A Relevance Feedback Perspective to Image Search Result Diversification

Bogdan Boteanu[1], Ionuţ Mironică[1], Bogdan Ionescu[1,2]
[1]LAPI, University "Politehnica" of Bucharest, 061071, Romania,
[2]DISI, University of Trento, 38123 Povo, Italy,
Email: {*bboteanu,imironica,bionescu*}@*alpha.imag.pub.ro*

***Abstract*—An efficient information retrieval system should be able to provide search results which are in the same time *relevant* for the query but which cover different aspects, i.e., *diverse*, of it. In this paper we address the issue of image search result diversification. We propose a new hybrid approach that integrates both the automatization power of the machines and the intelligence of human observers via an optimized multi-class Support Vector Machine (SVM) classifier-based relevance feedback (RF). In contrast to existing RF techniques which focus almost exclusively on improving the relevance of the results, the novelty of our approach is in considering in priority the diversification. We designed several diversification strategies which operate on top of the SVM RF and exploit the classifiers' output confidence scores. Experimental validation conducted on a publicly available image retrieval diversification dataset show the benefits of this approach which outperforms other state-of-the-art methods.**

I. Introduction

Current photo search technology is mainly relying on employing text annotations, visual, or more recently on GPS information to provide users with accurate results for their queries. Retrieval capabilities are however still below the actual needs of the common user, mainly due to the limitations of the content descriptors, e.g., text tags tend to be inaccurate (e.g., people may tag entire collections with a unique tag) and annotation might have been done with a goal in mind that is different from the searchers goals. Automatically extracted visual descriptors often fail to provide high-level understanding of the scene while GPS coordinates capture the position of the photographer and not necessarily the position of the query.

Until recently, research focused mainly on improving the *relevance* of the results. However, an efficient information retrieval system should be able to *summarize* search results and give a global view so that it surfaces results that are both relevant and that are covering different aspects (i.e., *diverse*) of a query, e.g., providing different views of a monument rather than duplicates of the same perspective showing almost identical images. Relevance was more thoroughly studied in existing literature than diversification [1][2][3] and even though a considerable amount of diversification literature exists (mainly in the text-retrieval community), the topic remains important, especially in multimedia [4][5].

The problem of retrieval results diversification was addressed initially for text-based retrieval as a method of tackling queries with unclear information needs [6]. A typical retrieval scenario that focuses on improving the relevance of the results is based on the assumption that the relevant topics for a query belong to a single topic. However, this is not totally accurate as 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, objects have different shapes, points of interest can be photographed from different angles and so on. Therefore, one should consider equally the diversification in a retrieval scenario.

A typical text retrieval diversification approach involves two steps [7]. First, a ranking candidate set S with elements that are relevant to the user's query is retrieved. Second, a subset R of S is computed by retaining only the very relevant elements and at the same time a set that is as diverse as possible, i.e., in contrast to the other elements from the set R. The key of the entire process is to mitigate the two components (relevance and diversity — a bi-optimization process) which in general tend to be antinomic, i.e., the improvement of one of them usually results in a degradation of the other. Too much diversification may result in losing relevant items while increasing solely the relevance will tend to provide many near duplicates.

In the context of image retrieval, many approaches have been investigated. For instance, [8] addresses the visual diversification of image search results with the use of lightweight clustering techniques in combination with a dynamic weighting function of visual features to best capture the discriminative aspects of image results. Diversification is achieved by selecting a representative image from each obtained cluster. [9] jointly optimizes the diversity and the relevance of the images in the retrieval ranking using techniques inspired by Dynamic Programming algorithms. [10] aims to populate a database with high precision and diverse photos of different entities by re-evaluating relational facts about the entities. Authors 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). [4] addresses 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.

Despite the advances in the field, research on automatic image analysis techniques reached the point where further improvement of the retrieval performance may require the use of user expertise. More and more research is focused now towards the new concept of "human in the loop", i.e., including human computation in the processing chain. In its early stages, this was carried out by conducting user studies on the systems' results. However, this approach is very time consuming and far from being able to perform in real time, usually taking even months to complete. A recent perspective is to take advantage of the potential of crowdsourcing platforms [11] in which humans (i.e., users around the world) act like a computational machine that can be accessed via a computer interface. Although it shows great potential, issues such as validity, reliability, and quality control are still open to further investigation especially for high complexity tasks, such as our search diversification problem. Due to the involvement of untrained people (crowd), tackling complex tasks is less effective.

In this paper, we exploit the benefits of this concept from the perspective of hybrid approaches that integrate both the automatization power of the machines and the intelligence of human observers. Relevance Feedback (RF) techniques attempt 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. Relevance feedback proved itself efficient in improving the relevance of the results but more limited in improving the diversification. We therefore propose a classification-based relevance feedback that uses Support Vector Machines (SVM) and some diversification strategies to specifically address in priority the diversification and relevance of the results.

The remaining of the paper is organized as follows: Section II reviewers the literature on image retrieval relevance feedback and positions our approach, Section III describes the proposed approach and the diversification strategies, Section IV and Section V discusses the experimental setup and results, respectively; while Section VI concludes the paper.

