Hierarchical Clustering Pseudo-Relevance Feedback for Social Image Search Result Diversification

Bogdan Boteanu, Ionuţ Mironică, Bogdan Ionescu,
LAPI, University "Politehnica" of Bucharest, 061071 Bucharest, Romania,
Email: {*bboteanu, imironica, bionescu*}*@alpha.imag.pub.ro*

***Abstract*—This article addresses the issue of social image search result diversification. We propose a novel perspective for the diversification problem via Relevance Feedback (RF). Traditional RF introduces the user in the processing loop by harvesting feedback about the relevance of the search results. This information is used for recomputing a better representation of the data needed. The novelty of our work is in exploiting this concept in a completely automated manner via pseudo-relevance, while pushing in priority the diversification of the results, rather than relevance. User feedback is simulated automatically by selecting positive and negative examples from the initial query results. Unsupervised hierarchical clustering is used then to re-group images according to their content. Diversification is finally achieved with a re-ranking approach. Experimental validation on Flickr data shows the advantages of this approach.**

I. Introduction

An efficient retrieval system should be able to *summarize* search results and give a global view so that it surfaces results that are both *relevant* and covering different aspects, i.e., *diverse*, of the query. Most of the queries involve many declinations such as for instance sub-topics, e.g., animals are of different species, cars are of different types and producers, points of interest can be photographed from different angles and so on. By widening the pool of possible results, one can increase the likelihood of the retrieval system to provide the user with information needed and thus to increase its efficiency. Relevance was more thoroughly studied in existing literature than diversification [1] and even though a considerable amount of diversification literature exists (mainly in the text-retrieval), the topic remains important, especially in multimedia [2].

The key of the entire diversification process is to mitigate the two components, relevance and diversity, which in general tend to be antinomic: too much diversification may result in losing relevant items while increasing solely the relevance will tend to provide many near duplicates. For instance, authors in [3] use lightweight clustering in combination with a dynamic weighting function of visual features to best capture the discriminative aspects of image results. Several diversification scenarios are investigated: folding — appreciates the original ranking by assigning a larger probability of being a representative to higher ranked images; Max-Min — tries to get as visually diverse representatives as possible by using a max-min heuristic on the distances between sub-topic representatives; election — interleaves the processes of representative selection and cluster formation and uses the idea that every image decides by which image (besides itself) it is best represented, which in the end determine its chances of being elected as representative. Authors in [4] aim to populate a database with high precision and diverse photos of different entities by re-evaluating relational facts about the entities. They use a model parameter that is estimated from a small set of training entities. Visual similarity is exploited using the classic Scale-Invariant Feature Transform (SIFT). Authors in [2] address the problem of image diversification in the context of automatic visual summarization of geographic areas and exploits user-contributed images and related explicit and implicit metadata collected from popular content-sharing websites. The approach is based on a Random walk scheme with restarts over a graph that models relations between images, visual features, associated text, as well as the information on the uploader and commentators.

In this paper, we exploit a novel perspective of the diversification via the use of Relevance Feedback techniques (RF). Traditional RF attempts to introduce the user in the loop by harvesting feedback about the relevance of the search results. This information is used as ground truth for recomputing a better representation of the data needed. We propose an alternative pseudo-relevance solution for rending this process completely automatic, while still maintaining a high performance.

The reminder of the paper is organized as following. Section II presents the current RF state-of-the-art and positions our contribution. The proposed approach is explained in Section III. Section IV and V deal with the experimental validation. Finally, Section VI concludes the paper.

II. Previous work

Relevance feedback has proven to increase retrieval accuracy and gives more personalized results for the user. One of the earliest and most successful RF algorithms is the Rocchio's algorithm [5] (which is still used at the present time). Using the set of relevant and non-relevant documents selected from the current user relevance feedback window, the Rocchio's algorithm modifies the features of the initial query by adding the features of positive examples and subtracting the features of negative examples to the original feature. Another relevant approach is the Relevance Feature Estimation (RFE) algorithm [6]. It assumes that for a given query, according to the user's subjective judgment, some specific features may be more important than others. A re-weighting strategy is adopted which analyzes the relevant objects in order to understand which dimensions are more important than others in determining "what makes a result relevant". Features with higher variance with respect to the relevant queries lead to lower importance factors than elements with reduced variation.

