The University of Glasgow at ImageClefPhoto 2009

Guido Zuccon, Teerapong Leelanupab, Anuj Goyal, Martin Halvey, P. Punitha, Joemon M. Jose
University of Glasgow, Glasgow, G12 8RZ, United Kingdom
{guido,kimm,anuj,halvey,punitha,jj}@dcs.gla.ac.uk

Abstract

In this paper we describe the approaches adopted to generate the five runs submitted to ImageClefPhoto 2009 by the University of Glasgow. The aim of our methods is to exploit document diversity in the rankings. All our runs used text statistics extracted from the captions associated to each image in the collection, except one run which combines the textual statistics with visual features extracted from the provided images. The results suggest that our methods based on text captions significantly improve the performance of the respective baselines, while the approach that combines visual features with text statistics shows lower levels of improvements.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Interdependent Document Relevance, Diversity, Novelty

1 Introduction

In this paper we provide an overview of our participation in ImageClefPhoto 2009 and describe the methods adopted to generate the submitted runs. The ultimate aim of this year's retrieval task was to promote diversity against redundancy in document retrieval ranking. The need for diversity in a result list is empirically motivated by several studies [1, 4, 11]. Addressing diversity issues aims to allow retrieval systems coping with poorly specified or ambiguous queries, maximizing the chance to retrieve relevant images to the user's information need.

We recognize that diversity is a broad and under-specified concept: it is possible to refer to topical diversity, source diversity, presentation diversity, etc. In the context of ImageClefPhoto 2009 we identified the following types of diversity: (i) topical diversity, (ii) semantic diversity, (iii) visual diversity.

In four out of five submitted runs, we have tackled the problem of diversity considering topicality and semantics. Under these approaches, the content of the documents have been assumed to coincide with the text (the caption) associated with each image. All the four runs based on text aim to exploit the statistical features associated to the captions, promoting topically and

semantically different documents. In the last of the five runs, however, we considered also visual features, combining these with the topical information provided by the matching between the textual queries associated to the topics and the textual captions. Thus, in this approach both topical and visual diversity were captured in our ranking function.

The evaluation of the participants' results is a central step in a retrieval campaign. In order to evaluate diversity, several measures have been proposed, such as s-recall (also called cluster recall), s-precision, ws-precision [16], α-NCGD [2], subMRR [15]. Of these, only s-recall has been used in ImageClefPhoto 2009. Submitted runs have been discriminated by the performances they obtained in s-recall at 10 and precision at 10. We believe, however, that a thorough evaluation of the runs by means of the diversity measures suggested in the literature would provide more insights for the understanding of how well systems are diversifying document rankings.

The paper continues as follows. In section 2 we describe the approaches implemented to generate the submitted runs. Section 3 reports the results obtained by evaluating the runs against the ground truth and further discussions related to our approaches. Finally, the paper concludes in section 4 where we state the achievements obtained by participating to ImageClefPhoto2009 and future lines of research.

2 Approaches

In the following we illustrate the approaches adopted to tackle the challenging task of promoting diversity in this year's ImageClefPhoto. All our approaches assume a common initial document retrieving and ranking step. In this step, a query is matched against each document in the collection and a ranked list is returned along with the scores associated to each reported document. We refer to this step as the initial ranking process. Afterwards each method processes the common input in order to generate a re-ranking of the initial results. Our approach based on the Quantum Probability Ranking Principle (QPRP) is motivated employing quantum theory, while the others start from empirical observation to derive a model. The approach called "VisualDiversity" is our only method which uses both the visual features of the images and the text features of the associated captions. The remaining methods instead employ exclusively statistics drawn from the provided textual captions.

The remaining of this section is structured as follows. In section 2.2 we illustrate the first of our approaches based exclusively on textual features, the Maximal Marginal Relevance (MMR), derived from [1]. Afterwards we present the methods implemented in $Glasgow - run - 3$ and $Glasgow - run - 4$ (section 2.3), which rely on a clustering procedure to individuate topically and semantically diverse facets of the retrieved documents and re-ranks the initial ranking combining those evidences and the MMR methodology. The main intuitions behind the QPRP and details of the suggested ranking procedure are presented in section 2.4. Finally, in section 2.5 the method called "VisualDiversity", which employs statistics derived both from text captions and from images, is illustrated.

