955	32,340
testset	84	Google Trends	111,095	24,986

coverage, we collected queries from both a user study and a thorough analysis of the worldwide Google Trends[1] for image search within the last five years (2012–2017). In order to increase the real-world applicability, a potential query of interest was accepted if it represented a general, multi-topic term (e.g. multi-topic: *accordion player* vs. single topic: *accordion* or *player*) and if it was not related to a single location (e.g. *hanging bridge* vs. *Golden Gate Bridge*).

For each of the collected queries, we acquired up to 300 images and their corresponding metadata from Flickr using the Flickr API[2]. The raw data were retrieved using the query text formulation and ranked with Flickr's default relevance algorithm providing a current state-of-the-art technology as a baseline for the dataset. All retrieved images have Creative Commons[3] licenses of type 1 to 7 permitting their redistribution. The metadata for each image include 1) *general photo information*, such as the photo ID, the date it was taken, its rank within the Flickr results, and the license type; 2) *user information* including user ID and user name; 3) *user-provided image descriptions*, such as title, description, and tags; and 4) *social information* providing the number of photo views and the number of posted comments. We divided the data into development (*devset*) and test sets (*testset*) according to the origin of the queries. Table 1 summarizes the general information on the collected data.

3 DATA ANNOTATION

The acquired images were annotated with respect to both their *relevance* and *diversity* regarding the underlying query. The annotation process was carried out by experienced (trusted) annotators. The following definitions of relevance and diversity were adopted:

Relevance: an image is considered to be relevant to the query if it is a common visual representation of the query (all query terms at once). Images with low quality (e.g. severely blurred, out of focus) are not considered relevant in this scenario;

Diversity: a set of images is considered to be diverse if it depicts different visual characteristics of the query terms with a certain degree of complementarity, i.e. most of the perceived visual information is different from one image to another.

The definitions were determined and validated in the multimedia and information retrieval communities via feedback gathered from more than 200 respondents to the MediaEval[4] benchmarking surveys 2013–2017.

3.1 Relevance Annotation

The annotation of image relevance was carried out by 17 annotators covering distinct parts of the dataset. Annotators were asked to label each image (one at a time) as being relevant to the underlying query (score 1), non-relevant (0), or with "don't know" (−1). The definition of relevance was available to the annotators during the entire process. The annotation process was not time restricted and annotators were recommended to consult any external information source in case of uncertainty about the relevance of the image or about the interpretation of the query. Additionally, a master annotator reviewed the annotations focusing on the elimination of ambiguity (i.e. images annotated with −1). The final relevance score was determined using a simple majority voting scheme.

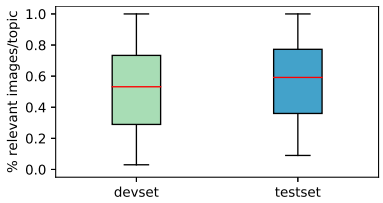
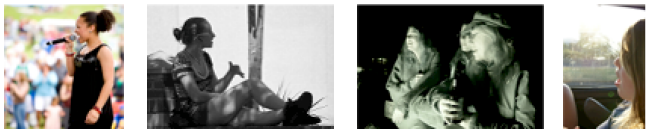


Figure 1: Distribution of the ratio of finally relevant images.

(a) Query: *bus stop*



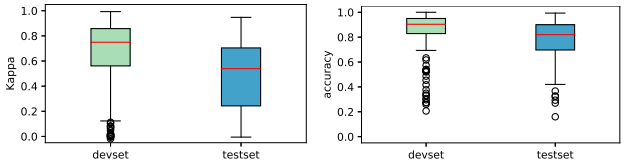
(b) Query: *girl singing*

Figure 2: Challenges in relevance estimation.

Figure 1 shows the distribution of the percentages of relevant images per topic using a boxplot visualization. The distributions for the development and test data are highly comparable and show no significant difference (Mann-Whitney-U test [10], $p = 0.2091$). On average, 53% of the images of the topics in the development set and 57% of those in the test set are considered as relevant to the underlying query. This indicates that state-of-the-art approaches for relevance estimation – as the one currently employed by Flickr – still have a great potential for improvement. In general, there seems to be no obvious pattern for the performance of the underlying relevance estimation algorithm by Flickr. For example, the query *citroen vintage car* results in 95% relevant images returned by Flickr while a comparable query *three wheeled car* has only 13%. Similarly, the query *tree with flowers* has 83% relevant images and the query *tree without leaves* only 14%. These observations suggest that a reliable estimation of relevance is a challenging task for multi-topic queries.

