Improving Diversity in Image Search via Supervised Relevance Scoring

Eleftherios Spyromitros-Xioufis
CERTH-ITI / AUTH, Greece
espyromi@iti.gr

Symeon Papadopoulos
CERTH-ITI, Greece
papadop@iti.gr

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

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

Yiannis Kompatsiaris
CERTH-ITI, Greece
ikom@iti.gr

Ioannis Vlahavas
AUTH, Greece
vlahavas@csd.auth.gr

ABSTRACT

Results returned by commercial image search engines should include relevant and diversified depictions of queries in order to ensure good coverage of users' information needs. While relevance has drastically improved in recent years, diversity is still an open problem. In this paper we propose a reranking method that could be implemented on top of such engines in order to provide a better balance between relevance and diversity. Our method formulates the reranking problem as an optimization of a utility function that jointly considers relevance and diversity. Our main contribution is the replacement of the unsupervised definition of relevance that is commonly used in this formulation with a supervised classification model that strives to capture a query and application-specific notion of relevance. This model provides more accurate relevance scores that lead to significantly improved diversification performance. Furthermore, we propose a stacking-type ensemble learning approach that allows combining multiple features in a principled way when computing the relevance of an image. An empirical evaluation carried out on the datasets of the MediaEval 2013 and 2014 "Retrieving Diverse Social Images" (RDSI) benchmarks confirms the superior performance of the proposed method compared to other participating systems as well as a state-of-the-art, unsupervised reranking method.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback*

Keywords

image retrieval, diversity, relevance feedback, image classification, image reranking

ICMR'15, June 23–26, 2015, Shanghai, China.

ACM 978-1-4503-3274-3/15/06 ...$15.00.
http://dx.doi.org/10.1145/2671188.2749334.

1. INTRODUCTION

The importance of presenting a set of results that are at the same time relevant to the query but also exhibit diversity has been pointed out long ago in the Information Retrieval (IR) community [3]. Diversity in top results provides a more comprehensive and concise answer to the query which in turn enables faster access to the desired information and ultimately results in increased user satisfaction. Despite this fact, existing image search engines (either operating on web scale, e.g., Google Images, or within media sharing platforms, e.g., Flickr) still focus primarily on relevance. As a consequence, top results usually contain many similar images and the user has to go deeper down the list of results in order to discover diverse views of the query. In addition, the deeper one goes down the list, the higher the probability to encounter irrelevant results becomes, thus impeding the discovery of diverse views. This focus on relevance is perhaps due to the limitations imposed by relying mostly on the textual modality of the images [15] (e.g., surrounding web page text in the case of Google Images, tags and textual descriptions in the case of Flickr). Obviously, ignoring the visual content of the images limits the ability of a search engine to provide either relevant or diverse results. Figure 1 shows the first 10 results returned by Flickr (top) and Google images (middle) in response to a query about "La Madeleine" church in Paris. Both result sets are not optimal in the sense that they contain irrelevant and/or similar images.

In this paper, we propose a method that aims at refining the initial ranking of existing image search engines so that both the relevance and the diversity of the top results are improved, as shown on Figure 1 (bottom). Our method builds upon a diversification method [7] that casts the above problem into the optimization of a utility function that jointly considers relevance and diversity. In contrast to [7] that used a generic, unsupervised definition for relevance, we propose a query and application-specific definition that is directly learned from user feedback. This supervised definition alleviates several shortcomings of the unsupervised counterpart and manages to provide a significantly more accurate relevance scoring which, as we show, has a direct positive impact on the overall diversification performance of the algorithm. Although our method assumes the existence of relevance annotations, we explain that such annotations do not necessarily need to be given by experts but can also be acquired via implicit user feedback. Another important property of the proposed diversification approach is that it allows the com-

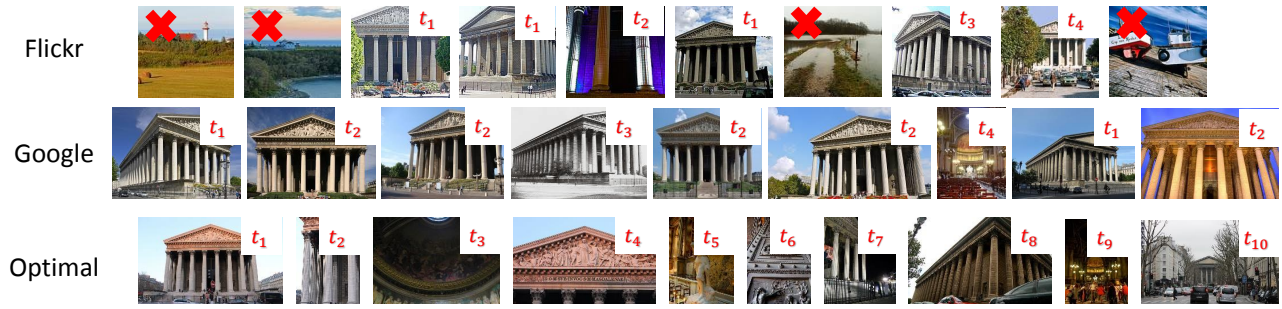

Figure 1: The first 10 results returned by Flickr (top) and Google (middle) in response to a query about "La Madeleine" church in Paris. The bottom row corresponds to an optimal result set where each image captures a different aspect of the query. t_i denotes the topic that an image belongs to.

bination of multiple types of features in a principled and effective way when computing the relevance of an image.