2.1 Initial Ranking

In this section, we briefly describe the data and method employed to generate an initial ranking for the topics of ImageClefPhoto 2009. For each of the 50 topics, a set of subtopics have been individuated by the organizers. For example, topic 16, which title is "queen" has as associated subtopics "queen silvia", "queen rania", "queen sofia", and an "other" cluster, which contains facets not related to the previous clusters. Although the topics where provided with a description and an image example, we decided to use merely the topic's titles as queries. Doing so, we aimed to capture the situation of a user posing a very broad or ambiguous query. The topics have been matched against a collection of nearly 500,000 pairs of images and textual annotations from the Belga News Agency.

In order to retrieve the initial document ranking for our runs *Glasgow - run - 1,3,4,5* we

employed Terrier[1] and we developed our method using Java. For the *Glasgow - run - 2*, instead, we obtained the initial document ranking from Lemur[2], and the re-ranking procedure has been implemented in C++. In both cases, stemming and stop word removal has been applied using the same dataset, and the internal implementation of the Okapi's retrieval function [12] has been used to retrieve the initial set of documents and to estimate the probability of relevance of each document. The initial ranking has been also used as baseline to evaluate the performances gains our methods obtained.

2.2 MMR

Maximal Marginal Relevance (MMR) has been one of the first approaches that aimed to exploit the diversity between documents by combining it with the relevance score of each document [1]. We implemented the MMR approach in *Glasgow - run - 1* with the intention of using this as a benchmark against the approaches we propose in the following. More recent and sophisticated methods, such as [15], might have been implemented as well. However, they require a thorough and complex exploration of the parameters space. This is not possible in the context of ImageClefPhoto2009 since there are not either a training set or the subtopical qrels to be employed in the learning of the optimum parameter values. The MMR approach we adopted is characterized by the following equation:

$$MMR_{J+1} \equiv \underset{x_i \in I \setminus J}{\operatorname{argmax}} [\alpha S(x_i; q) + (1 - \alpha) D(x_i; (x_1, ...x_J))] \quad (1)$$

where I is the set of initial results retrieved by the IR system; J is a set of re-ranked results at iteration J; q is a query; x_i is a candidate document in $I \setminus J$, which is the set of documents that have not been ranked yet; and x_J is a document in J, i.e. the set of documents that have been already ranked. Function $S(x_i; q)$ is a normalised similarity metric used for document retrieval, such as Okapi BM25, while $D(x_i; (x_1, ...x_J))$ is a diversity metric, which we implemented as the opposite of the cosine similarity between document vectors which components are the BM25 weights of the respective terms. Intuitively, the MMR method maximizes marginal relevance and diversity in document retrieval at each iteration by linearly combining the relevance of the documents with the dissimilarity to previous retrieved documents. The parameter $\alpha > 0.5$ means that similarity to the query is more important than novelty/diversity, while $\alpha < 0.5$ represents situations where novelty/diversity is more important than relevance to the query. In previous experiments, we found that $\alpha = 0.3$ yields satisfying results since the obtained ranking still contains a great number of relevant documents. We therefore selected this value as a constant for all our runs based on MMR.

Following the MMR method, the first image to be inserted in the final ranking is the one that is most similar to the query, since no previous results are in the ranking and the diversity component of eq. 1 does not contribute to the document's final score. Afterwards, the image $x_i \in I \setminus J$ which obtains the highest MMR score in each iteration is added to the results list until all images are selected. In our *Glasgow - run - 1*, we re-ranked the top 500 images.