II. Previous work

A general relevance feedback scenario can be formulated as: for a certain retrieval query, the user gives his opinion by marking the results as relevant or non-relevant. Then, the system automatically computes a better representation of the information needed based on this information and retrieval is further refined. Relevance feedback can go through one or more iterations of this sort. This basically improves the system response based on query related ground-truth.

Relevance feedback has proven to increase retrieval accuracy and gives more personalized results for the user [12][13][14][15][16]. Recently, a relevance feedback track was organized by TREC to evaluate and compare different relevance feedback algorithms for text descriptors [17]. However, relevance feedback was successfully used not only for text retrieval, but also for image features [12][14][15][16] and multimodal video features [13][18]. In general there are two different strategies for relevance feedback: changing the feature's representation and using a re-learning strategy via a classifier.

One of the earliest and most successful RF algorithms is the Rocchio's algorithm [19][13]. Using the set of R relevant and N 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 [12]. 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". Every feature has an importance weight computed as $w_i = 1/\sigma$ where σ denotes the variance of relevant retrievals. Therefore, features with higher variance with respect to the relevant queries lead to lower importance factors than elements with reduced variation.

More recently, machine learning techniques found their application in relevance feedback approaches. In these 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 [14][16], or classified as relevant or irrelevant depending on some output functions [20]. Some of the most successful techniques use Support Vector Machines [14], Nearest Neighbor approaches [15], classification trees, e.g., use of Random Forests [16]; or boosting techniques, e.g., AdaBoost [20].

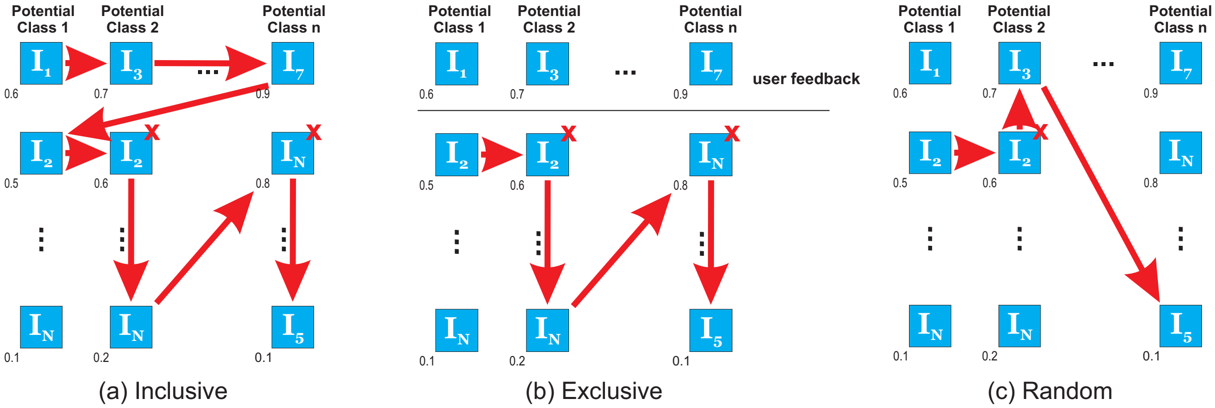
Almost all the existing relevance feedback techniques focus exclusively on improving the relevance of the results. The novelty of our approach is in considering in priority a diversification strategy on top of the classic relevance feedback approach. Experimental validation conducted on a publicly available image retrieval diversification dataset, i.e., Div400 [21], show the benefits of this approach which outperforms other state-of-the-art approaches. The proposed approach is presented in the sequel.

III. Proposed approach

The proposed approach involves a classifier-based relevance feedback and consists of two steps. The first step is an optimized multi-class Support Vector Machine (SVM) classifier-based relevance feedback. The objective is to use user input to categorize the images in a number of distinct classes (i.e., sub-topics). The second step is the actual diversifier and consists of an intra and inter-class image diversification strategy which operates on the SVM class output confidence scores. Several strategies are proposed and evaluated. Each processing step is presented in the following.

A. Multi-class Support Vector Machine relevance feedback

The proposed relevance feedback is a classifier-based feedback approach which works as following: given the results for a certain image retrieval system, the user provides a categorization of the top n ranked results (n is usually a small number) in two classes: relevant vs. non-relevant (for the current query). Then, we use this information as ground truth to train a certain classifier to respond to these two classes. In this classification process, images are represented with content descriptors, i.e., numeric representations of the discriminative underlying image contents.



(a) Inclusive

(b) Exclusive

(c) Random

Fig. 1: The proposed diversification strategies: (a) Inclusive, (b) Exclusive and (c) Random (the small numbers represent some simulated SVM output confidence scores).

Equipped with such tool, we then feed to the freshly trained classifier all the returned images. The classifier will return for this new data some confidence scores which represent the class appurtenance probability. The higher the score, the more likely is that the image belongs to the target class, i.e., relevant images in our case. Using these scores, we then re-rank all the returned images following several strategies which are presented in the next section. The idea is to put in priority the diversification of the relevant results. This represents one relevance feedback iteration. The process can be iterated a number of times until results do not change anymore.