Our approach can be applied to different retrieval scenarios. However, in this paper we focus on landmark retrieval mainly due to the fact that a social image retrieval benchmark is available for this scenario (RDSI benchmark of MediaEval 2013 [9] and 2014 [10]). Specifically, we consider the case of a tourist who wants to plan a trip to an unfamiliar location. Knowing only the name of the landmark, the person uses it to learn more about it by visiting the corresponding Wikipedia article where he/she can get a textual description and some example photos. Before deciding whether this landmark is worth visiting, the tourist wants to get an informative visual summary of it. A query using existing search engines tends to return many similar images or holiday photos that prominently feature people as the main subject. Other results may include low quality images or images of completely different locations. Ideally, the person would like to receive photos capturing different visual characteristics of the target location, e.g., different viewpoints, architectural details, creative views, etc., so that most of the perceived visual information is different from one photo to another as in Figure 1 (bottom).

Using an early version of our method (briefly described in [5] and [14]), we achieved top results in the RDSI benchmarks of MediaEval 2013 and 2014. In this paper, we develop the method further and provide an extensive analysis of its performance. In particular, this paper makes the following contributions:

- We highlight the connection between the accuracy of the relevance scoring component of the diversification method of [7] and its diversification performance.
- We show that relevance scoring accuracy can be significantly improved by exploiting relevance annotations and performing the scoring in a supervised manner.
- We evaluate supervised relevance scoring models built from training sets of various compositions and show that a small number of query-specific positive annotations combined with application-specific positive and negative annotations yield very good results.
- We develop a novel ensemble learning algorithm that combines multi-dimensional and one-dimensional features in a principled and effective way.

As a result, our method achieves significantly better diversification performance than a state-of-the-art reranking method that uses an unsupervised definition of relevance and outperforms the best performing systems of the 2013 and 2014 RDSI benchmarks by 7.5% and 5.7% respectively.

2. RELATED WORK

One of the first and seminal works on diversity in information retrieval is the work of Carbonell and Goldstain [3]. Recognizing that in the context of text retrieval and summarization, pure relevance ranking is not sufficient, they proposed Maximal Marginal Relevance (MMR). MMR is a reranking method that linearly combines independent measurements of relevance and diversity (their relative weight is a user-tunable parameter) into a single metric that is maximized in a greedy, iterative fashion. In [17], MMR was combined with a language modelling framework to solve the problem of subtopic retrieval where the task is to find documents that cover as many different subtopics of a query topic as possible. In [4], a greedy algorithm is developed for maximizing the probability of finding at least one relevant document in top n. This algorithm is similar to MMR and is also shown to be a greedy maximizer of the instance or cluster recall metric, i.e. the number of different subtopics covered by a given set of results. More recently, a similar formulation of the diversification problem was given by Deselaers et al. [7] and was found to outperform a common clustering-based diversification approach, in the context of diverse image retrieval. As in MMR, diversification is achieved via the optimization of a criterion that linearly combines relevance and diversity. However, [7] gives a more general formulation and uses dynamic programming algorithms to perform the optimization in addition to the greedy, iterative algorithm presented in [3]. In our work, we adopt the formulation of [7] but combine it with a supervised definition of relevance that leads to significantly better performance. Also, compared to [7], where different modalities were combined in an ad-hoc way, the use of learning allows us to develop a more principled and effective way of combining multiple features.

Diversity in social image retrieval was the focus of the MediaEval 2013 and 2014 RDSI benchmarks that attracted the interest of many groups working in this area. Most participants developed diversification approaches that combined clustering with a strategy to select and return representative images from each cluster. Our MMR-based approach has the advantage of targeting the diversification problem

in a more straightforward way compared to clustering-based approaches which first try to solve a different and presumably more difficult problem (i.e. finding groups of similar images). Also, despite the fact that most systems involved a mechanism to improve relevance before enforcing diversity, the majority did not exploit relevance annotations. Instead, top-performing solutions ([11], [6]) used specialized filters (e.g., face and blur detectors) and hand-coded rules (distant images are irrelevant) in order to discard irrelevant images according to the verbal definitions of relevance and irrelevance given by the task organizers. By learning the concept of relevance through the use of query and application-specific relevance annotations, our method can adapt automatically to different queries and retrieval scenarios and thus represents a more general solution.

3. METHOD

3.1 Problem Definition

Let q be a query[1] and $I = \{im_1, \ldots, im_N\}$ be a ranked list of images that have been retrieved by an existing search engine in response to q. Although the quality of the results depends on the specific query and search engine, we expect that for values of N up to a few hundreds, I will typically comprise both relevant and irrelevant images and that some of the relevant images might contain duplicate information[2] (see Figure 1). The goal of a diversification method, is to refine the initial ranking of the images in I so that relevant images are ranked higher than irrelevant and top positions contain as little duplicate images as possible. Since users usually inspect only the top few results, reranking the whole list is not needed and we instead request a K-sized subset of images from I that are as relevant (to the query) and as diverse (with each other) as possible. Among the many measures that have been proposed in order to quantify the above qualitative goal of a diversification method is, for instance, the subtopic or cluster recall at K (CR@K) [17] that measures the percentage of different topics/aspects retrieved in the first K results. Note that a perfect CR@K requires all K results to be relevant. A typical cut-off is $K = 20$ (adopted in MediaEval 2014 [10] and ImageCLEF 2008 [1]), as most search engines present around 20 results on their first page.