2.3 Semantic Clustering

2.3.1 Clustering as pre-processing for MMR improvement

Although MMR can adequately combine relevance and diversity in the ranking results, the selection of the first document to retrieve might represent a problem. In fact, the score for a new document to be added depends upon the documents ranked at the previous steps. As a result, an unfavorable choice at an early position in the result list may adversely deteriorate the entire ranking. The re-ranking procedure might get trapped in a local maximum of distance (diversity). MMR suggests that the first document retrieved should be the one with highest similarity score. However, there have been several methods proposed to leverage the problem of non-optimal ranking due to initial

[1] http://ir.dcs.gla.ac.uk/terrier/.
[2] http://lemurproject.org/.

bad selection such as dynamic programming [3], requiring high computation cost. In our *Glasgow - run - 3*, we strive to deal with this problem using a light-weight approach by incorporating within a clustering technique.

In our experiment, we assumed that top 100 ranked documents from an initial set of results are highly relevant to the user's information need. Note that this is in some extent similar to pseudo relevance feedback, but in our approach re-ranking is performed without an extended interaction. We then cluster the top 100 documents using Hierarchical Agglomerative Clustering due to its low computational cost. Subsequently, we consider only the biggest cluster in order to extract the first document to rank. This choice is motivated by the assumption that the biggest cluster in the top list could be used to characterize the most popular subtopic users identify. The medoid of the biggest cluster is selected as a representative and as an initial choice for re-ranking. Later re-ranking iterations are processed following the MMR method on the top 500 images from the initial ranking obtained by Okapi.

2.3.2 Revisiting the integration of Clustering and MMR techniques

A popular method to obtain diverse but relevant results in image and information retrieval is re-ranking based on clustering techniques. Most methods deal with the diversity problem using a two-step approach. In the first step the set of potentially relevant documents is acquired using standard retrieval methods. In the second step these documents are clustered, assuming that individual clusters have the potential to represent different sub-topics of results amongst the first positions. To build the final ranked list based on this approach, a representative document is then taken from each cluster in a round-robin fashion and added to the result list [5].

There are three common approaches used to select the representative document of each cluster. The first is to select the cluster's document, with the highest similarity score to the given query [3]. In this case, relevance is considered solely after clustering. A second approach is to select the medoid[3], which is assumed being the best representative of the cluster [9, 14]. The third method selects the most similar document to the members of the selected cluster [6]. Therefore, all the presented methods merely rely on clustering and consequently ignore the diversity of documents in the final list. As a result, the top ranked results are still similar or semantically close to each other. To the best of our knowledge, no existing *cluster-based* methods consider the diversity and relevance criterion aiming to find the optimal balance between similarity of the results to queries and diversity between candidate documents within the documents ranking.

In this paper, we propose an approach that takes into consideration the diversity criterion on the selection of documents within clusters. This approach has been implemented in *Glasgow - run - 4* on the top 500 images from the initial ranking employing the Expectation Maximisation (EM) clustering algorithm, which has been observed to perform best in clustering similar images in our previous experiments. We set the number of clusters to be 20, aiming at producing a ranking that, in the top 20, holds as many relevant images that are representative of the different sub-topics within the results. The standard deviation is set at 6×10^{-6}. The seed number is set at 100 with 100 maximum iterations. After clustering the dataset of the initial ranking, clusters are ranked according to average relevance score of instances within clusters as defined by

$$S(C_k; q) = \frac{1}{I_k} \sum_{i=1}^{I} s(x_{ki}, q) \tag{2}$$

where $S(C_k; q)$ is the average similarity of the cluster C_k to query q; I_k is the number of documents in cluster C_k; and $X = \{x_1, ..., x_n\}$ is the initial set of potential relevant documents. Subsequently, the retrieval consists of two steps: in the first step, we select clusters according to the means of their similarity scores. We used a round robin approach to simplify the selection at cluster level in our preliminary experiment. In the second step, MMR is employed to select a document to be added to the results list from a particular cluster in each iteration. We assume that applying diversity based re-ranking on clusters can enhance the overall diversity of the document ranking,

[3]The document closest to the centroid of the cluster.