We selected for classification the Support Vector Machines (SVM), which are very well known to perform best in image/multimedia retrieval scenarios. In its basic form, the binary SVM builds a margin that maximizes the distance between two data classes. Several kernel functions can be used to model that margin, from linear to non-linear approximations (Radial Basis Function, Chi-Square, etc). Apart from its general efficiency, this classification scheme provides an important advantage for our specific relevance feedback scenario, i.e., SVM is remarkably intolerant to the number of training examples for the two classes [32], while most learning algorithms tend to correctly classify the class with the larger number of examples. Obviously, for the relevance feedback, the number of positive and negative examples tend to be significantly disproportioned (being recorded in a small result window).

However, our diversification problem is better to be modeled as a multi-class classification problem, where the different classes correspond to the diverse sub-topic representations of the results. We therefore implemented a multi-class SVM classification framework which works as follows. For each target image class (provided by user) we train an individual binary SVM classifier. After training all the SVMs, each classifier will generate a confidence score for each of the output classes. The final fusion of those scores to achieve the multi-class attribution of the images is to be carried out using the diversification strategies presented in the following section (Section III-B).

Finally, to improve even more the classification results, we propose an optimized version of the SVM which consists on optimizing the parameter C that controls the tradeoff between margin maximization and error minimization during the training process. Instead of considering a global value for C, we optimize it for each query in particular. The idea is to divide the relevance feedback training samples in two parts and use one part for training the classifier and the second part to assess its performance. The process is repeated for various values of C until optimal performance is achieved. This process ensures both the optimization of the parameters and the training of the classifier.

B. Diversification strategies

To put diversification in priority, we propose and investigate several diversification strategies. These strategies operate on the SVM output relevance scores for the images. Images are re-ranked by analyzing intra and inter-class relevance scores thus to return in first place the relevant and in the same time diverse representations of the query.

Firstly, for each of the SVM output classes, the images are sorted in descending order according to their output confidence scores. Then, the following diversification strategies are adopted (see Figure 1):

Inclusive: We maintain the number of classes that resulted from the user's feedback and we aim to keep in each class at least one image. Considering the order described above and starting with the first image in the each potential class (i.e., a candidate class for the current image), images are visited to determine to which class they should be assigned to. Each image is first checked to see if it was previously visited. If not, the image is assigned to the current potential class which becomes the final class and returned as a relevant and diverse result. If it was previously considered, then the first unvisited image in the current potential class, according to the confidence score, is assigned to this class and further returned as the next relevant and diverse result. The algorithm repeats until the required number of images is reached (see Figure 1.a);

Exclusive: This approach doesn't take into account the images provided via the user's feedback. In addition to the inclusive strategy described above, the images are extra checked to see if they were used in the training process of the SVM as

Fig. 2: Div400 [21] location picture examples (photo credits from Flickr, from left to right and top to bottom: Andwar, Ipoh kia, Marvin (PA), photoAtlas, Julie Duquesne, Jack Zalium and kniemla).

user input. If so, another image in the current potential class, which wasn't earlier assigned to another class, is searched and returned as a relevant and diverse result. The process is repeated until the required number of images is reached (see Figure 1.b);

Random: This is based on selecting images randomly from the ordered list described above, according to a pseudo-random number generator. The same principle as in the inclusive strategy is considered when the image is found to be already selected, the algorithm searches for the next unvisited image of the current potential class and the one indicated by a new randomly generated number is selected. This means that there is at least one number generated for each class (see Figure 1.c).

IV. Experimental Setup

In this section we detail the evaluation framework for the proposed relevance feedback techniques.

A. Data

For conducting the experiments, we selected a publicly available image retrieval diversification dataset, namely Div400 [21], that was validated within the 2013 MediaEval benchmark [25][24]. This dataset is built around a photo with landmark locations retrieval scenario. It provides for 396 locations up to 150 photos and associated metadata retrieved from Flickr[1] and ranked with Flickr's default "relevance" algorithm. Locations are diverse (e.g., museums, archeological sites, cathedrals, roads, bridges, etc) and spread over 34 countries around the world. An example is presented in Figure 2. Data is collected from Flickr with both textual (i.e., location name) and GPS queries. Provided location metadata consists of Wikipedia links to location webpages and GPS information and photo metadata includes social data, e.g., author title and description, user tags, geotagging information, time/date of the photo, owner's name, the number of times the photo has been displayed, number of posted comments, rank, etc.

Data are annotated for both relevance and diversity of the photos using the following definitions: *relevance* — a photo is relevant if it is a common photo representation of the location, e.g., different views at different times of the day/year and under different weather conditions, inside views, creative views, etc, which contain partially or entirely the target location (bad quality photos are considered irrelevant) — photos are tagged as relevant, non-relevant or with "don't know"; *diversity* — a set of photos is considered to be diverse if it depicts complementary visual characteristics of the target location (e.g., most of the perceived visual information is different) — relevant photos are clustered into visually similar groups. Annotations were determined mainly by experts with advanced knowledge of location characteristics and are provided with the dataset.