3.2 Maximal Marginal Relevance

MMR formulates the goal of a diversification method as an optimization problem where one tries to maximize a linear combination of relevance and diversity, the so called "marginal relevance". According to the formulation given in [7], the objective is to find the K-sized set $S \subset I$ that maximizes the following utility function:

$$\arg\max_{S \subset I, |S|=K} U(S|q) = w * R(S|q) + (1-w) * D(S), \quad (1)$$

where $R(S|q)$ is a measure of the relevance of S to the query, $D(S)$ is a measure of the diversity in S, and w is a parameter that controls the relative importance of relevance and diversity. w can be either adjusted by the user or tuned (on a diversity annotated validation set) to optimize a particular quantitative measure (e.g., CR@20). Note that as implied by Equation 1, diversity is independent of the query.

[1]Note that q can be expressed as a textual, visual or mixed query without loss of generality.

[2]These assumptions are shown to hold (Subsection 4.1) when Flickr is used as search engine for landmark queries.

Figure 2: Wikipedia image of "Angkor Wat" (left), a relevant inside view (center) and an irrelevant image with a person in front of the monument (right).

3.3 Relevance

3.3.1 *Unsupervised Relevance*

The relevance of a set of images S was defined in [7] as

$$R(S|q) = \sum_{im_i \in S} R(im_i|q) = \sum_{im_i \in S} s(im_i, im_q), \quad (2)$$

where $R(im_i|q)$ denotes the relevance of each individual image to the query and $s(im_i, im_q)$ is a normalized similarity measure between im_i and im_q. Depending on the type of representation chosen for the query and the database images, several different similarity functions may be used. For instance, when both the query and the database images are represented as text, $s(im_i, im_q)$ could be the cosine similarity between bag of word vectors, while in cases where the query is expressed via an example image, $s(im_i, im_q)$ could be the (normalized) inverse of the Euclidean distance between the respective visual feature representations.

A limitation of the unsupervised definition of relevance described above is that similarity, as quantified by common textual or visual representation-measure combinations, does not attempt to capture the concept of relevance as conceived by users. In addition, similarity does not imply relevance and vice versa. Consider, for instance, the landmark retrieval use case described in Section 1. As exemplified in Figure 1 the optimal set of results might contain images that are dissimilar to a reference, visual representation of the query (e.g., the Wikipedia image) but are still considered relevant to it e.g., inside views, architectural details, etc. On the other hand, images that are visually similar to the query might be considered irrelevant due to people being the main focus of the image (Figure 2). The situation is even more problematic when a textual representation is used because textual similarity is often a poor proxy for visual similarity (noisy or missing tags/descriptions). Although the use of multiple modalities is expected to improve the accuracy of relevance scoring, their combination is usually performed with arbitrarily chosen weights [7].

3.3.2 *Learning Relevance from User Feedback*

Motivated by the shortcomings of unsupervised relevance scoring described above and by the fact that improving the relevance scoring quality within the MMR framework has a direct positive impact on common diversification performance measures (Subsection 5.1), we propose a supervised relevance scoring method that exploits relevance annotations in order to induce a more accurate definition of relevance.

More specifically, for each query q, we build a probabilistic model $h_q : X \rightarrow [0, 1]$ that takes a n-dimensional representation of the image $X = R^n$ as input and outputs the probability that the image is relevant to the query given the input. These probabilistic outputs ($h_q(im_i)$) replace the unsupervised similarity measurements ($s(im_i, im_q)$) in Equation 2. In order to train this model, we assume the existence of a set $D_q = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of m training examples where $x_i \in X$ is the input vector and $y_i \in Y = \{0, 1\}$ is the class value with $y_i = 1/0$ denoting a relevant/irrelevant example.

Ideally, D_q will be composed of examples that have explicitly been annotated as relevant/irrelevant to q by the user and will therefore constitute an accurate representation of what the user considers relevant. This, however, requires considerable effort from the user. A more realistic scenario is the collection of relevance annotations by exploiting implicit relevance feedback given by multiple users for the same query q. A search engine can count the number of times that each image $im_i \in I$ is clicked (viewed) by users when returned in response to a query q. Then, we can reasonably assume that images with high click counts are relevant to the query and images with lower click counts are irrelevant. In fact, even a single click (or few to exclude noise) will be sufficient to claim relevance of an image. On the other hand, multiple feedback signals will be required to confidently claim irrelevance of an image due to the implicit and partial nature of the feedback: a) an image might be relevant but contain duplicate information with another image and therefore not clicked or b) an image far down the list might have not been inspected at all. Therefore, exploiting implicit user feedback as described above could result in a set of query-specific positive (relevant) examples and, in case of popular queries, negative (irrelevant) examples.

In this paper, we focus on infrequent queries for which negative (irrelevant) examples are unavailable. To train h_q in this case, we combine few query-specific positive examples with positive and negative examples from other (popular) queries of similar type (e.g., landmark queries). Despite being a seemingly counter-intuitive choice, the inclusion of positive and negative examples from other queries of the same type, increases the generalization ability of the classifier (Subsection 5.2) because it helps capture an application specific notion of relevance/irrelevance (e.g., out-of-focus or human-in-focus images are irrelevant and drawings are relevant in the RDSI task scenario).