yet maintaining relevancy to the query. In each iteration, the number of candidate documents to be compared is reduced from the entire set of initial rank to a set of a specific cluster. MMR can be used to fuse similarity and diversity:

$$R(x_k^*; q, (x_{r1}, ...x_{rJ})) = \alpha S(x_k^*; q) + (1 - \alpha) D(x_k^*; (x_{r1}, ...x_{rJ})) \quad (3)$$

where R is the retrieval score for a candidate document x_k^* in a specific cluster k given query q and S is the similarity of the candidate document given the query. α is a weight used to balance similarity and diversity, and it has been set to 0.3 in this experiment. For partial results $(x_{r1}, ...x_{rJ})$ in the final ranked list at iteration J, we define the *dissimilarity* score for a candidate image x_k^* to be inserted in the ranking at iteration J+1 as

$$D(x_k^*; (x_{r1}, ...x_{rJ})) = \frac{1}{J} \sum_{j=1}^{J} d(x_k^*, x_{rj}) \quad (4)$$

2.4 QPRP

For the run $Glasgow - run - 2$ we employed an approach based on the theoretical framework of the Quantum Probability Ranking Principle (QPRP), introduced in [17]. The framework has justification from quantum probability theory and posits that documents have to be ranked not only based on the probability of relevance of each document to the user's information need, but also according to the interdependent relationships between documents at relevance level. This approach then aims to capture and model interdependent document relevance and diversity in document ranking from first principles. Although the QPRP presents still several open questions, in particular related to the estimation of the probability of relevance and the interference effects, we identified in ImageClefPhoto 2009 a suitable testing scenario for the current developments accomplished towards a full understanding of the capabilities of the QPRP. For a complete presentation of the QPRP framework and of the open questions related to it the reader is referred to [17]. In this paper, instead, we provide a brief introduction to the principle and the main intuitions underlying it.

Consider the following scenario. A set of documents has been retrieved in response to an information need and a probability of relevance to the information need is attached to each document. The system is asked to provide a documents ranking to present to the user. Assume the document with highest probability of relevance to the information need is ranked at the first position[4].

Which document the system should rank next? The widely used Probability Ranking Principle (PRP) [13] posits that documents should be ranked in descendent probability of relevance to the information need. Thus, according to the PRP the system should present at the second position in the rank the document with the second highest probability of relevance. Consider this example. The set of retrieved documents is $R = \{d_1, d_2, d_3\}$ and the probability of relevance of each document to the information need is respectively $P(d_1) = 0.9$, $P(d_2) = 0.8$, $P(d_3) = 0.7$. Following our assumption, d_1 is presented at the first rank; a system implementing the PRP ranks at the second position in the rank document d_2, since $P(d_2) > P(d_3)$.

Conversely, the QPRP suggests that interdependent document relationships at relevance level should be accounted for in the document ranking. In particular, the QPRP ranks d_2 at the second position in the rank if and only if $P(d_2) + I_{d_1,d_2} > P(d_3) + I_{d_1,d_3}$, where I_{d_1,d_i} is the quantum interference between the documents that have been already ranked, d_1 in our example, and document d_i. The QPRP posits that the interference between documents occurs at relevance level and that it models interdependent document relevance.

Regarded to what do PRP's and QPRP's rankings differ? Intuitively, the QPRP is a generalization of the PRP: in particular when the interference term equals zero for each pair of documents, then the two principles provide the same ranking list. However, when I_{d_1,d_2} (or

[4]This is a quite intuitive assumption that is justified both from the PRP and the QPRP.

equally I_{d_1,d_3}) is not null, the ranking suggested by the PRP is subverted by the interference term if

$$I_{d_1,d_3} > P(d_2) - P(d_3) + I_{d_1,d_2} \tag{5}$$

Although there is no guarantee that this happens in document ranking, a follow-up paper to [17] will empirically illustrate the presence of such phenomena and thus that PRP and QPRP provide significant different rankings.