Div400 is divided into a development set containing 50 locations (5,118 photos, in average 102.4/location) that is intended to be used for designing and validating the approaches and a test set containing 346 locations (containing 38,300 photos, in average 110.7/location) for the actual evaluation. In consequence, we conducted all the experimentations on the test set.

B. Testing

To test the diversification approaches, we use the same scenario and evaluation conditions as in the 2013 MediaEval benchmark [25][24]. Given the dataset above, the proposed approaches should be able to refine (for each of the locations) the initial Flickr retrieval results by selecting a ranked list of up to 50 photos that are equally relevant and diverse representations of the query (according to the previous definitions).

For the relevance feedback approaches, we consider the scenario where user feedback is automatically simulated with the known class membership of each photo retrieved from the ground truth. This approach allows a fast and extensive simulation which is necessary to evaluate different methods and parameter settings, otherwise impossible with realtime user studies. Such simulations represent a common practice in evaluating relevance feedback scenarios [12][14][20]. Although this is not a real live user feedback experience and some of its constraints may be neglected (e.g., user fatigue, the influence of inter-user agreement), previous experiments from the literature show that results are very close, given the fact that the ground truth was collected in a similar way from real users.

Relevance feedback is recorded in a limited result window. We use a common setting which consists of considering only the first 20 retrieved images. In practice, this provides a good compromise between relevance feedback's efficiency and the users' fatigue.

C. Metrics

To assess performance for both diversity and relevance, we compute the following standard metrics. Diversity of the results is assessed with cluster recall at X ($CR@X$) [22], defined as:

$$CR@X = \frac{N}{N_{gt}} \tag{1}$$

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 20 clusters from the dataset). Defined this way,

[1] http://www.flickr.com/services/api/

$CR@X$ 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 get a clearer view of relevance, precision at X ($P@X$) is also computed:

$$P@X = \frac{N_r}{X} \tag{2}$$

where N_r is the number of relevant pictures from the first X ranked results. Therefore, $P@X$ 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 $P@X$:

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

Evaluation is conducted for different cutoff points, X∈{10,20,30,40,50}. Results are reported as overall average values over all the locations in the test dataset.

V. Experimental results

To assess the performance of the proposed relevance feedback scenarios we conducted several experiments which are presented in following.

A. Content descriptors

One of the key parameters of the proposed relevance feedback scheme is the image content representation, i.e., the employed content descriptors. Although the proposed approach is not dependent on a certain content description scheme, the choice of the descriptors influence significantly the results and should be adapted to the specificity of the data (optimized).

Our first experiment consists on assessing the performance of several descriptors with the objective of determining the best descriptor combination. 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 on image retrieval tasks, namely: 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” [26]; global Histogram of Oriented Gradients (HoG, 81 values) — represents the HoG feature computed on 3 by 3 image regions [27]; global color moments computed on the HSV Color Space (CM, 9 values) — represent the first three central moments of an image color distribution: mean, standard deviation and skewness [28]; global Locally Binary Patterns computed on gray scale representation of the image (LBP, 16 values) [29]; global Color Structure Descriptor (CSD, 64 values) — represents the MPEG-7 Color Structure Descriptor computed on the HMMD color space [30]; 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 [31]; and spatial pyramid representations of these (denoted 3x3) — each of the previous descriptors is computed also locally; the image is divided into 3 by 3 non-overlapping blocks and descriptors are computed on each patch; the global descriptor is obtained by the concatenation of all values. Apart from the individual descriptors, we test also the early fusion of all the descriptors, i.e., normalization and concatenation of all values.

TABLE I: Relevance feedback results for various descriptors (binary SVM, RBF kernel, one relevance feedback session; best results are represented in bold).

descriptor	P@10	P@20	P@30	CR@10	CR@20	CR@30
all together	**0.8681**	**0.8268**	**0.7875**	**0.4388**	**0.6543**	**0.7807**
CN 3x3	0.8474	0.8095	0.7778	0.4106	0.6169	0.7395
LBP 3x3	0.8082	0.7762	0.7519	0.4085	0.6163	0.7498
GLRLM 3x3	0.7833	0.7522	0.7346	0.4043	0.6091	0.7362
CM 3x3	0.812	0.7776	0.7541	0.3892	0.5854	0.7184
HoG	0.7912	0.7607	0.7415	0.3969	0.5872	0.7091
CSD	0.798	0.7721	0.7486	0.3859	0.5839	0.7196
CM	0.7529	0.7358	0.7262	0.3804	0.563	0.6839
CN	0.7746	0.7488	0.7354	0.373	0.5575	0.6908
GLRLM	0.7544	0.732	0.7198	0.3696	0.5664	0.6944
LBP	0.7582	0.7374	0.7247	0.3668	0.562	0.6917

For this preliminary test, we have selected as baseline the use of the binary SVM with a Radial Basis Function (RBF) kernel. No diversification scheme is employed. However, to account for diversity, the user feedback is simulated with the diversity ground truth, i.e., the selected images are visually diverse sub-topic examples. We use only one relevance feedback session. Results are presented in Table I.