In many cases, the relevance of an image is not unambiguously determinable. Figure 2 shows examples for images retrieved from Flickr for two queries: *bus stop* and *girl singing*. Due to the clearly visible bus sign and overall settings, we can label the first two images retrieved for the query *bus stop* as relevant (with a high degree of probability). However, the third and the fourth images require for highly specialized expert knowledge – either of the locations where the photos were taken, or from the language the sign in the third

[1] http://trends.google.com/
[2] http://www.flickr.com/services/api/
[3] https://creativecommons.org/
[4] http://www.multimediaeval.org/



(a) Kappa scores (b) Accuracy

Figure 3: Relevance annotation agreements between any two annotators for each query.

image is written in. The second example in Figure 2 shows some images retrieved for the query *girl singing*. Again, the relevance of the images is difficult to assess using the visual information only and one might argue that the depicted girls are simply speaking. Finally, the relevance of images capturing an activity (e.g. *playing*, *dancing*, *reading*) is commonly challenging to determine due to the nature of a photo capturing a single moment in time. These difficulties are partially indicated by Cohen's Kappa [8] statistics visualized in Figure 3(a). In general, the Kappa statistic measures the agreement between any two annotators on the same topic discarding agreement by chance. The scores range between -1 and 1 with values below 0 indicating disagreement worse than chance. Despite the significant difference in the score distributions of the development and test sets (Mann-Whitney-U test [10], $p \ll 0.0001$), on average, the Kappa scores show a good agreement between annotators (mean Kappa score of 0.68 for the development and 0.48 for the test set) with the test data having a broader range of Kappa scores and, thus, indicating partially controversial relevance annotation opinions. Despite its popularity, Cohen's Kappa coefficient was proven flawed on numerous occasions and was commonly criticised for its excessively high chance agreement correction [12, 27]. Therefore, similar to the work by Nowak and Rüger [25], we additionally investigate the accuracy between any two sets of annotations. For each query, the accuracy corresponds to the ratio of identically labeled images with respect to the total number of images. The results are visualized in Figure 3(b). Overall, both the annotations on the development and test sets achieve a high accuracy: in 75% of all pairwise comparisons, the accuracy exceeds 83% and 70%, achieving an average accuracy of 86% and 78% for the development and test sets respectively.

3.2 Diversity Annotation

Diversity annotation was performed only for images which were considered to be relevant to the underlying query. In total, 16 annotators were involved and processed distinct parts of the data. For each query, the annotation process passed two steps. First, the annotators were asked to familiarize themselves with the (relevant) images by analyzing them for about 5 minutes. Next, annotators were asked to group the images in clusters based on their visual similarity. For each of the clusters, the annotators provided keywords reflecting their intuition in building the particular clusters. Similar to the relevance annotation, the definition of diversity was available to the annotators during the entire process. Again, the annotation process was not time-restricted.

The assignment of images to groups is a nontrivial task. Different persons tend to focus on different visual aspects ranging from low-level characteristics such as colors, shape and general scene settings

(a) Pot type (b) Flower color

Figure 4: Different clustering possibilities for example images retrieved for the query *flower pot*.

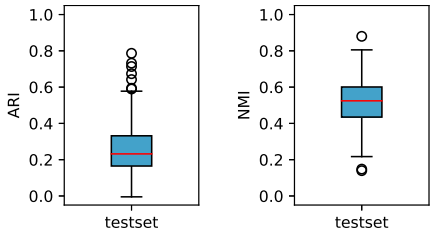


Figure 5: Diversity annotation agreements between any two annotators on the test data.