Part of this work was supported under InnoRESEARCH POSDRU/159/1.5/S/132395 and POSDRU/174/1.3/S/149155.

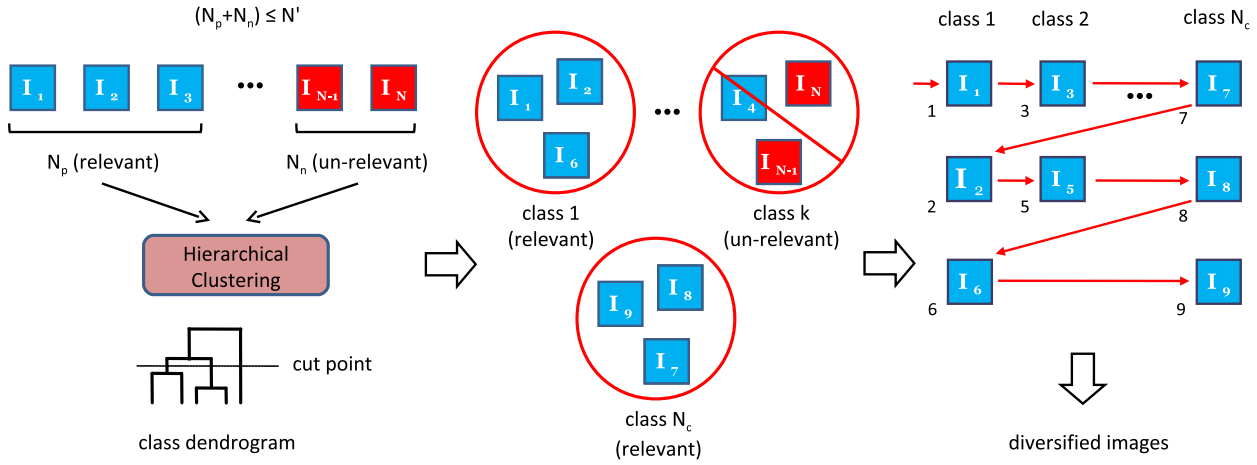

Fig. 1: General scheme of the proposed approach: Selection of positive and negative examples (N_p and N_n, respectively; N' is the total number of returned images), Clustering and pruning (N_c is the number of resulting classes), Diversification.

More recently, machine learning techniques found their application in relevance feedback approaches. The relevance feedback problem can be formulated either as a two class classification of the negative and positive samples; or as an one class classification problem, i.e., separate positive samples by negative samples. After a training step, all the results are ranked according to the classifiers's confidence level [7], or classified as relevant or irrelevant depending on some output functions [8]. Some of the most successful techniques use Support Vector Machines, Nearest Neighbor approaches, classification trees, e.g., use of Random Forests; or boosting techniques, e.g., AdaBoost.

Almost all the existing relevance feedback techniques focus exclusively on improving the relevance of the results. In this paper we propose a novel pseudo-relevance perspective that exploits the concept of relevance feedback while pushing in priority the diversification, in an automated manner. User feedback is simulated automatically by selecting positive and negative examples from the initial query results. Unsupervised hierarchical clustering is used to re-group images according to their contents. Diversification is finally achieved with a re-ranking approach. Experimental validation shows the benefits of this approach which outperforms other relevance feedback and state-of-the-art diversification approaches.

III. Proposed Approach

In this paper we propose a different perspective for improving image search result diversification and relevance. It exploits the concept of *pseudo relevance feedback*. The method's diagram is presented in Figure 1. The proposed approach operates on top of an existing retrieval system and works as a re-ranking step that refines the initial query results. In a first step, some positive and negative examples are selected from the query results (see Section III-A). Then, an unsupervised classification step is used to cluster these examples. Each obtained cluster is further evaluated based on the number of the relevant and un-relevant images within. This step will ensure the relevance of the refined images (see Section III-B). The final step is the actual diversification. Following the initial ranking of the retrieved images, cluster images are progressively selected to form the refined diversified query results (see Section III-C). Each of the processing steps is detailed in the following.

A. Selection of positive and negative examples

The first step of the proposed approach consists of selecting a number of positive and negative query examples. Instead of using a classic relevance feedback strategy where the user is supposed to provide these examples, we use a pseudo-relevance feedback assumption [9].