Ranking documents with the QPRP. In the following we describe the complete ranking strategy suggested by the QPRP. The first document that is ranked is the one that has higher probability of relevance given the information need. This is the same as the PRP suggests, and is also the best we can do from a utility theory point of view: given the evidence that we have before starting to rank, the best document to be returned to the user at rank one is the one that is expected to be most relevant to his information need. We indicate this document as $d@1$, meaning the document that is ranked at position 1. The document that has to be returned at second position in the ranking is the one which maximizes

$$P(d_i) + I_{d@1,d_i} \tag{6}$$

with $d_i \in \mathcal{RE} \setminus \{d@1\}$, being $\mathcal{RE}$ the set of documents retrieved in response to the query[5]. Let $\mathcal{RA}$ contain the documents that have been already ranked; then the document that has to be returned at rank n is the one that maximizes

$$P(d_i) + \sum_{d_x \in \mathcal{RA}} I_{d_x,d_i} \tag{7}$$

with $d_i \in \mathcal{RE} \setminus \mathcal{RA}$.

The interference term. At this stage it appears clear the central role of the interference term in the formulation of the QPPR. However, what is the interference term governed by? The interference term is expressed (refer to [17] for the mathematical justification) as:

$$I_{d_x,d_y} = 2\sqrt{P(d_x)P(d_y)}cos\theta_{d_x,d_y} \tag{8}$$

where θ_{d_x,d_y} is the difference of the phases[6] associated to the probabilities of d_x and d_y. Thus the interference's behaviour depends from the phase difference θ. A correct estimation of either θ or of the probability (of relevance) amplitudes associated to documents is then essential for an effective modeling of interdependent document relevance using the QPRP framework: this is still argument of research. Thus, we employed a rather naïve estimation of θ based on the opposite of the Pearson's correlation between vectors associated to documents. Each component of the vectors is associated to a term of the collection's vocabulary and is characterized by the correspondent Okapi weight.

In *Glasgow - run - 2* we implemented the QPRP paradigm, by re-ranking an initial document ranking containing the 100 most relevant documents to a query retrieved using Lemur. The parameters of the Okapi's score function were set to the values suggested in [12].

2.5 Visual Diversity

In $Glasgow - run - 5$ we combined text statistic with visual features. Before generating a ranking encoding diversity exploiting the visual content of the images, we processes for each query the top 100 documents retrieved employing the Okapi scoring function. Afterwards we apply factor analysis and bi-clustering to the visual features of the gathered documents. These two techniques are presented in the next sections.

[5]Thus $\mathcal{RE} \setminus \{d@1, \ldots, d@x\}$ represents the set of documents retrieved but not yet ranked after x positions.

[6]Recall that the QPRP assumes the presence of a complex probability (of relevance) amplitude ϕ_i associated to each document d_i, where probability amplitudes and probabilities are related by $P(d_i) = |\phi_i|^2$.

2.5.1 Factor Analysis

In the following, we illustrate factor analysis in details. Factor analysis (FA) is a linear method for dimensionality reduction based on the second-order data summaries (covariances) [8]. The underlying assumption of Fa is that measured variables depend on some unknown, and often unmeasurable, common factors. The aim of FA is to uncover such relations, and thus it can be employed to reduce dataset's dimensions.

The main intuition under FA is that a set of p standardized variables (in our case images in the result set) $x = \{x_1, ..., x_p\}$ can be represented as a linear combination of latent orthogonal uncorrelated variables $f = \{f_1, ..., f_m\}$ called common factors, summed to unique (specific) factors $e = \{e, \ldots, e_p\}$. The unique factors $e_1, \ldots, e_p$ express the part of x that cannot be explained by the common factors. Formally,

$$x_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + ... + \lambda_{im} f_m + e_i \tag{9}$$

where coefficients $\lambda_{ij}, i \in \{1, ..., p\} j \in \{1, ..., m\}$, are called *factor loadings* and represent the correlation between variable x_i and factor f_j. Each common factor represents some common characteristics of the data under consideration.

Eq. (9) can be rewritten in matrix form, with obvious notation as

$$x = \Lambda f + e \tag{10}$$

where $x = [x_1\ x_2\ x_3\ ....\ x_p]^T, e = [e_1\ e_2\ e_3\ ....\ e_p]^T$, and

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & ... & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & ... & \lambda_{2m} \\ & . & . & ... & . \\ \lambda_{p1} & \lambda_{p2} & ... & \lambda_{pm} \end{pmatrix}$$

Λ is also called the loading matrix.