3.3.3 A Multimodal Ensemble Classifier

When relevance annotations are used, a further advantage over unsupervised approaches is that the combination of multiple features can be incorporated into the learning process. Here, we present Multimodal Stacking (MMS), an ensemble classification scheme that learns how to combine the outputs of multiple, independently trained classifiers (each one using a different type of features) in order to make a better relevance prediction. The algorithm is inspired from stacked generalization [16], a method for the fusion of heterogeneous classifiers, widely known as stacking. The training of MMS consists of the following steps: Initially, k independent probabilistic classifiers $h_{q_i} : X_i \rightarrow [0, 1], i = \{1, \ldots, k\}$ are built, one for each multi-dimensional feature representation $X_i \in R^{n_i}, n_i > 1$. Each of these single-modality classifiers is then used to predict the classes of all training examples and their predictions are gathered to form a meta training set $D'_q = \{(x'_1, y_1), \ldots, (x'_m, y_m)\}$, where the input vectors $x'_i = [h_{q_1}(x_1), \ldots, h_{q_k}(x_k)]$ consist of the outputs of the single-modality classifiers. This meta training set is used to train a meta classifier $h'_q : X' \rightarrow [0, 1]$, where $X' \in R^k$ is the meta input space and its output is the probability that the image is relevant. At prediction time, the single-modality classifiers are first applied to classify the unknown instance and their outputs are used to form a meta instance that is fed to the meta classifier which makes the final prediction. Compared to early fusion approaches, MMS has the advantage that features of different dimensionalities can be easily combined since all models contribute one-dimensional features to the meta classifier. Furthermore, additional features can be easily incorporated into the final model by directly augmenting the input space of the meta classifier.

3.4 Diversity

Assuming a ranking $im_{r1}, \ldots, im_{rK}$ of the images in S, Deselaers et al. [7] define diversity as

$$D(S) = \sum_{i=1}^{K} \frac{1}{i} \sum_{j=1}^{i} d(im_{ri}, im_{rj}), \quad (3)$$

where $d(im_{ri}, im_{rj})$ is the dissimilarity between the images ranked at positions i and j. Thus, high diversity scores are given to image sets with a high average dissimilarity. We notice that with this definition of diversity, an image set that contains pairs of highly similar (and therefore not diverse) images is allowed to receive a high diversity score if the average dissimilarity is high. This results in a direct negative impact on diversification measures such as CR@K. Therefore, we adopt the following stricter definition:

$$D(S) = \min_{im_i, im_j \in S, i \neq j} d(im_i, im_j), \quad (4)$$

where the diversity of a set S is defined as the dissimilarity between the most similar pair of images in S.

3.5 Optimization

An exhaustive optimization of U has high complexity as it would require computing the utility of all $\frac{N!}{K!(N-K)!}$ K-subsets of I. The computational cost becomes prohibitive even if N is small (e.g., $N \approx 300$) and one is interested in identifying $K = 20$-sized MMR-optimal subsets. Therefore, we apply a greedy, iterative optimization algorithm that was also used in [3], with appropriate changes to reflect our new definitions for relevance and diversity. This algorithm starts with an empty set S and sequentially expands it by adding at each step $J = 1, \ldots, K$ the image im^* that scores highest (among the unselected images), to the following incremental utility function:

$$U(im^*|q) = w \cdot h_q(im^*) + (1-w) \cdot \min_{im_j \in S^{J-1}} d(im^*, im_j), \quad (5)$$

where S^{J-1} represents S at step $J-1$. In the first step, the diversity term of Equation 5 is undefined (S is empty) and thus the image with the highest relevance score is selected.

4. EXPERIMENTAL SETUP

4.1 Retrieving Diverse Social Images Task

During the task, participants were provided with a reference annotated development set of 30 queries - to build

Table 1: Dataset statistics: number of queries (#q), avg. number of relevant/irrelevant images (#r/#i), avg. number of clusters (#c), F1@20 of Flickr ranking (#fp).

Set	#q	#r	#i	#c	#fp
Dev	30	208.5	88.8	23.2	0.477
Test	123	199.9	96.4	22.6	0.470

their approaches - as well as a test set of 123 queries - upon which they were evaluated - of which the ground truth was disclosed only after the end of the task. Ground truth consisted of relevance and diversity annotations provided by experts for all images of each POI. Specifically, each image was first labelled as either relevant or irrelevant and then visually similar relevant images were grouped together into clusters. Performance on each query was assessed using the F1@20 metric that is equal to the harmonic mean of CR@20 and P@20 (the percentage of relevant images in the top 20). Table 4.1 provides some key statistics of the RDSI 2014 dataset. In Section 5.4 we also apply our method on the RDSI 2013 task that is structured similarly to RDSI 2014. The main difference is that RDSI 2013 uses CR@10 as the main evaluation metric. Further details about RDSI 2013 can be found at the corresponding overview paper [9].