to higher level characteristics commonly originating from expert knowledge such as specimen type. Figure 4 shows two possible groupings of example images retrieved for the query *flower pot*. Both groupings are valid and legitimate: the first one using the pot type as grouping criterium and the second one the color of the flowers. A further possible grouping might focus on the flower type itself. As result, the annotation process seems to be highly subjective. To a certain degree, this subjectivity also reflects the variation in the potential intents of the end user of the system. In order to allow for future research on this context, for each topic of the test set, we collected annotations by three annotators. We investigate two broadly employed clustering evaluation measures to estimate the degree of agreement between any two annotations: the Adjusted Rand Index (ARI) [28] and the Normalized Mutual Information (NMI) [31]. ARI measures the similarity between two groupings with correction for chance agreement. ARI scores range between -1 and 1. Positive ARI values indicate similarity with a score of 1 corresponding to a perfect agreement. NMI is bound to the range $[0, 1]$ with values close to 1 indicating significant agreement. Figure 5 shows the distribution of the achieved scores. Both evaluation measures show the same tendency and a broad variability of the level of agreement between two annotations. ARI ranges between 0 and 0.79 ($mean$=0.26 ± 0.08). NMI ranges between 0.14 and 0.88 ($mean$=0.51 ± 0.06). These results confirm the factual variation (subjectivity) in the annotations.

Finally, Table 2 summarizes the general statistics of the visual diversity annotations for the development and test sets. Overall, both data sets show similar characteristics in terms of number of clusters and number of images per cluster. Noteworthy is the strongly varying number of images per cluster from a single image to more than 100 images in a cluster. This fact is especially relevant when considering a clustering method for solving the diversification problem as the prospective clustering approach should be able to cope with partially strongly imbalanced data.

Table 2: Diversity statistics.

	#clusters			#images/cluster		
	min	max	mean±std	min	max	mean±std
devset	4	25	17±6	1	142	9±8
testset	3	25	14±4	1	159	14±13

4 PROVIDED DESCRIPTORS

To facilitate participation from both information retrieval and multimedia retrieval communities, we provide a broad range of precomputed content-based descriptors:

- *General purpose, visual-based descriptors* extracted using the LIRE library[5] [21]: auto color correlogram (ACC, 256 dimensions) [23]; color and edge directivity descriptor (CEDD, 144 dimensions) [4], fuzzy color and texture histogram (FCTH, 192 dimensions) [5], Gabor texture (60 dimensions), joint composite descriptor (JCD, 168 dimensions) [6], several MPEG7 features including color layout (33 dimensions), edge histogram (80 dimensions), and scalable color (64 dimensions) [22], pyramid of histograms of orientation gradients (PHOG, 630 dimensions) [2], and bag-of-words of speeded up robust features (SURF, 2,000 dimensions) [1].
- *Convolutional neural network (CNN)-based descriptors* based on the reference model provided by the Caffe framework[6]. This model is learned with the 1,000 ImageNet classes used for the ImageNet challenge. The descriptors are extracted from the last fully connected layer (fc7, 4,096 dimensions).
- *Text-based features* including term frequency (TF), document frequency (DF) information, and their ratio (TF-IDF) calculated simply as TF/IDF. The text-based features are computed per image, per topic (query), and per user following three different interpretations of a document [15]. The default interpretation considers each image as a document. In this case, TF is the number of occurrences of a term in the metadata of an image (title, description, or tags) and DF the number of images mentioning this term in their metadata. The second interpretation considers the topic itself as a document. In this case, the metadata of all images associated with a topic are merged (concatenated) and TF represents the number of occurrences of each term in this combined topic description. Finally, the third interpretation puts the user in focus, i.e. each user is considered as a document. Again, all metadata of the images of a particular user are merged and used for the calculation of TF, DF, and TF/IDF. Although the topic- and the user-based interpretations cannot be employed to rank the underlying images directly, they provide additional context for image ranking and diversity estimation.
- *User annotation credibility descriptors* providing an estimation of the quality of the users' tag-image content relationships [11, 14, 15]. The following descriptors are provided: *visualScore* (measure of user's image relevance), *tagSpecificity* (average tag specificity per user), *photoCount* (total number of images the user shared), *uniqueTags* (proportion of unique tags), *uploadFrequency* (average time between two consecutive uploads), *bulkProportion* (the proportion of tags that appear identical for at least two distinct images), *meanPhotoViews* (mean view numbers of the user's images), *meanTitleWordCounts* (mean number of words in the title of the user's images), *meanTagsPerPhoto* (mean number of tags users add for their images), *meanTagRank* (mean rank of a user's tags in a list in which the tags are sorted in descending order according to the number of appearances in a large subsample of Flickr images), and *meanImageTagClarity* (adaptation of the Image Tag Clarity [32] using a TF-IDF language model as individual tag language model).