In general, the actual retrieval systems are capable of providing high quality results in terms of relevance, e.g., see Google Image Search, Flickr, Panoramio, etc. Therefore, most of the very first returned results tend to be relevant to the query. In contrast, the very last of the results are highly likely to be noisy and un-relevant. For instance, if we consider as example Flickr's state-of-the-art retrieval system, results in [10], [11] show that, in average, among the first 50 returned images, at least 37 images are relevant to the query (i.e., 75.37% — estimate obtained for 549 location related queries; the query is formulated as keywords using the location's name). These results are in support of our relevance assumption.

Therefore, we retain the first N_p images from the initial ranking as positive examples and the last N_n images as negative examples (denoted *hypothesis 1*). This leads to a total number of N examples ($N = N_p + N_n$) that constitutes an automatic ground truth. To address the borderline case when the number of returned images, $N^{'}$, is lower than N, we adopt the following approximations where the positive examples are a fraction of the total number of examples:

$$N_p^{'} = \frac{N_p}{N} \cdot N^{'}, \quad N_n^{'} = N^{'} - N_p^{'} \tag{1}$$

The immediate advantage of this strategy is in the complete automation of the relevance feedback process. No real user interaction is actually required, which reduces significantly the processing time as well as the need for conducting complex user studies.

B. Clustering and pruning

Equipped with the ground truth, we use a clustering strategy to group similar appearance images. We selected a Hierarchical Clustering (HC) scheme that proved highly efficient in various diversification scenarios [12], [10]. The HC scheme uses the "bottom up" approach (agglomerative)[1], thus starting with each of the images assigned to an individual cluster and ending with a single cluster. Besides its low complexity, HC has the advantage of providing a dendrogram of classes by grouping images iteratively based on a certain distance metric. This allows for adapting the number of output classes to the target scenario based on the selection of a cutting point of the dendrogram. HC is applied only to the selected positive and negative examples.

Once we achieve the clustering, we adopt a supplementary pruning step. A class is declared un-relevant if it contains only negative examples or if the number of negative examples is higher than the positive ones, namely: $N_n^{(i)} \geq 0.5 \cdot N^{(i)}$, where $N_n^{(i)}$ is the number of negative examples in class i and $N^{(i)}$ is the total number of examples in class i (denoted *hypothesis 2*). This assumption is based on the fact that cluster images are supposed to be similar with each other. Therefore, if a significant number of negative examples is present, there is a high probability that all the images are in fact negative examples and were assigned wrongly to the positive category.

C. Diversification

The final step is the actual diversification of the results. To improve also the relevance, we take into account the initial ranking of the results, as the first retrieved images have a higher probability to be relevant than the last ones. To enforce the diversity, we restrict the output to contain at least one image from each HC generated cluster. The algorithm is the following.

Firstly, for each of the HC output relevant classes (the classes declared as un-relevant are discarded from diversification), the images are sorted according to their initial ranking, so that the first image in a class is the one which has the highest rank in the initial retrieval results. Considering the order described above and starting with the first class, i.e., the class labeled as the first one by the HC scheme, we select as output each class first ranked image. This leads to N_c images, where N_c is the total number of classes. The process is repeated iteratively, and classes are covered again by selecting the second ranked images, third ranked and so on.

If in a certain class, the number of images is lower than the number of the current iteration (e.g., in the third iteration we attempt to select images from a class that has only two images), then that particular class is disregarded in the current and further iterations. The process is repeated until the desired number of images is achieved.

IV. Experimental Setup

A. Dataset

To test our approach, we selected a publicly available image retrieval diversification dataset, namely the Div150Cred dataset [10], that was used in the MediaEval 2014 Retrieving Diverse Social Images Task[2]. It contains 153 location related queries (e.g., museums, bridges, parks, monuments, etc) with up to 300 photos per query and associated metadata retrieved from Flickr using Flickr's default "relevance" algorithm (a total of 45,375 images). Images are annotated for both relevance and diversity by human assessors. In particular, for diversity, images are clustered into similar appearance classes. The data is divided into a development set containing 30 queries (8,923 images) intended for designing/training the approaches and a test set of 123 queries (36,452 images) for the actual evaluation. To be able to compare to the benchmarking results, we use the same experimenting conditions and perform the evaluation on test set.