A note on ground truth: Although the relevance annotations provided in the RDSI task come from expert annotators, the same type of ground truth could be acquired (for popular queries) from implicit user feedback as described in Subsection 3.3.2. Also, note that relevance annotated images from only 30 queries are used when building the model that estimates the relevance of the images of each test query. In addition to these images, we use the reference Wikipedia page and images provided for each POI as query-specific positive examples. As we explained in Subsection 3.3.2, positive examples can be easily acquired even for rare queries.

4.2 Features

The different features used to represent the dataset are grouped in three categories: a) visual - computed directly from the image content, b) textual - computed from textual annotations, and c) meta - computed from the metadata associated with the images. For all types of multi-dimensional features unit length normalization is applied and cosine similarity/distance is used for relevance/diversity computations.

Visual: After initial experiments with the features made available by the task organizers [9], we extracted the following state-of-the-art features that lead to significantly better performance:

VLAD: $d = 24,576$-dimensional VLAD+CSURF vectors [13] are computed using a 128-dimensional visual vocabulary and then projected to d' dimensions with PCA and whitening. Using $d' = 128$ leads to near-optimal results for both relevance and diversity.

CNN: Convolutional neural network features adapted for the landmark retrieval domain. These features were computed using the Caffe framework [12], with the reference model architecture but using images of 1,000 landmarks instead of ImageNet classes. We collected approximately 1,200 images for each landmark and fed them directly to Caffe for training after preliminary experiments that validated the resilience of the CNN architecture to noisy images. This change of training classes was inspired by recent domain adaptation work presented in [2] that demonstrated higher effectiveness in feature transfer when the training classes are conceptually close to the datasets used in experiments. The outputs of the *fc7* layer, which includes $4,096$ dimensions, constitute the initial features. Similarly to *VLAD*, the dimensionality of *CNN* was reduced to $d' = 128$ with PCA, a value which gives near-optimal results.

Textual: To generate textual features we first transformed each query and each Flickr image into a text document. For queries, we used a parsed version of the corresponding Wikipedia page and for Flickr images we used a concatenation of the words in their titles, descriptions and tags. Bag-of-words features (*BOW*) were then computed for each document using the 10K most frequent terms of the collection as the dictionary and term frequencies as term weights. We found that by repeating the terms in the image titles and descriptions two and three times respectively to increase their contribution in the similarity compared to the terms in the tags that are usually more noisy, lead to increased performance.

Meta: The following one-dimensional features were computed from the textual metadata and used as additional features in the meta input space of the MMS algorithm: distance from the POI, Flickr rank, number of views.

5. EXPERIMENTS

5.1 Relevance Scoring Impact on Quality

In this experiment we want to study the relationship between the quality of the scores provided by the relevance scoring component and the diversification performance as measured by F1@20. A natural way to characterize the quality of relevance scoring is to measure the ability of the scoring function (either supervised or unsupervised) to assign higher relevance scores to relevant images than irrelevant ones. A suitable measure that is commonly used to asses the performance of probabilistic classifiers (as well as scoring functions that assign higher scores to instances that are considered more representative of the class) is the area under ROC curve (AUC). AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative. Thus, an AUC score of 0.5 corresponds to a random classifier and an AUC score equal to 1 corresponds to a perfect classifier that ranks all positives examples higher than negative.

Classifiers of various AUC scores are generated as follows: A positive $r_p = [p_{low}, 1], p_{low} \geq 0$ and a negative $r_n = [0, n_{high}], n_{high} \leq 1$ relevance score range are defined and each positive/negative example is assigned a random score, drawn uniformly from the corresponding range. When $p_{low} > n_{high}$ we have a perfect class separation and AUC equals to 1, while for $p_{low} \leq n_{high}$ classifiers of lower AUC scores can be generated. Using the above process we generated classifiers with average (across all test queries) AUC scores of $\{0.5, 0.6, \ldots, 1.0\}$. Figure 3 shows the average (across all test queries) F1@20 performance when each of the above classifiers is plugged into the MMR algorithm in combination with VLAD (red line), CNN (blue line) and BOW (green line) features for diversity computation. In all cases, we fix $w = 0.5$. We observe that with all types of diversification features, F1@20 shows a strong positive correlation with AUC, suggesting that improving the relevance scoring quality is important for good diversification.

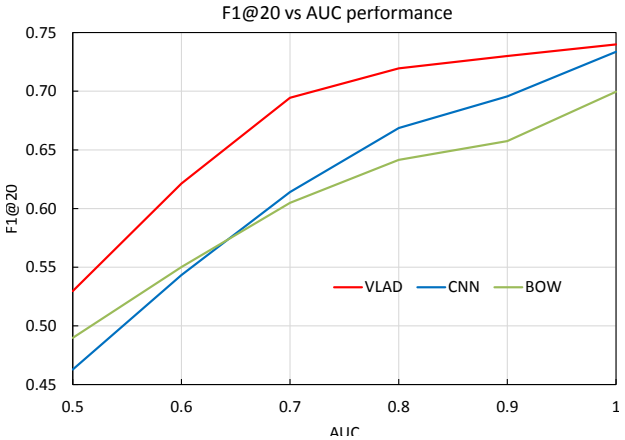

Figure 3: F1@20 vs AUC.

5.2 Unsupervised vs Supervised Relevance

In this section we compare the unsupervised variant of the MMR method [7] (*uMMR*), with the proposed supervised variant, named *sMMR*. In both cases, we use the definition of diversity presented in Equation 4 as it systematically led to better results. We show experiments when either a visual (VLAD)[3] or a textual (BOW) representation is used for both relevance and diversity to highlight potential differences between visual and textual features.