One can observe that the best precision and cluster recall is achieved when combining all the descriptors together, e.g., $P@10$ is 86.81% which is an improvement of 2% over the closest individual descriptor score, $CR@10$ is 43.88% with an improvement of almost 3%. The improvement is consistent also when increasing the number of images, at 20, 30 and so on (for brevity reasons we provided the results up to 30 images). An interesting result is the fact that computing also the spatial pyramid representations of the descriptors provide better results compared to the global versions (see the descriptors marked with 3x3). Based on this findings, the remaining experiments are to be conducted using all the descriptors, which gives the best performance.

B. Evaluation of the proposed diversification strategies

The next experiment consists on evaluating the proposed diversification strategies (see Section III-B). For all the strategies, user feedback was simulated using diversity ground truth from the dataset. We use a linear kernel and only one relevance feedback iteration. Results for different cutoff points are synthesized in Figure 3. Results reveal in the first place the fact that the precision tends to decrease with the higher precision cut-offs. This is motivated by the fact that increasing the number of results also increases the probability of including non-relevant pictures as in general the best matches tend to accumulate among the first returned results. Second, in contrast to the precision, cluster recall and thus diversity, increases

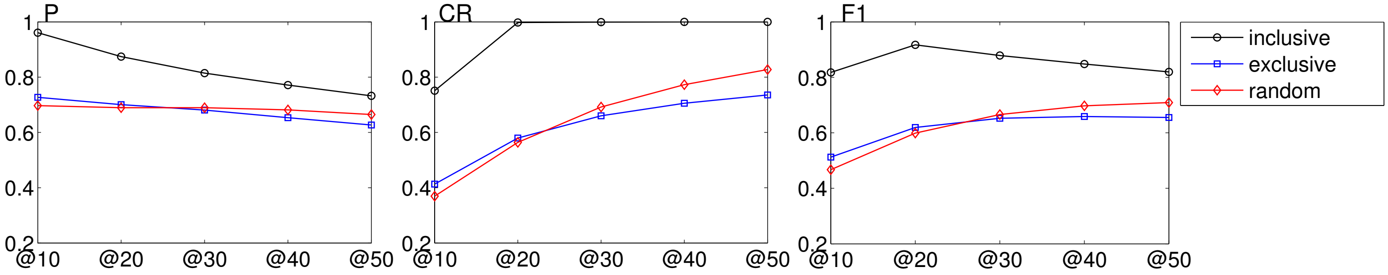


Fig. 3: Precision, cluster recall and F1 scores for the proposed SVM multi-class relevance feedback diversification strategies (inclusive, exclusive and random — see Section III-B). Results are obtained with a linear kernel for one relevance feedback iteration.

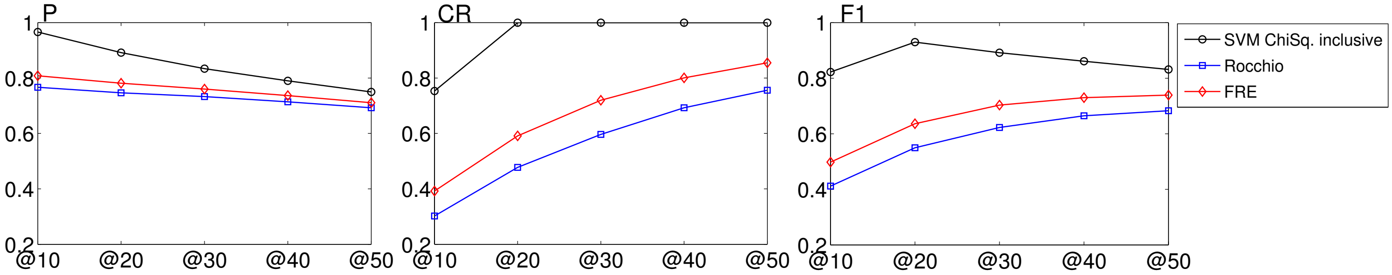


Fig. 4: Comparison of the proposed SVM multi-class relevance feedback with the inclusive diversification strategy (Chi-Square kernel, see Section III-B) with Rocchio and FRE relevance feedback. Results are depicted for one relevance feedback iteration.

TABLE II: Various kernel SVM relevance feedback with inclusive diversification strategy (best results are in bold).

SVM kernel	P@10	P@20	P@30	CR@10	CR@20	CR@30
linear	0.9608	0.8743	0.8145	0.751	0.9977	0.9992
RBF	0.9658	0.8906	0.8327	**0.753**	**1**	**1**
Chi-Square	**0.9661**	**0.8918**	**0.8339**	**0.753**	**1**	**1**

with the number of pictures. This result is intuitive as the more pictures we retrieve, the more likely is to include a representative picture from each of the annotated categories.