The *communality* in FA represents the extent of overlap between the variables and the factors. In detail, communality is the proportion of variance of a particular variable (i.e. images in the initial ranking) that is due to common factors, that is, the proportion of variance that each variable has in common with other variables. The communality for each variable is equal to the sums of squares of the loadings for the variable. Let the communality of x_i is c_i then:

$$c_i = \sqrt{\lambda_{i1}^2 + \lambda_{i2}^2 + ... + \lambda_{im}^2} \tag{11}$$

2.5.2 Bi-clustering

Bi-clustering is a technique for two-way data analysis. A two-way dataset is a matrix with rows r_i, column c_j and entries a_{ij}. The purpose of bi-clustering is to find subgroups of rows and columns, which are as similar as possible to each other and as different as possible to the rest.

Data analysis by using simple clustering techniques such as K-Means clustering makes several a-priori assumptions that may not be adequate in all circumstances. In fact, clustering can be applied to either images or factors, implicitly directing the analysis to a particular aspect of the system (*e.g.* groups of factors or groups of images). Also, clustering algorithms usually try to get disjoint sets of elements, constraining an image or factor to belong exclusively to one cluster.

The concept of a bi-cluster allows for a more flexible framework for data analysis. A bi-cluster is defined as a submatrix containing a set/subset of images and a set/subset of factors. Given a loading matrix, we can distinguish the data patterns it represents by a collection of bi-clusters, each representing a different type of joint characteristics of a subset of images in a corresponding subset of factors. In brief, there are no a-priori constraints on the organization of bi-clusters and in particular, images or factors can be part of more than one bi-cluster or of no bi-cluster.

Many algorithms like: BiMax, Plaid, Spectral, XMotifs have been proposed for bi-clustering; a comparison of these algorithms is provided in [7, 10]. In our diversity based retrieval technique,

the BiMax algorithm is implemented because of its lower computational complexity. In a binary matrix, the BiMax algorithm finds sub-matrix where all entries are one. The algorithm iterates through the following two steps:

1. Rearrange the rows and columns to concentrate ones in the upper right of the matrix;
2. Divide the matrix into two submatrices.

Whenever only ones are found in one of the submatrices, this submatrix is returned. In order to obtain satisfying results the method is restarted several times with different starting points.

In the next section we detail our proposed approach for re-ranking of text based results to promote visual diversity in document ranking.

2.5.3 Visual Features based Diversity

Our main goal is to re-rank text based results to make the ranked list more diverse based on visual features. For this purpose, we use two low-level features defined in MPEG-7 standards i.e. Color Structure and Edge Histogram. While color structure gives information about the color distribution, edge histogram provides information about edge distribution. We combined these two features into one feature by a concatenation to process an image further. To apply factor analysis, a matrix F is generated by using the top 100 results from text based retrieval, in which each row contains a particular feature component of all top 100 images. Let the component number (dimensionality) in a low-level feature be d, the size of F will be $(d \times 100)$.

Factor analysis is applied on the covariance matrix of F which generates the loading matrix Λ. In the loading matrix, each row represents an image by its factor loadings for all common factors from 100 top results. Further, we calculate communality for each image by using eq. 11. The Communality tells us about the characteristics of an image which are common with other images in the result set.

Since each common factor represents some specific visual characteristics of the result set, combinations of different common factors are also an efficient representation of result set characteristics. For diversity based retrieval, our main goal is to create subgroups of the loading matrix such that a fixed number of factors behave in a similar pattern for all images in that subgroup. A similar pattern means that a factor is either highly correlated or less correlated with all components in a subgroup. This is a two-fold requirement: (1) images within a subgroup will represent some common characteristic; and (2) images from two different subgroups will contain somewhat different characteristics.