When a visual representation is used, the unsupervised definition of relevance needs to be adapted because the RDSI benchmark provides up to five (instead of 1) example images per query. Rather than arbitrarily choosing one of the available images, we use two alternative definitions of relevance that lead to better results. Given a composite query $q = \{im_{q1}, \ldots, im_{qm}\}$, uMMR$_{avg}$ defines the relevance of each image im_i as the average similarity $R(im_i|q) = \frac{1}{m}\sum_{im_{qi}\in q} s(im_i, im_{qi})$ and uMMR$_{max}$ defines relevance as the maximum similarity $R(im_i|q) = \max_{im_{qi}\in q} s(im_i, im_{qi})$. The two variants coincide when $m = 1$. sMMR can be instantiated with any classification algorithm. We choose L2-regularized Logistic Regression (the LibLinear implementation of [8]) as it provided a good trade-off between efficiency and accuracy compared to other state-of-the-art classifiers in preliminary experiments. To select the regularization parameter c, we perform leave-one(-query)-out cross-validation on the development set and choose the value $c \in \{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$ that gives the best average AUC. Depending on how the training set is composed for each query, three different variants of the sMMR method are created:

sMMR$_q$: The training set contains only the Wikipedia images or the textual representation of the Wikipedia page. Since these are positive examples, we add few randomly chosen Flickr images from other queries as negative examples. sMMR$_q$ represents the case where only few query-specific positive examples are available and attempts to capture the query-specific notion of relevance.

sMMR$_a$: The training set is composed of Flickr images from other queries as positive and negative examples. This variant represents the case where query-specific training examples are unavailable and we can only use positive and negative training examples from other queries, for which sufficient user feedback is available. sMMR$_a$ attempts to capture the application-specific notion of relevance.

sMMR$_{aq}$: The training set is composed of Flickr images from other queries as positive and negative examples as well as the Wikipedia page/images as positive examples. This variant represents the case where few query-specific positive examples are combined with positive and negative examples from other queries for which an adequate amount of user feedback is available. We found that simply combining the few query-specific positive examples with a significantly larger number of application-specific positive examples generates very similar models with the sMMR$_a$ variant. Therefore, we experimented with assigning higher weights to the query-specific positive examples in order to increase their contribution to the formation of the classification boundary. sMMR$_{aq}$ attempts to capture both the query and the application-specific notion of relevance.

[3]Similar results were obtained with CNN features.

Table 2: AUC and F1@20 performance (averaged over test set queries) of the uMMR and sMMR variants using visual and textual features.

	VLAD		BOW	
Method	AUC	F1@20	AUC	F1@20
uMMR$_{avg}$	0.622	0.536	0.660	0.503
uMMR$_{max}$	0.613	0.528	0.660	0.503
sMMR$_q$	0.624	0.535	0.666	0.553
sMMR$_a$	0.664	0.536	0.571	0.481
sMMR$_{aq\times1}$	0.665	0.536	0.575	0.487
sMMR$_{aq\times10}$	0.669	0.532	0.589	0.514
sMMR$_{aq\times100}$	0.690	0.545	0.640	0.541
sMMR$_{aq\times1000}$	**0.693**	**0.561**	**0.670**	**0.555**

Table 2 presents the test set AUC and F1@20 scores of the variants of uMMR and sMMR presented above. AUC is calculated on the relevance rankings produced by each variant (without diversification). The reported F1@20 scores correspond to using a value for $w \in \{0.0, 0.1, \ldots, 1.0\}$ that was tuned to optimize F1@20 on the development set.

When VLAD features are used, we observe that all sMMR variants outperform the uMMR variants in terms of AUC. Among the uMMR variants uMMR$_{avg}$ obtains better AUC than uMMR$_{max}$. Interestingly, despite the absence of any query-specific relevance information, the sMMR$_a$ variant obtains a significantly better AUC than both uMMR$_{avg}$ and sMMR$_q$, suggesting that capturing an application-specific notion of relevance is important in relevance scoring using visual features. As expected, when the query and application-specific relevance information are combined in the sMMR$_{aq}$ variant, AUC performance improves even further, especially when a large weight is assigned to query-specific examples. AUC performance increases from 0.665 when equal weight is given to the query-specific examples (sMMR$_{aq\times1}$) to 0.693 when a 1000 times higher weight is given (sMMR$_{aq\times1000}$). With respect to F1@20, although a strong correlation with AUC is still observed, there are cases where a better AUC does not necessarily lead to better F1@20. This is attributed to a poor tuning of the w parameter as a result of using a small amount of training queries. Nevertheless, we notice that sMMR$_{aq\times1000}$, the best performing variant in terms of AUC, also obtains the best F1@20 score (0.561) that is about 5% better than the 0.536 score obtained by uMMR$_{avg}$.