B. Evaluation metrics

To assess performance, we compute the standard cluster recall at a cutoff at X images ($CR@X$) [13] and the precision at X images ($P@X$), given by:

$$CR@X = \frac{N}{N_{gt}}, \quad P@X = \frac{N_r}{X} \tag{2}$$

where N is the number of image clusters represented in the first X ranked images and N_{gt} is the total number of image clusters from the ground truth (N_{gt} is limited to a maximum of 25 clusters for this dataset), N_r is the number of relevant images among the first X ranked results. $CR@X$ assesses how many clusters from the ground truth are represented among the top X results provided by the retrieval system. Since clusters are made up of relevant photos only, relevance of the top X results is implicitly measured by $CR@X$, along with diversity. To have a clearer view of relevance, $P@X$ measures the number of relevant photos among the top X results.

Finally, to account for an overall assessment of both diversity and precision, we also report $F1@X$, i.e., the harmonic mean of $CR@X$ and $P@X$:

$$F1@X = 2 \cdot \frac{CR@X \cdot P@X}{CR@X + P@X} \tag{3}$$

Results are reported as overall average values over all the queries in the dataset.

C. Content description

In the clustering process images are represented with content descriptors. Although the approach is not dependent on a certain type of description scheme, the choice of the descriptors influence significantly the results and should be adapted to the specificity of the evaluation data.

Given the specificity of the task, i.e., diversifying visual contents, we tested a broad category of visual descriptors which are known to perform well in image retrieval tasks [14]: global color naming histogram (CN, 11 values) — maps colors to 11 universal color names: "black", "blue", "brown", "grey", "green", "orange", "pink", "purple", "red", "white", and "yellow" [15]; global Histogram of Oriented Gradients (HoG, 81 values) — represents the HoG feature computed on 3 by 3 image regions [16]; global color moments computed on the HSV Color Space (CM, 9 values) — represent the

[1] http://www.mathworks.com/help/stats/hierarchical-clustering.html/

[2] http://ceur-ws.org/Vol-1263/

first three central moments of an image color distribution: mean, standard deviation and skewness [17]; global Locally Binary Patterns computed on gray scale representation of the image (LBP, 16 values) [18]; global Color Structure Descriptor (CSD, 64 values) — represents the MPEG-7 Color Structure Descriptor computed on the HMMD color space [19]; global statistics on gray level Run Length Matrix (GLRLM, 44 dimensions) — represents 11 statistics computed on gray level run-length matrices for 4 directions: Short Run Emphasis, Long Run Emphasis, Gray-Level Non-uniformity, Run Length Non-uniformity, Run Percentage, Low Gray-Level Run Emphasis, High Gray-Level Run Emphasis, Short Run Low Gray-Level Emphasis, Short Run High Gray-Level Emphasis, Long Run Low Gray-Level Emphasis, Long Run High Gray-Level Emphasis [20]; global descriptor which is obtained by the concatenation of all values.

In addition to visual information, we experimented also with text descriptors. In particular we use histogram representations of term frequency (TF), document frequency (DF) and term frequency - inverse document frequency (TF-IDF) information computed on image metadata (descriptor average size of 500 values); as well as an estimate of the user image annotation credibility (denoted Credibility — an estimation of the global quality of tag-image content relationships for a user's contributions; 9 values descriptor). Descriptors are detailed in [10].

Descriptors were experimented individually or in combination. Fusion is carried out with an early fusion approach preceded by a max-min value normalization.

D. Pre-filtering

To improve more the relevance of the results, we pass the initial retrieved images through several pre-filtering steps. Firstly, we use the Viola-Jones [21] *face detector* to filter out images with persons as the main subject. These images are in general un-relevant for the common user. The output of the filter for an image consists in a number of pairs of coordinates, indicating where it is most likely to have the face of a person. If this number is greater than a threshold T_f, then the image is considered to contain faces and it is removed.

Secondly, we use an *image blur detector* to remove the out of focus images. Regardless their content, severely blurred images are in general not satisfactory results for a query. We use the aggregation of 10 state-of-the-art blur indicators as implemented by Said Pertuz[3], namely: Brenner's indicator, graylevel variance, normalized GLV, energy of gradient, thresholded gradient, energy of Laplacian, modified Laplacian, variance of Laplacian, Tenengrad, and sum of wavelet coefficients. An image is rejected if the average of the normalized values is lower than a threshold T_b.