Strategy-wise, the best diversification approach is by far the inclusive strategy which reaches 100% diversification when assessing more than 30 images (in average, the dataset provides 13 different sub-topics per location). The other two approaches provide more or less similar results for both precision and cluster recall. Surprisingly, the random selection strategy seems to provide better results than the exclusive selection. A possible explanation for this may be the fact that uniformly shuffling the images within the classes tend to produce diversified results.

For the best strategy, we tested also the influence of the choice of the kernel function. We experimented with linear, RBF and Chi-Square kernels. Results are presented in Table II (we present the results up to 30 images). Given the already competitive results, the choice of the kernel function influences very little the performance. However, the use of non-linear kernels such as RBF and Chi-Square seems better adapted to our task and allows for increasing the precision with several percents for some of the cutoff points.

C. Comparison with other relevance feedback strategies

In this experiment, we compare the performance of the proposed relevance feedback using the inclusive diversification strategy (which gave the best results) with other relevance feedback approaches from the literature. We selected the Rocchio's algorithm [19] and the Relevance Feature Estimation (RFE) algorithm [12] (see Section II). All the user feedback is simulated using the diversity ground truth from the dataset. Results are presented in Figure 4.

One may observe that the proposed strategy is still more efficient in contrast with other relevance feedback approaches. This proves again that the proposed approach is better adapted to the diversification task than classical relevance feedback techniques which tend to prioritize the relevance of the results.

D. Comparison with state-of-the-art diversification techniques

A final comparison of the results is conducted in the context of the current state-of-the-art search result diversification literature. For comparison, we have selected the three best performing approaches proposed at the 2013 MediaEval Retrieving Diverse Social Images benchmark [24] (teams SOTON-WAIS, SocSens and CEA), as well as the three approaches presented in [33] which are highly relevant for our diversification strategy. In addition to these, we report also to the initial retrieval results provided by Flickr which are ranked with Flickr's default "relevance" algorithm. These methods are automatic, in the sense that the entire diversification process is computed by a machine.

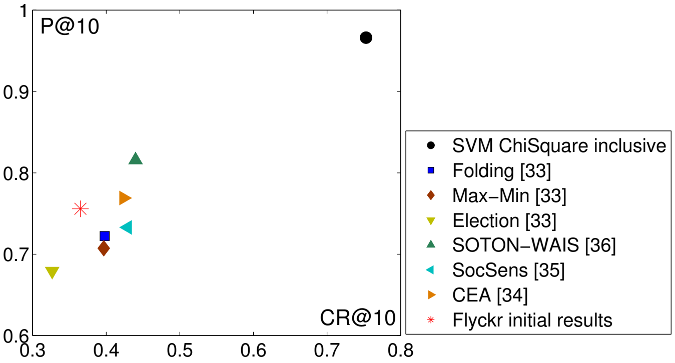
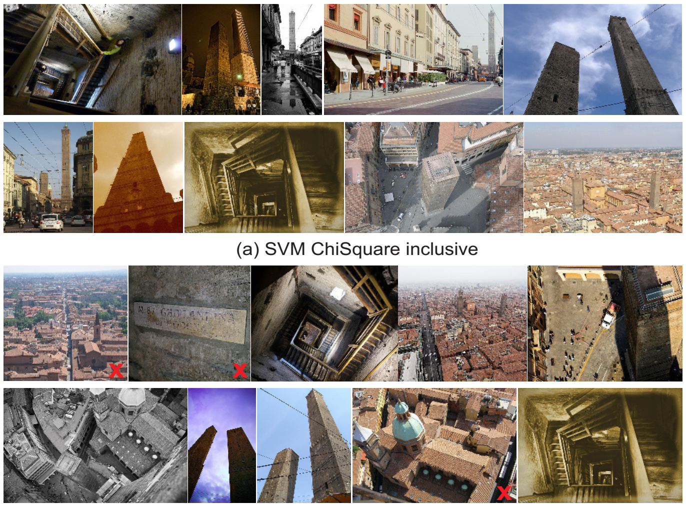


Fig. 5: Precision vs. cluster recall at 10 images.

The method proposed by SOTON-WAIS run3 [36] uses a re-ranking with a proximity search to improve precision followed by a Greedy Min-Max diversifier. Images are represented with both text and visual descriptors. SocSens run1 [35] involves a Greedy optimization of a utility function that weights both relevance and diversity scores. Images are represented with visual descriptors. CEA run2 [34] diversifies the query results by considering the images from different users or that were taken by the same user on different days. Images are represented with text descriptors. The approach in [33] considers three diversification scenarios: 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 subtopic 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. Images are represented with visual descriptors. All the image descriptions (text and visual) use the representations proposed in Div400 [21] which are also adopted in this paper (see Section V-A).