When textual features are used, the uMMR method (the two variants coincide in this case) is only slightly outperformed by $sMMR_q$ and $sMMR_{aq\times1000}$ in terms of AUC. Contrarily to when visual features are used, $sMMR_a$ obtains a significantly lower AUC compared to both uMMR and the sMMR variants that use query-specific information. This suggests that application-specific information is not sufficient to produce a good relevance scoring when this scoring is based only on the textual modality. Nevertheless, the relevance scoring produced by $sMMR_a$ is better than random scoring (AUC=0.5) and the $sMMR_{aq\times1000}$ variant that combines query and application-specific information is again the best performer. With respect to F1@20, we observe that, with the exception of the $sMMR_a$ and $sMMR_{aq\times1}$ variants, all other variants produce better F1@20 scores than uMMR. When $sMMR_{aq\times1000}$ is used, F1@20 increases from 0.503 to 0.555, a 10% improvement.

To test whether there is statistical significance in the observed differences in F1@20 between uMMR and sMMR we performed the Wilcoxon signed-ranks test between their best performing variants separately for each feature. The null hypothesis is rejected in all cases (VLAD, CNN and BOW) with a $p \leq 0.01$ confirming the superiority of sMMR.

5.3 Multimodal Fusion

So far, we compared instantiations of the uMMR and sMMR methods that used only a single type of features for relevance scoring. In this section, we want to evaluate the performance of the two methods when more features are combined to assign relevance scores.

In particular, to create a multi-feature instantiation of the uMMR method given k types of features, we compute the relevance of each image to the query as $R(im|q) = \frac{1}{k}\sum_{i=1}^{k} R_i(im|q)$, where $R_i(im|q)$ is computed according to equation 2. That is, unsupervised relevance scores are computed independently for each feature and then averaged to produce an overall relevance score. In the case of composite queries, the $uMMR_{avg}$ variant is used as it gave slightly better results than $uMMR_{max}$ in Subsection 5.2. The resulting method is called *uMMR-M*.

To combine multiple features in sMMR, we use the MMS method presented in Subsection 3.3.3. More precisely, the $sMMR_{aq\times1000}$ configuration is used for the single-modality models since it was found superior to other configurations in Subsection 5.2. L2-regularized Logistic Regression is used as classification algorithm for both the single-modality models and the meta model and the regularization parameter is tuned as described in Subsection 5.2 but searching in a wider range of values ($c \in \{10^{-4}, 10^{-3}, \ldots, 10^{-4}\}$) for the meta model. The resulting method is called *sMMR-MMS*.

Table 3 presents the AUC and F1@20 scores obtained with uMMR-M and sMMR-MMS using pairs of the three multi-dimensional features presented in Subsection 4.2 (rows 4-6). We also present results for sMMR-MMS when the three one-dimensional features (meta) presented in Subsection 4.2 are used within the MMS algorithm as described in Subsection 3.3.3 (rows 7-9). Results using each of the three features alone for both relevance and diversity are also reported to facilitate comparison (rows 1-3). AUC scores are calculated on the relevance rankings produced by each variant and the reported F1@20 scores correspond to using a value for $w \in \{0.0, 0.1, \ldots, 1.0\}$ that was tuned to optimize F1@20 on the development set. In all multi-feature instantiations, VLAD features were used for diversification as they gave the best F1@20 scores when tested with artificial classifiers of various AUC performances in Subsection 5.1.

Table 3: AUC and F1@20 performance (averaged over test set queries) of uMMR-M and sMMR-MMS using various combinations of features.

Feature(s)	AUC	F1@20	AUC	F1@20
	$uMMR_{avg}$		$sMMR_{aq\times1000}$	
VLAD	0.622	0.536	*0.693*	*0.561*
CNN	0.733	0.530	***0.808***	***0.572***
BOW	0.660	0.503	*0.670*	*0.555*
	uMMR-M		sMMR-MMS	
VLAD+CNN	0.729	0.580	*0.806*	*0.594*
VLAD+BOW	0.697	*0.586*	*0.727*	0.578
CNN+BOW	0.765	0.588	***0.814***	***0.608***
VLAD+CNN+meta	N/A	N/A	**0.828**	0.619
VLAD+BOW+meta	N/A	N/A	0.740	0.590
CNN+BOW+meta	N/A	N/A	0.827	**0.631**

Looking at the results obtained with single-modality instantiations we observe that the AUC performance obtained with CNN features is significantly better compared to VLAD and BOW when either $uMMR_{avg}$ or $sMMR_{aq\times1000}$ is used for relevance scoring. This superiority of CNN for relevance scoring explains also the better F1@20 performance obtained with CNN features despite the fact that VLAD were found better for diversity scoring in Section 5.1.

With respect to two-feature instantiations we observe that sMMR-MMS obtains better AUC scores than uMMR-M in all cases. The situation is similar with respect to F1@20 where uMMR-M obtains better F1@20 only in one case (VLAD+BOW). We also observe that with either uMMR-M or sMMR-MMS, both AUC and F1@20 are always better than using any of the two features alone with the exception of the VLAD+CNN combination where slightly worse performance is obtained in terms of AUC compared to using the CNN features alone. This can be attributed to the facts that a) CNN and VLAD are both visual representations and thus have a small degree of complementarity and b) CNN significantly outperforms VLAD in terms of AUC. Using the CNN+BOW feature combination and the sMMR-MMS method, a 0.631 F1@20 score is obtained that is about 6% better than the F1@20 obtained by the best single-modality instantiation. To test whether there is statistical significance in the observed differences in F1@20 between uMMR-M and sMMR-MMS we performed the Wilcoxon signed rank test between their best two-feature instantiations (CNN+BOW). The null hypothesis is rejected with a $p \leq 0.01$.