Finally, in particular for this dataset, we use a *GPS-based filter*. The filter rejects the images that are positioned too far away from the query location, and therefore which cannot be relevant shots for that location. We use a tolerance radius of T_d Kms. For accurate results, distance between GPS coordinates is computed using the Harvesine formula[4].

[3]http://www.mathworks.com/matlabcentral/fileexchange/27314-focus-measure/content/fmeasure/fmeasure.m

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

TABLE I: Diversification results for various descriptor - HC parameter combinations (best results are represented in bold).

descriptor	N_p-N_n-N_c	$P@20$	$CR@20$	$F1@20$
HoG	130-3-31	0.7549	0.4064	0.5199
CSD	80-3-32	0.7663	0.4216	0.5345
CM	110-15-35	0.7711	0.4115	0.5288
CN	80-3-20	0.761	0.4181	0.5315
GLRLM	90-3-35	**0.7972**	0.4133	0.5368
LBP	80-15-34	0.7699	0.4098	0.5263
all visual (early fusion)	120-21-35	0.7598	0.4245	0.5349
TF	110-18-29	0.787	**0.4515**	**0.5657**
DF	90-9-24	0.7874	0.4335	0.5518
TF-IDF	120-3-24	0.7837	0.4457	0.5588
all text (early fusion)	90-12-26	0.7862	0.4377	0.5544
Credibility	110-3-20	0.6846	0.4296	0.5209
all (early fusion)	160-9-33	0.6841	0.4191	0.5136

V. Experimental results

This section presents the experimental validation results. We have conducted the following experiments: Section V-A deals with method's parameter tuning (e.g., pre-filtering of data, choice of descriptors, choice of HC parameters); Section V-B demonstrates the usefulness of the pseudo-relevance hypothesis; Section V-C compares the proposed method to reference relevance feedback approaches from the literature; finally, Section V-D situates our results in the context of the state-of-the-art diversification approaches.

A. Parameter tuning

The performance of the proposed approach depends on the choice of several parameters.

The first test consists of determining the best descriptor - HC parameter combination (i.e., choice of N_p, N_n, N_c — see Section III-A). No pre-filtering is used. We experiment with varying N_p (number of positive examples) from 80 to 160 with a step of 10, N_n (number of negative examples) from 0 to 21 with a step of 3, and N_c (number of diversity classes) from 20 to 35 with a step of 1. We use the HC's standard Euclidean distance metric. Table I presents the results for the optimal configurations. We report the metrics at a cutoff at 20 images which is the official metric for this dataset. The best results in terms of $F1@20$ are obtained with TF textual descriptors and $N_p = 110$, $N_n = 18$, $N_c = 29$, which yielded $F1@20 = 0.5657$. Surprisingly, text descriptors proved to be very efficient for diversification, maintaining in the same time a good performance for relevance. An explanation for this may be the fact that social metadata provide a higher level of description that automatic visual descriptors.

The next test consists of assessing the influence of pre-filtering. We vary T_b from 0 to 0.6 with a step of 0.02, T_f from 1 to 3 with a step of 1 and T_d from 0 to 10 with a step of 5 (see Section IV-D). The best performance is achieved by taking $T_f = 1$, $T_b = 0.42$, and $T_d = 5$, which leads to $F1@20 = 0.5863$. Pre-filtering allows for a gain of around 2 percentage points over the previous best result.

The final experiment was conducted for selecting the best distance metric - centroid selection combination for the HC. For the previously selected parameters, we experiment now

TABLE II: Diversification results for various HC metrics - centroid combinations (best results are represented in bold).

distance-centroid	P@20	CR@20	F1@20
Euclidean-single (default)	0.8102	0.4682	0.5863
Euclidean-median	0.8154	0.4686	0.5882
Chebyshev-centroid	0.8106	0.4671	0.5851
Chebyshev-average	**0.8191**	**0.4752**	**0.5948**
cosine-centroid	0.8069	0.4641	0.5816

with various HC metrics, namely: Euclidean, sEuclidean, city-block, Minkowski, Chebyshev, cosine, correlation, Spearman, Hamming, and Jaccard, combined with several types of cluster centroids: average, centroid, complete, median, single, ward, and weighted. The best combination results are reported in Table II. The highest $F1@20$ is achieved for Chebyshev and average centroid, $F1@20 = 0.5948$.