What is interesting to see is whether (and to what extent) the use of a hybrid approach that involves human judgments on the diversity of the results allows for an improvement of the results over the automated methods. In Figure 5 we plot the precision vs. cluster recall for the above systems. For comparison purpose, results are reported for a cutoff at 10 images which was the official metrics of the Retrieving Diverse Social Images benchmark (submitted systems were optimized with respect to this cutoff point). All the methods (except for the Election) provide an improvement of the diversity compared to the Flickr baseline (initial retrieval results). However, the automated methods tend to provide a limited diversification improvement, at most 7% over the baseline, while some of them sacrifice the precision of the results, e.g., SocSens's approach or Folding and Max-Min which lead to lower precision than the Flickr's initial results. The use of the human feedback, and specifically not only for relevance but for the diversification of the results, significantly boost the performance leading to an improvement of at least 33% over the best automated method while keeping the precision of the results at a very high rate.



(a) SVM ChiSquare inclusive

(b) Flickr initial results

Fig. 6: Visual comparison of the results for "Asinelli Tower" (Italy). Flickr image credits (from left to right and top to bottom): (a) kondrag, sdhaddow, Alessandro Capotondi, roy.luck, Argenberg, greenblackberries, Sim Dawdler, Funchye, Argenberg, magro_kr, (b) lorkatj, leonardo4it, kyle NRW, Viaggiatore Fantasma, kyle NRW, sara zollino, Alessandro Capotondi, magro_kr (2 images), Funchye. Un-relevant images are marked with a red X. Only the first 10 ranks are displayed.

E. Visual example

Finally, we present a visual example of how the diversified results look after the use of the proposed relevance feedback. We selected one of the locations, namely the "Asinelli Tower" in Italy. In Figure 6 we illustrate for comparison the images provided by the initial Flickr results (as baseline) and the best proposed diversification strategy, namely SVM with Chi-Square kernel and inclusive diversification. One may observe that the diversification is first of all capable of reducing the irrelevant images — this is also visible from the precision value — but as well as to diversify the results by providing very different perspectives.

VI. Conclusions and Future Work

In this paper we addressed the issue of image search result diversification from the perspective of including the human expertise in the computational process. We discussed a relevance feedback approach that relies on an locally optimized multi-class SVM classification-based approach. Diversification of the results is achieved by employing a dedicated strategy that exploits the classifiers' output confidence scores for diversifying the selection of images. Experimental validation was carried out on a dedicated dataset that proposes a 346 geographical locations scenario with more than 38,300 Flickr photo search results to diversify. Several experimental tests were carried out: comparison of different diversification strategies, comparison with other relevance feedback approaches from the literature and comparison with state-of-the-art image search diversification techniques from the 2013 MediaEval Retrieving Diverse Social Images benchmark and literature. Results show the true benefits of including the human in the loop which allows for a great deal of improvement over the automated methods or relevance oriented traditional relevance feedback approaches. However, on the downside, relevance feedback still depends on the presence of humans. To compensate this, future work will consist of investigating the benefits of the pseudo-relevance feedback approaches that substitute the human input by assuming the top ranked images as relevant.

Acknowledgment

The authors would like to thank the 2013 Retrieving Diverse Social Images Task, MediaEval Benchmarking Initiative for Multimedia Evaluation [25], for providing the dataset. Part of this work was supported under ExcelDOC POSDRU/159/1.5/S/132397 (2014-2015).

References

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, “Content-based image retrieval at the end of the early years”, IEEE Transaction on Pattern Analysis and Machine Intelligence, 22(12), pp. 1349 - 1380, 2000.

[2] R. Datta, D. Joshi, J. Li, J.Z. Wang, “Image Retrieval: Ideas, Influences, and Trends of the New Age”, ACM Comput. Surv., 40(2), pp. 1-60, 2008.

[3] R. Priyatharshini, S. Chitrakala, “Association Based Image Retrieval: A Survey”, Mobile Communication and Power Engineering, Springer Communications in Computer and Information Science, 296, pp 17-26, 2013.

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

[5] A.-L. Radu, B. Ionescu, M. Menéndez, J. Stöttinger, F. Giunchiglia, A. De Angeli, “A Hybrid Machine-Crowd Approach to Photo Retrieval Result Diversification”, Multimedia Modeling, Ireland, LNCS 8325, pp. 25-36, 2014.

[6] Dang, Van and Croft, W. Bruce, “Diversity by Proportionality: An Election-based Approach to Search Result Diversification”, ACM International Conference on Research and Development in Information Retrieval, pp. 65-74, Portland, Oregon, USA, 2012.

[7] M.R. Vieira, H.L. Razente, M.C.N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. Traina Jr., V.J. Tsotras, “On Query Result Diversification”, IEEE International Conference on Data Engineering, pp. 1163 - 1174, 11-16 April, Hannover, Germany, 2011.

[8] R.H. van Leuken, L. Garcia, X. Olivares, R. van Zwol, “Visual Diversification of Image Search Results”, ACM World Wide Web, pp. 341-350, 2009.

[9] T. Deselaers, T. Gass, P. Dreuw, H. Ney, “Jointly Optimising Relevance and Diversity in Image Retrieval”, ACM Int. Conf. on Image and Video Retrieval, 2009.