We can employ three different methods to cluster the loading matrix. The image clustering only considers the overall distance between two images by using all factors loadings, which is questionable as we have to think about factor combinations as well. In factor clusters, we have a group of different factors which behave almost the same for all images. However, it will miss some factor combinations that behave similarly only for some images, due to the constraint that each cluster should contain all images.

To overcome the problems in these one-way clustering, we turn to bi-clustering over the loading matrix. Bi-clustering results in subgroups in which some particular factors are highly correlated with all images in that subgroup. Each bi-cluster hence contains factors or factor components as wished. Moreover, each bi-cluster represents distinct characteristics of the result set.

We used the BiMax algorithm [10] for bi-clustering the loading matrix Λ. BiMax works on a binary matrix. Let a_{ij} be an original element in Λ before binarization and b_{ij} the binary value, the binarizarion function is defined as follows:

$$b_{ij} = \begin{cases} 0 & if \lambda_{ij} \leq \text{Median of all loadings for factor j} \\ 1 & if \lambda_{ij} > \text{Median of all loadings for factor j} \end{cases} \tag{12}$$

Factors will be encoded as a 1 if they have loading greater than the median, 0 otherwise. The median is used for partitioning the loading columns in two equal parts.

After bi-clustering, each bi-cluster contains some distinct characteristics of the result set which is used to generate diversity in the final ranking. To create a final document ranking, we first calculate the ranking within a cluster based on the ranking in text based results and the communality of each image. The text ranking (R_{text}) characterizes relevance while the communality based rank (R_{comm}) returns the most common images from that cluster. Finally, we re-rank images within a cluster in increasing order of their final rank R_f:

$$R_f = (\alpha)R_{text} + (1-\alpha)R_{comm} \tag{13}$$

Lesser the value of R_f, higher the image will be in the final document ranking. Note that a high α means that similarity to the query is more important than diversity, while low α represents situations where diversity is more important than similarity. After some experiments we fixed value of α to 0.4.

To generate a final document ranking, we first rank the clusters based on the best R_{text} among all images within a cluster and then we select one image from each cluster in descending order of cluster ranks repeatedly till all clusters are completely exhausted. To select an image from a cluster, we just choose the image having minimum R_f in that cluster. In this way we generate a final document ranking containing more visual diversity than the initial ranking.

3 Results and Discussions

The performances of our runs against the values obtained by the initial rankings given as input to each approach are reported in Tables 1, 2, and 3. Since the ultimate aim of this year's campaign was to promote diversity in document ranking, we report and discuss only the values relative to cluster recall [16]. This measure is defined as follow:

$$CR@k = \frac{\cup_{i=1}^{k} subtopics(d_i)}{n_A} \tag{14}$$

where the function $subtopics(d_i)$ returns those subtopics or facets that are covered by document d_i, n_A is the total number of subtopics for the given topic and k is a rank position. The evaluation of our runs using the usual IR measures such as precision, recall, MAP, etc, can be found in the campaign summary released by the organizers. Note that the results reported for $Glasgow-run-2$ differ from the one reported by the ImageClefPhoto's organizers. The submitted run, in fact, contained a mistake being thus wrongly evaluated. We corrected this problem after the submission and the organizers kindly agreed to evaluate the correct rankings against the ground truth for the subtopics. This evaluation is reported in Table 2, together with the evaluation of the initial ranking obtained with Okapi (which has been given as input to our approach). The values reported in the table, however, cannot be fairly compared to the one obtained in the other runs or by other participants, since those rankings provided by the QPRP were not included in the pool that have been manually judged by the ImageClefPhoto assessors.

Run id	CR@5	CR@10	CR@15	CR@20	CR@30	CR@50	CR@100
Glasgow 1	0.4638 +1.37%	0.6091 +11.00%	0.6662 +11.15%	0.6929 +8.26%	0.7422 +7.57%	0.8079 +8.11%	0.8492 +1.90%
Glasgow 3	0.5658 +19.15%	0.6691 +18.98%	0.7175 +17.50%	0.7775 +18.25%	0.8089 +15.19%	0.8455 +12.20%	0.9046 +7.91%
Glasgow 4	0.5654 +19.10%	0.6401 +15.31%	0.7314 +19.07%	0.7551 +15.82%	0.7751 +11.49%	0.8265 +10.18%	0.8839 +5.75%
baseline	0.4574	0.5421	0.5919	0.6356	0.6860	0.7423	0.8330

Table 1: Overview of the cluster recall (s-recall) evaluation and improvements with respect to the baseline for the runs Glasgow 1, Glasgow 3, and Glasgow 4.