B. Proof of hypotheses

In this experiment we show the benefits of two of the adopted hypotheses. Experiments were conducted using previous best performance parameter tuning.

Hypothesis 1: the adoption of the positive and negative examples (see Section III-A). We test the impact of taking as negative examples the very last of the returned images. Therefore, for the best results, we set $N_n = 0$ (no negative examples). In this case we achieve $F1@20 = 0.571$. This is lower than the use of negative examples by 2 percentage points.

Hypothesis 2: the adoption of un-relevant classes in HC (see Section III-B). To test the usefulness of building the un-relevant classes and thus removing them from the results, we experiment by considering all the classes as relevant. This yields an $F1@20 = 0.5765$ which is almost 2 percentage points lower that the result with removing un-relevant classes.

C. Comparison to relevance feedback approaches

In this section we compare our results to other relevance feedback approaches from the literature, namely: Rocchio [5] that changes the initial query point according to user's feedback, Relevance Feature Estimation [6] (RFE) that alters the feature representation by assessing features' importance and some classification-based approaches: Support Vector Machines (SVM) [7] and AdaBoost [8], which formulate the relevance feedback as a two class classification of the negative and positive samples. User relevance feedback is simulated with the images' ground truth in a window of 20 images (this is a common setting that allow good results [22]). We experimented with two situations: (1) feedback is simulated with the relevance ground truth (*relevance*); (2) feedback is simulated with the diversity ground truth by selecting one image from each image class in the initial feedback window (*diversity*). This should allow for more emphasis on the diversification. The approaches were tuned to best performing parameters.

Results are presented in Table III. The first observation is the fact that the use of diversified feedback instead of only relevance allows for improvement over the last one. However, regardless the use of actual image ground truth, the best traditional relevance feedback result in terms of $F1@20$ is 0.5172,

TABLE III: Comparison to relevance feedback approaches (RBF - Radial Basis Function kernel; best results are represented in bold).

RF approach	*feedback*	*descriptor*	$P@20$	$CR@20$	$F1@20$
proposed	pseudo-rel.	TF text	0.8191	**0.4752**	**0.5948**
Rocchio [5]	relevance	CN	**0.8549**	0.3385	0.4718
Rocchio [5]	diversity	CSD	0.7126	0.3429	0.455
RFE [6]	relevance	CN	0.828	0.3239	0.4526
RFE [6]	diversity	CN	0.787	0.3561	0.4773
SVM RBF [7]	relevance	GLRLM	0.8508	0.369	0.505
SVM RBF [7]	diversity	all visual	0.75	0.4086	0.5172
AdaBoost [8]	relevance	GLRLM	0.8077	0.3666	0.4934
AdaBoost [8]	diversity	LBP	0.7463	0.3779	0.4935

TABLE IV: Comparison to diversification approaches (best results are represented in bold).

Approach	*pre-filtering*	*modality*	$P@20$	$CR@20$	$F1@20$
proposed	yes	text	0.8191	**0.4752**	0.5948
PRa-MM [23]	yes	text-visual	**0.8512**	0.4692	**0.5971**
SocSens [24]	no	text-visual	0.815	0.4747	0.5943
CEALIST [25]	yes	visual	0.7931	0.4563	0.571
TUW [26]	yes	visual	0.7687	0.4497	0.5602
UNED [27]	no	text	0.7772	0.4343	0.5502
Folding [3]	no	visual	0.778	0.4527	0.5639
Max-Min [3]	no	visual	0.7435	0.4229	0.5311
Election [3]	no	visual	0.7106	0.3808	0.4875

achieved with SVM and Radial Basis Function (RBF) kernel. This is almost 8 percentage points less than the proposed approach. These results are very promising considering the fact that the proposed approach uses automatically generated feedback.

D. Comparison to diversification approaches

The final validation experiment consisted on comparing the results against state-of-the-art diversification approaches from the literature.