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

[11] M. Soleymani, M. Larson, “Crowd-sourcing for affective annotation of video: Develop ment of a viewer-reported boredom corpus”, SIGIR Workshop on Crowd-sourcing for Search Evaluation, 2010.

[12] Y. Rui, T. S. Huang, M. Ortega, M. Mehrotra, S. Beckman: “Relevance feedback: a power tool for interactive content-based image retrieval”, IEEE Trans. on Circuits and Video Technology, pp. 644-655, 1998.

[13] N. V. Nguyen, J.-M. Ogier, S. Tabbone, A. Boucher: “Text Retrieval Relevance Feedback Techniques for Bag-of-Words Model in CBIR”, International Conference on Machine Learning and Pattern Recognition, 2009.

[14] S. Liang, Z. Sun, “Sketch retrieval and relevance feedback with biased SVM classification”, Pattern Recognition Letters, 29, pp. 1733-1741, 2008.

[15] G. Giacinto, “A Nearest-Neighbor Approach to Relevance Feedback in Content-Based Image Retrieval”, ACM Int. Conf. on Image and Video Retrieval, 2007.

[16] Y. Wu, A. Zhang, “Interactive pattern analysis for relevance feedback in multimedia information retrieval”, Multimedia Systems, 10(1), pp. 41-55, 2004.

[17] A. F. Smeaton, P. Over, W. Kraaij, “High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements”, Multimedia Content Analysis Theory and Applications, pp. 151-174, 2009.

[18] B. Ionescu, K. Seyerlehner, I. Mironica, C. Vertan, P. Lambert, ”An Audio-Visual Approach to Web Video Categorization”, MTAP, 2012.

[19] J. Rocchio: ”Relevance Feedback in Information Retrieval”, The Smart Retrieval System Experiments in Automatic Document Processing, G. Salton, 1971.

[20] J. Yu, Y. Lu, Y. Xu, N. Sebe, Q. Tian: ”Integrating Relevance Feedback in Boosting for Content-Based Image Retrieval”, ICASSP, pp. 965-968, 2007.

[21] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, *Div400: A Social Image Retrieval Result Diversification Dataset*, ACM Multimedia Systems - MMSys2014, 19-21 March, Singapore, 2014.

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

[23] A. Popescu, G. Grefenstette, “Social Media Driven Image Retrieval”, ACM ICMR, April 17-20, Trento, Italy, 2011.

[24] B. Ionescu, M. Menéndez, H. Müller, A. Popescu, “Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation”, MediaEval Benchmarking Initiative for Multimedia Evaluation, vol. 1043, CEUR-WS.org, ISSN: 1613-0073, October 18-19, Barcelona, Spain, 2013.

[25] MediaEval 2013 Workshop, Eds. M. Larson, X. Anguera, T. Reuter, G.J.F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, M. Soleymani, co-located with ACM Multimedia, Barcelona, Spain, October 18-19, CEUR-WS.org, ISSN 1613-0073, Vol. 1043, http://ceur-ws.org/Vol-1043/, 2013.

[26] Weijer, Van de, Schmid, C., Verbeek, J., Larlus, D. “Learning color names for real-world applications.”, IEEE Trans. on Image Processing, 18(7), pp. 1512-1523, 2009.

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

[28] Stricker, M., Orengo, M. “Similarity of color images”, SPIE Conference on Storage and Retrieval for Image and Video Databases III, vol. 2420, 1995, 381 - 392.

[29] Ojala, T., Pietikinen, M., Harwood, D. “Performance evaluation of texture measures with classification based on Kullback discrimination of distributions”, IAPR International Conference on Pattern Recognition, vol. 1, 1994, 582 - 585.

[30] Manjunath, B. S., Ohm, J. R., Vasudevan, V. V., Yamada, A. “Color and texture descriptors”, IEEE Trans. on Circuits and Systems for Video Technology, vol. 11(6), 2001, 703 - 715.

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

[32] H. Drucker, B. Shahrary,D. C. Gibbon, “Relevance feedback using support vector machines”, In International Conference on Machine Learning, pp. 122-129, 2001.

[33] R. H. van Leuken, L. Garcia, X. Olivares, R. van Zwol, “Visual Diversification of Image Search Results”, ACM WWW World Wide Web Conference. ACM Madrid, Spain, 2009, 341-350.

[34] A. Popescu, “CEA LISTs Participation at the MediaEval 2013 Retrieving Diverse Social Images Task”, Working Notes Proceedings [25], http://ceur-ws.org/Vol-1043/mediaeval2013_submission_43.pdf, 2013.

[35] D. Corney, C. Martin, A. Göker, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, L. Aiello, B. Thomee, “SocialSensor: Finding Diverse Images at MediaEval 2013”, Working Notes Proceedings [25], http://ceur-ws.org/Vol-1043/mediaeval2013_submission_24.pdf, 2013.

[36] N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, P. Lewis, “Experiments in Diversifying Flickr Result Sets”, Working Notes Proceedings [25], http://ceur-ws.org/Vol-1043/mediaeval2013_submission_18.pdf, 2013.