5 MEDIAEVAL 2017 VALIDATION

We validated the proposed dataset during the 2017 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative for Multimedia Evaluation[7]. The goal of the task was to refine the images retrieved as a result to a given text-based query by providing a ranked list of up to 50 photos that are both relevant and visually diversified representations of the query [35]. In total, 29 runs were submitted for the final evaluation on the test set. Performance was assessed for both diversity and relevance using cluster recall at X ($CR@X$), precision at X ($P@X$), and their harmonic mean $F1@X$. $CR@X$ reflects the diversification quality of a given image result set and corresponds to the ratio of the number of clusters from the ground truth that are represented in the top X results. For each query, we compute $CR@X$ for each one of the three available ground truth diversity annotations and select the one which maximizes $CR@X$. Since the clusters in the ground truth consider relevant images only, the relevance of the top X results is implicitly measured by $CR@X$. Nevertheless, $P@X$ provides a more precise view on the relevance of a particular image set since it directly measures the relevance among the top X images. We considered various cut off points, i.e. $X = \{5, 10, 20, 30, 40, 50\}$. Additionally, we provided two further evaluation metrics, which are well-established in the information retrieval community, the *intent-aware expected reciprocal rank* ($ERR\text{-}IA@X$) [3] and the *α-normalized discounted cumulative gain* ($\alpha\text{-}nDCG@X$) [7] metrics. The final official ranking metric was $F1@20$ 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 a common user behavior inspecting the first page of results.

Table 3 summarizes the distribution of the employed features and feature combinations by the submitted approaches showing a slight tendency towards the combinations of deep learning technologies and text-based analysis. Figure 6 visualizes the achieved performance results in terms of $CR@20$ and $P@20$. Overall, only 55% of all submitted runs (top right area in Figure 6) succeeded in improving both precision and recall in comparison to the Flickr baseline showing the performance of the raw (original) Flickr retrieval results. Moreover, 17% of all submitted runs resulted in lower performance than the Flickr baseline (bottom left area in Figure 6). The top performance (red square in Figure 6) was achieved by a pseudo-relevance feedback approach using a cross-media similarity measure [29]: $F1@20 = 0.7045$, $CR@20 = 0.6786$, and $P@20 = 0.7821$.

Finally, we investigate the differences between the highest and the lowest achievable $F1@20$ scores according to the three available annotations for each topic of the test set. Figure 7 shows the

[5] http://www.lire-project.net/
[6] http://caffe.berkeleyvision.org/
[7] http://www.multimediaeval.org/

Table 3: Employed features and feature combinations.

	id	visual general	visual CNN	text	credibility	#runs
single	1	x				3
	2		x			4
	3			x		6
	4				x	1
multi	5	x		x		4
	6		x	x		8
	7			x	x	2
	8	x		x	x	1
						29

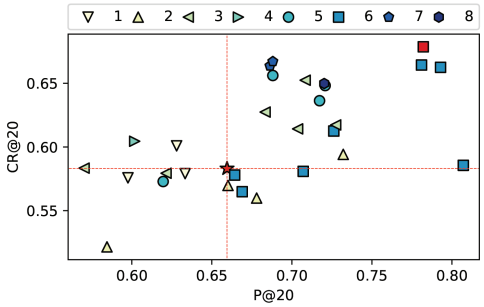


Figure 6: Performance results of the submitted runs. The red star symbol corresponds to the Flickr baseline. The further shapes are assigned to ids of the different features/feature combination as listed in Table 3. The red square indicates the top achieved performance in terms of $F1@20$ of 0.7045 [29].

distribution of the differences across all 29 submitted runs for each topic. The results show high sensitivity to the employed annotation and underlying topic/query and stress again the role of subjectivity in the annotation and retrieval process. For the query *firefighter helmet*, for example, the differences in the $F1@20$ score range up to 0.56. In contrast, the variation in the performance on the queries *bus stop* or *double door* is significantly lower but may still result in a difference of 0.1 (or 10%) in the final $F1@20$ score. Overall, the difference in the $F1@20$ score exceeds 0.1 in 57% and even 0.2 in 19% of all possible cases.

Given the multiple available annotations per topic, another possibility to evaluate the results is to consider the average performance per topic rather than selecting the top one. Such an evaluation aims at the satisfaction of as many end users as possible. Surprisingly, the order of the runs' performance does not change much (except for a few switches between runs with minor differences) although the overall scores are notably lower. Table 4 summarizes the top achieved results in comparison to the Flickr baseline as a reference for comparison with future work.