We compare to the 2014 MediaEval Retrieving Diverse Social Images benchmarking [28], [10] (for brevity reasons, we present only the top 5 best performing teams), namely: PRa-MM [23] — uses face, blur, GPS and user credibility-based pre-filtering. Diversification is achieved with BIRCH Hierarchical Clustering, isolated cluster removing and re-ranking. Data is represented with text (TF-IDF), visual and user credibility descriptors; SocSens [24] — uses a weighted combination of relevance scores assessed with an unsupervised classification model and diversity scores based on dissimilarity between the most similar pair of images of a subset. Data is represented with visual and text (bag of words) descriptors; CEALIST [25] — uses face and GPS pre-filtering. Diversification is achieved by ranking images according to their distance to Wikipedia images followed by k-means classification. Data is represented with visual Caffe convolutional descriptors; TUW [26] — uses GPS and description length pre-filtering. Diversification is achieved with an optimal combination of Metis, Spectral and Hierarchical Clustering. Only visual information is used; UNED [27] — uses a Formal Concept Analysis (FCA) in order to infer latent topics and apply an Hierarchical Agglomerative Clustering approach. Diversification is achieved by selecting images with highest rank-based score from each cluster. Only text information is used. We also compare

to the three visual diversification strategies proposed in [3] (introduced in Section I).

Results are presented in Table IV. The proposed approach achieves the second best result in terms of $F1@20$, after PRa-MM [23] — $F1@20 = 0.5971$, with $F1@20 = 0.5948$. However, we achieve the best diversification performance, with $CR@20 = 0.4752$. Results show that methods achieving the highest relevance are not necessarily the ones with the highest diversification, e.g., PRa-MM [23] has $P@20 = 0.8512$ compared to $P@20 = 0.8191$ achieved with the proposed approach, but their diversification is lower. In terms of modality, exploiting the text social information allows for the best performance. The use of pre-filtering techniques do improve performance in certain cases but is not strictly necessary, e.g., SocSens [24] achieves $F1@20 = 5943$ without any filtering.

VI. Conclusions

In this article we addressed the problem of social image search result diversification from the perspective of relevance feedback techniques. We proposed a novel perspective that rends the feedback process completely automatic via pseudo-relevance feedback and considers in priority the diversification, instead of the relevance of the results. The method operates on top of an existing retrieval system.

Experimental validation on Flickr data (from the 2014 MediaEval Retrieving Diverse Social Images task) show the potential of this approach. It outperforms other traditional relevance feedback approaches by as much as 8 F1-measure percentage points, even when feedback is diversified and simulated with actual ground truth. Moreover, the proposed approach achieves similar or better performance than other state-of-the-art diversification approaches from the literature. It allows in particular to achieve better diversification of the results. We therefore proved the benefits of the pseudo-relevance assumption in the context of result diversification opening new perspectives for this area of research.

Future work will mainly address exploring more complex diversification scenarios, such as the ones involving multi-concept queries where results tends to be less accurate.

References

[1] R. Priyatharshini, S. Chitrakala, "Association Based Image Retrieval: A Survey," in Mobile Communication and Power Engineering, Springer Communications, Computer and Information Science, 2013, pp. 17-26.

[2] S. Rudinac, A. Hanjalic, M.A. Larson, "Generating Visual Summaries of Geographic Areas Using Community-Contributed Images," in IEEE Transactions on Multimedia, 15(4), 2013, pp. 921-932.

[3] R. H. van Leuken, L. Garcia, X. Olivares, R. van Zwol, "Visual Diversification of Image Search Results," in International Conference on World Wide Web, 2009.

[4] B. Taneva, M. Kacimi, G. Weikum, "Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity," in ACM Web Search and Data Mining, 2010, pp. 431-440.

[5] J. Rocchio, "Relevance Feedback in Information Retrieval," in The Smart Retrieval System Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs NJ, 1971, pp. 313-323.

[6] Y. Rui, T. Huang, S.-F. Chang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues," in Visual Communication and Image Representation, 10(1), 1999, pp. 39-62.

[7] S. Liang, Z. Sun, "Sketch Retrieval and Relevance Feedback with Biased SVM Classification," in Pattern Recognition Letters, 29, 2008, pp. 1733-1741.

[8] J. Yu, Y. Lu, Y. Xu, N. Sebe, Q. Tian, "Integrating Relvance Feedback in Boosting for Content-based Image Retrieval," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, USA, 2007, pp. 965-968.