Finally, we see that when the meta features are used, both AUC and F1@20 improve further. In particular, the CNN+BOW+meta combination obtains the highest F1@20 followed by VLAD+CNN+meta and VLAD+BOW+meta.

Time efficiency: Compared to uMMR, sMMR has the additional computational cost of training the relevance detection models. However, these models will typically be trained offline. During the online reranking phase, uMMR has to compute similarity scores between the query image and the top-N results of a search engine while sMMR has to classify each of the top-N images. When a linear model is used (as done here), the computational cost is similar.

Table 4: Comparison to best RDSI task results.

Method	2013 (CR@10)	2014 (F1@20)
Best in RDSI	0.440 [11]	0.597 [6]
This paper	**0.473** +7.5%	**0.631** + 5.7%

Using one core of an i5 2.4 GHz processor, the training of a $\text{sMMR}_{aq\times 1000}$ model on 8919 images took ∼0.5 sec with VLAD and CNN and ∼9 sec with BOW features while using such a model to predict the relevance score of a single image took ∼0.01 ms and ∼0.3 ms respectively. A sMMR-MMS model combining the predictions of any two such models took an additional time of 200 ms to train and ∼0.007 ms to make a prediction. Reranking 300 images to select a 20-sized subset with the greedy algorithm of Subsection 3.5 using VLAD features for diversity took ∼50 ms.

5.4 Comparison to Best RDSI Results

Table 4 compares the test performance of our best diversification method, sMMR-MMS, with that of the best performing systems of the MediaEval 2013 and 2014 RDSI tasks. As in previous experiments, the c parameter of the logistic regression models used in the MMS algorithm was tuned to optimize AUC on the development set. Using this procedure, we built MMS relevance scoring models with the feature combinations presented above and then applied diversification using VLAD. For each sMMR-MMS instantiation, we evaluated its performance on the development set using values for $w \in \{0.0, 0.05, \ldots, 1.0\}$ and selected the instantiation that obtained the maximum F1@20 score and the corresponding w value, and applied it on the test set. For the RDSI 2013 task, CR@10 was used instead of F1@20 for model selection since it was the primary evaluation measure of this benchmark. In both cases, the best results were obtained using the CNN+BOW+meta combination. As shown in Table 4, our method outperforms the best systems of RDSI 2013 and 2014 by 7.5% and 5.7% respectively.

6. CONCLUSIONS - FUTURE WORK

We introduced a supervised version of the popular MMR diversification algorithm and showed that using few query-specific, positive relevance annotations combined with extra-query but application-specific, positive and negative relevance annotations leads to significant performance improvements of relevance scoring in terms of AUC. Furthermore, we captured the strong relationship between the AUC score and a common diversification measure (F1@K) within the MMR framework and showed that improved relevance results in better diversification. Finally, we presented a novel multimodal ensemble classifier (MMS) and showed that it effectively combines different types of multi- and one-dimensional features. The resulting diversification method significantly outperformed competing methods in two benchmarks.

Despite its computational efficiency, the greedy maximization algorithm that we used might fail to find the global optimum of the utility function in Equation 1. To deal with this problem, [7] developed dynamic programming algorithms that consider many alternative paths simultaneously, thus increasing the chance of finding an optimal solution. Although these variants were not considered here, they are directly applicable to our method and could further improve its performance. In this paper, we focused on the relevance scoring component of the MMR method, however, the quality of diversity scoring plays an equally important role. In the future, we want to investigate supervised techniques for improving the quality of diversity scoring and a principled way of combining multiple features to compute diversity scores. Application-wise, we would like to evaluate our method in a privacy-aware retrieval setting.

7. ACKNOWLEDGMENTS

This work is supported by the USEMP FP7 project, partially funded by the EC under contract number 611596.

8. REFERENCES

[1] T. Arni, P. Clough, M. Sanderson, and M. Grubinger. Overview of the imageclefphoto 2008 photographic retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 500–511. 2009.

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *ECCV*, 2014.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, 1998.

[4] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, pages 429–436, 2006.

[5] D. Corney, C. Martin, A. Göker, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, L. Aiello, and B. Thomee. Socialsensor: Finding diverse images at mediaeval 2013. In *MediaEval*, 2013.

[6] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. De Natale. Retrieval of diverse images by pre-filtering and hierarchical clustering. In *MediaEval*, 2014.

[7] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *ACM CIVR '09*, New York, USA, 2009.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[9] B. Ionescu, M. Menéndez, H. Müller, and A. Popescu. Retrieving diverse social images at MediaEval 2013: Objectives, dataset and evaluation. In *MediaEval*, 2013.

[10] B. Ionescu, A. Popescu, M. Lupu, A. Gînsca, and H. Müller. Retrieving diverse social images at MediaEval 2014: Challenge, dataset and evaluation. In *MediaEval*, 2014.

[11] N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, and P. H. Lewis. Experiments in diversifying flickr result sets. In *MediaEval*, 2013.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. A comprehensive study over vlad and product quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, 2014.

[14] E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, and I. Vlahavas. Socialsensor: Finding diverse images at mediaeval 2014. In *MediaEval*, 2014.

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

[16] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[17] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pages 10–17, 2003.