6 CONCLUSION

This paper introduces the SubDiv17 dataset that is specifically designed for benchmarking approaches aiming at the visual diversification of image search results. The dataset contains more than 50,000 real-world images and their metadata retrieved from Flickr for ca. 200 general-purpose, multi-topic queries. The dataset includes multiple ground truth annotations to facilitate the investigation of the subjectivity aspect in the general task of visual diversification of image search results.

Table 4: The top result achieved on the test data in comparison to the Flickr baseline. All performance measures are reported at cut-off 20 (@20).

	P	CR	F1	ERR-IA	α-n DCG
top $F1@20$/topic					
NLE, run#3 [29]	0.7821	0.6786	0.7045	0.7334	0.6894
Flickr baseline	0.6595	0.5831	0.5922	0.6096	0.5787
mean $F1@20$/topic					
NLE, run#3 [29]	0.7821	0.5578	0.6272	0.7093	0.6459
Flickr baseline	0.6595	0.4725	0.5263	0.5860	0.5457

The dataset including images, metadata, pre-computed descriptors, and ground truth annotations (both relevance and visual diversification) are publicly available[8]. To support the reproducibility of the exact conditions of the MediaEval task, we also provide the official annotation tool, a sample run file and a detailed description of the data and evaluation format.

ACKNOWLEDGMENTS

This work was partially supported by the Romanian Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002.

REFERENCES

[1] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding* 110, 3 (2008), 346–359. https://doi.org/10.1016/j.cviu.2007.09.014

[2] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Representing Shape with a Spatial Pyramid Kernel. In *ACM International Conference on Image and Video Retrieval (CIVR)*. ACM, New York, NY, USA, 401–408. https://doi.org/10.1145/1282280.1282340

[3] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *ACM Conference on Information and Knowledge Management (CIKM)*. ACM, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[4] Savvas A. Chatzichristofis and Yiannis S. Boutalis. 2008. CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. In *International Conference on Computer Vision Systems (ICCV)*. Springer-Verlag, Berlin, Heidelberg, 312–322. https://doi.org/10.1007/978-3-540-79547-6_30

[5] Savvas A. Chatzichristofis and Yiannis S. Boutalis. 2008. FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval. In *Internationsl Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE Computer Society, Washington, DC, USA, 191–196. https://doi.org/10.1109/WIAMIS.2008.24

[6] Savvas A. Chatzichristofis, Yannis S. Boutalis, and Mathias Lux. 2009. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Signal Processing, Pattern Recognition and Applications (SPPRA)*. ACTA Press, Calgary, Canada, 134–140.

[7] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[8] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[9] Giorgos Giannopoulos, Marios Koniaris, Ingmar Weber, Alejandro Jaimes, and Timos Sellis. 2015. Algorithms and criteria for diversification of news article comments. *Journal of Intelligent Information Systems* 44, 1 (2015), 1–47. https://doi.org/10.1007/s10844-014-0328-1

[10] Jean D. Gibbons and Subhabrata Chakraborti. 2010. *Nonparametric Statistical Inference (Statistics: a Series of Textbooks and Monogrphs)* (5 ed.). CRC, Boca Raton, FL, USA.