[9] G. Cao, J. Y. Nie, J. Gao, S. Robertson, "Selecting Good Expansion Terms for Pseudo-Relevance Feedback," in ACM International Conference on Research and Development in Information Retrieval, 2008.

[10] B. Ionescu, A. Popescu, M. Lupu, A.L. Ginscă, B. Boteanu, H. Müller, "Div150Cred: A Social Image Retrieval Result Diversification with User Tagging Credibility Dataset," in ACM Multimedia Systems - MMSys, Portland, Oregon, USA, 2015.

[11] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, "Div400: A Social Image Retrieval Result Diversification Dataset," in ACM Multimedia Systems - MMSys, Singapore, 2014.

[12] D. Cai, X. He, Z. Li, W.-Y. Ma, J.-R. Wen, "Hierarchical Clustering of WWW Image Search Results using Visual, Textual and Link Information," in ACM International Conference on Multimedia, 2004.

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

[14] S. Nowak, M. Huiskes, "New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageClef 2010," in Working Notes of CLEF, 2010.

[15] V. de Weijer, C. Schmid, J. Verbeek, D. Larlus, "Learning Color Names for Real-world Applications," in IEEE Trans. on Image Processing, 18(7), 2009, pp. 1512-1523.

[16] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection," in Conference On Intelligent Transportation Systems, 2009.

[17] M. Stricker, M. Orengo, "Similarity of Color Image," in SPIE Conference on Storage and Retrieval for Image and Video Databases III, vol. 2420, 1995, pp. 381-392.

[18] T. Ojala, M. Pietikinen, D. Harwood, "Performance Evaluation of Texture Measures with Classification based on Kullback Discrimination of Distributions," in IAPR International Conference on Pattern Recognition, vol. 1, 1994, pp. 582-585.

[19] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, A. Yamada, "Color and Texture Descriptors," in IEEE Trans. on Circuits and Systems for Video Technology, vol. 11(6), 2001, pp. 703-715.

[20] X. Tang, "Texture Information in Run-Length Matrices," in IEEE Trans. on Image Processing, vol.7(11), 1998.

[21] P. Viola, M. J. Jones, "Robust Real-Time Face Detection," in International Journal of Computer Vision, 57(2), pp. 137-154, 2004.

[22] B. Boteanu, I. Mironică, B. Ionescu, "A Relevance Feedback Perspective to Image Search Result Diversification," in International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, September 4-6, 2014.

[23] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, F. De Natale, "Retrieval of Diverse Images by Pre-filtering and Hierarchical Clustering," in Proceedings of the MediaEval Multimedia Benchmark Workshop, CEUR-WS.org, 1263, ISSN 1613-0073, Barcelona, Spain, 2014.

[24] E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, I. Vlahavas, "SocialSensor: Finding Diverse Images at MediaEval 2014," in Proceedings of the MediaEval Multimedia Benchmark Workshop, CEUR-WS.org, 1263, ISSN 1613-0073, Barcelona, Spain, 2014.

[25] A. L. Ginscă, A. Popescu, N. Rekabsaz, "CEA LIST's Participation at the MediaEval 2014 Retrieving Diverse Social Images Task," in Proceedings of the MediaEval Multimedia Benchmark Workshop, CEUR-WS.org, 1263, ISSN 1613-0073, Barcelona, Spain, 2014.

[26] J. R. M. Palotti, N. Rekabsaz, M. Lupu, A. Hanbury, "TUW @ Retrieving Diverse Social Images Task 2014," in Proceedings of the MediaEval Multimedia Benchmark Workshop, CEUR-WS.org, 1263, ISSN 1613-0073, Barcelona, Spain, 2014.

[27] A. Castellanos, A. Garca-Serrano, J. M. C. Recuero, "UNED @ Retrieving Diverse Social Images Task," in Proceedings of the MediaEval Multimedia Benchmark Workshop, CEUR-WS.org, 1263, ISSN 1613-0073, Barcelona, Spain, 2014.

[28] MediaEval 2014 Workshop, Eds. M. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. Sutcliffe, J. Choi, G.J.F. Jones, Barcelona, Spain, October 16-17, CEUR-WS.org, ISSN 1613-0073, 1263, http://ceur-ws.org/Vol-1263/