[8] https://doi.org/10.5281/zenodo.1219444

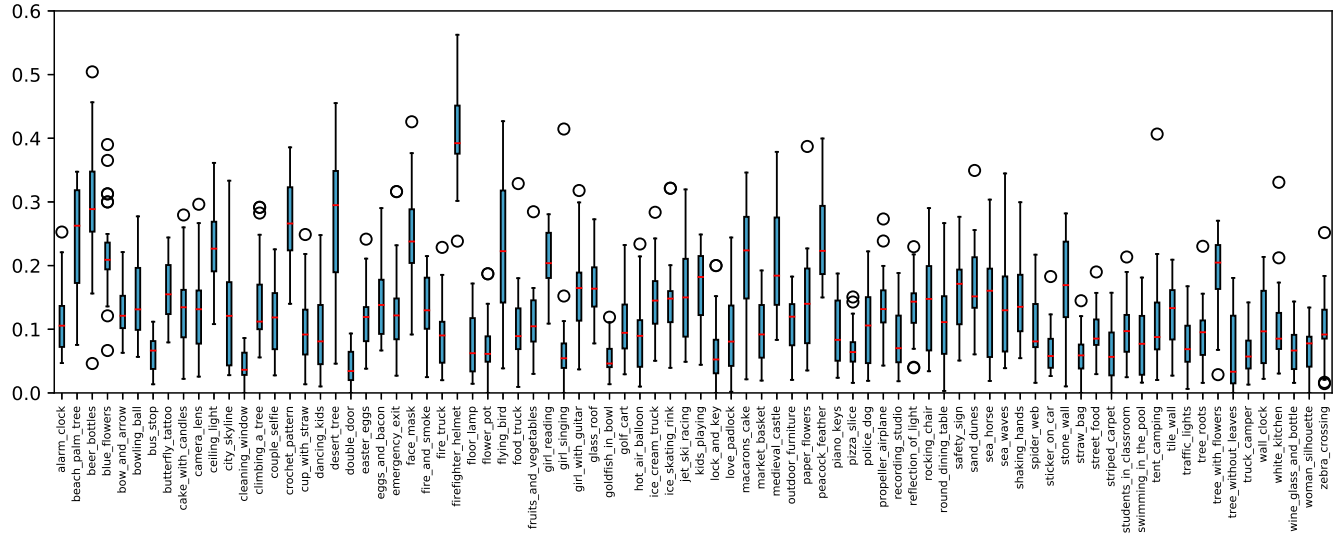


Figure 7: Distribution of the differences between the highest and lowest possible $F1@20$ score for each topic/query.

[11] Alexandru Lucian Gînscă, Adrian Popescu, Bogdan Ionescu, Anil Armagan, and Ioannis Kanellos. 2014. Toward an Estimation of User Tagging Credibility for Social Image Retrieval. In *ACM International Conference on Multimedia*. ACM, New York, NY, USA, 1021–1024. https://doi.org/10.1145/2647868.2655033

[12] Kilem Li Gwet. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring The Extent of Agreement Among Raters* (4 ed.). Advanced Analytics, LLC, Gaithersburg, MD, USA.

[13] Bogdan Ionescu, Alexandru Lucian Gînscă, Bogdan Boteanu, Mihai Lupu, Adrian Popescu, and Henning Müller. 2016. Div150Multi: A Social Image Retrieval Result Diversification Dataset with Multi-topic Queries. In *ACM Multimedia Systems Conference*. ACM, New York, NY, USA, 46:1–46:6. https://doi.org/10.1145/2910017.2910620

[14] Bogdan Ionescu, Anca-Livia Radu, María Menéndez, Henning Müller, Adrian Popescu, and Babak Loni. 2014. Div400: A Social Image Retrieval Result Diversification Dataset. In *ACM Multimedia Systems Conference*. ACM, New York, NY, USA, 29–34. https://doi.org/10.1145/2557642.2563670

[15] Bogdan Ionescu, Anca-Livia Radu, María Menéndez, Henning Müller, Adrian Popescu, and Babak Loni. 2015. Div150Cred: A Social Image Retrieval Result Diversification with User Tagging Credibility Dataset. In *ACM Multimedia Systems Conference*. ACM, New York, NY, USA, 207–212. https://doi.org/10.1145/2713168.2713192

[16] Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User Intent in Multimedia Search: A Survey of the State of the Art and Future Challenges. *Comput. Surveys* 49, 2 (2016), 36:1–36:37. https://doi.org/10.1145/2954930

[17] Marios Koniaris, Giorgos Giannopoulos, Timos Sellis, and Yiannis Vasileiou. 2014. *Diversifying Microblog Posts*. Springer International Publishing, Cham, 189–198. https://doi.org/10.1007/978-3-319-11746-1_14

[18] Shangsong Liang, Fei Cai, Zhaochun Ren, and Maarten de Rijke. 2016. Efficient Structured Learning for Personalized Diversification. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (2016), 2958–2973. https://doi.org/10.1109/TKDE.2016.2594064

[19] Shangsong Liang, Emine Yilmaz, Hong Shen, Maarten De Rijke, and W. Bruce Croft. 2017. Search Result Diversification in Short Text Streams. *ACM Transactions on Information Systems* 36, 1 (2017), 8:1–8:35. https://doi.org/10.1145/3057282

[20] Mengwen Liu, Yi Fang, Alexander G. Choulos, Dae Hoon Park, and Xiaohua Hu. 2017. Product review summarization through question retrieval and diversification. *Information Retrieval Journal* 20, 6 (2017), 575–605. https://doi.org/10.1007/s10791-017-9311-0

[21] Mathias Lux. 2011. Content Based Image Retrieval with LIRe. In *ACM International Conference on Multimedia*. ACM, New York, NY, USA, 735–738. https://doi.org/10.1145/2072298.2072432

[22] B.S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan, and Akio Yamada. 2001. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology* 11, 6 (2001), 703–715. https://doi.org/10.1109/76.927424

[23] Mandar Mitra, Ramin Zabih, Jing Huang, Wei-Jing Zhu, and S. Ravi Kumar. 1997. Image Indexing Using Color Correlograms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA, 762–768. https://doi.org/10.1109/CVPR.1997.609412

[24] Kaweh Djafari Naini, Ismail Sengor Altingovde, and Wolf Siberski. 2016. Scalable and Efficient Web Search Result Diversification. *ACM Transactions on the Web* 10, 3 (2016), 15:1–15:30. https://doi.org/10.1145/2907948

[25] Stefanie Nowak and Stefan Rüger. 2010. How Reliable Are Annotations via Crowdsourcing: A Study About Inter-annotator Agreement for Multi-label Image Annotation. In *International Conference on Multimedia Information Retrieval (ICMR)*. ACM, New York, NY, USA, 557–566. https://doi.org/10.1145/1743384.1743478

[26] Makbule Gulcin Ozsoy, Kezban Dilek Onal, and Ismail Sengor Altingovde. 2014. *Result Diversification for Tweet Search*. Springer International Publishing, Cham, 78–89. https://doi.org/10.1007/978-3-319-11746-1_6

[27] Robert Gilmore Pontius, Jr and Marco Millones. 2011. Death to Kappa: Birth of Quantity Disagreement and Allocation Disagreement for Accuracy Assessment. *International Journal of Remote Sensing* 32, 15 (2011), 4407–4429. https://doi.org/10.1080/01431161.2011.552923

[28] William M. Rand. 1971. Objective Criteria for the Evaluation of Clustering Methods. *J. Amer. Statist. Assoc.* 66, 336 (1971), 846–850. https://doi.org/10.2307/2284239

[29] Jean-Michel Renders and Gabriela Csurka. 2017. NLE@MediaEval'17: Combining Cross-Media Similarity and Embeddings for Retrieving Diverse Social Images. In *MediaEval 2017 Multimedia Benchmark Workshop*, Vol. 1984. CEUR-WS.org.

[30] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search Result Diversification. *Foundations and Trends in Information Retrieval* 9, 1 (2015), 1–90. https://doi.org/10.1561/1500000040

[31] Alexander Strehl and Joydeep Ghosh. 2003. Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research* 3 (2003), 583–617. https://doi.org/10.1162/153244303321897735

[32] Aixin Sun and Sourav S. Bhowmick. 2009. Image Tag Clarity: In Search of Visual-representative Tags for Social Images. In *SIGMM Workshop on Social Media*. ACM, New York, NY, USA, 19–26. https://doi.org/10.1145/1631144.1631150

[33] Duong Chi Thang, Nguyen Thanh Tam, Nguyen Quoc Viet Hung, and Karl Aberer. 2015. *An Evaluation of Diversification Techniques*. Springer International Publishing, Cham, 215–231. https://doi.org/10.1007/978-3-319-22852-5_19

[34] Jun Xu, Long Xia, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2017. Directly Optimize Diversity Evaluation Measures: A New Approach to Search Result Diversification. *ACM Transactions on Intelligent Systems and Technology* 8, 3 (2017), 41:1–41:26. https://doi.org/10.1145/2983921

[35] Maia Zaharieva, Bogdan Ionescu, Alexandru Lucian Gînscă, Rodrygo L.T. Santos, and Henning Müller. 2017. Retrieving Diverse Social Images at MediaEval 2017: Challenges, Dataset and Evaluation. In *MediaEval 2017 Multimedia Benchmark Workshop*, Vol. 1984. CEUR-WS.org.

[36] Kaiping Zheng, Hongzhi Wang, Zhixin Qi, Jianzhong Li, and Hong Gao. 2016. A survey of query result diversification. *Knowledge and Information Systems* 51, 1 (2016), 1–36. https://doi.org/10.1007/s10115-016-0990-4

[37] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu. 2014. Learning for Search Result Diversification. In *ACM SIGIR Conference on Research Development in Information Retrieval*. ACM, 293–302. https://doi.org/10.1145/2600428.2609634