Run id	CR@5	CR@10	CR@15	CR@20	CR@30	CR@50	CR@100
Glasgow 2	0.5371	0.6039	0.6519	0.6634	0.7143	0.7724	0.8137
	+17.63%	+10.73%	+8.45%	+3.70%	+5.83%	+5.29%	–
baseline	0.4424	0.5391	0.5968	0.6388	0.6726	0.7315	0.8137

Table 2: Overview of the cluster recall (s-recall) evaluation and improvements with respect to the baseline for the runs Glasgow 2. Note that CR@100 coincides both for the baseline and for QPRP since our method re-ranks only the top 100 documents for each query.

The performances obtained by $Glasgow-run-1$ (the MMR approach) are superior to the one obtained by the relative initial ranking. However, the empirical evaluation shows that our runs combining MMR and two different clustering approaches ($Glasgow-run-3$ and $Glasgow-run-4$) out-perform both the baseline and the original MMR approach. Also the approach based on the quantum probability framework (QPRP, $Glasgow-run-2$), obtains significantly better values of s-recall with respect to the initial ranking obtained employing the Okapi retrieval schema. Conversely, the results obtained in $Glasgow-run-5$ are not consistently superior to the relative baseline, although improvements are obtained for cluster recall at ranks less than 30. We still have to examine if this happens because of a flow in our methodology or because the way we combine visual features and text statistics is not effective for diversify document rankings. However, we hypothesize that the poor results obtained by our visual diversity run suggest that the diversity in the ImageClefPhoto 2009 dataset is preeminently topical rather than visual.

Run id	CR@5	CR@10	CR@15	CR@20	CR@30	CR@50	CR@100
Glasgow 5	0.4801	0.5481	0.6491	0.6571	0.6785	0.7422	0.8335
	+4.72%	+1.09%	+8.81%	+3.27%	-1.10%	-0.01%	+0.05%
baseline	0.4574	0.5421	0.5919	0.6356	0.6860	0.7423	0.8330

Table 3: Overview of the cluster recall (s-recall) evaluation and improvements with respect to the baseline for the runs Glasgow 5.

4 Conclusions

In this paper we have presented a summary of our participation in ImageClefPhoto 2009 evaluation campaign. The methods we proposed aimed to obtain diversified ranking in function of topical, semantical and visual diversity. We have proposed four new approaches (one of them combined both visual and textual feature) and we have tested a well known method for combining relevance and document dissimilarity (MMR). The obtained results are promising, in particular for the runs exploiting only textual statistics. However, we recognize the need to perform a through evaluation of our approaches employing a series of evaluation measures tailored to capture diversity and novelty citeClarke:2008,Zhai:2003. This will be subject of future work. Moreover, for each approach proposed in this paper, we have identified the following future directions of research:

- **Semantic clustering:** In our clustering and MMR approaches, the documents were represented by the Okapi weights associated to the terms occurring in each document. We claim that in doing so the semantic of each documents has been taken in consideration when diversity of the document ranking was required. Although we obtained a good level of diversity, effectiveness might be improved by employing a different document representation. In particular, we propose to employ a document representation based on co-occurrence statistics and adopt the methodology proposed in [18] to capture semantic relationships between documents so as to increase document diversity at semantic level.

- **QPRP:** although the run based on the QPRP framework provides satisfying results in terms of diversity, the research on the framework itself is still at its early stage of development. Several research questions have been stated during this paper and in [17]. The most stringent ones relate to the estimation of the initial probability (and in particular of the complex amplitudes) or alternatively to the approximation of the interference term.
- **VisualDiversity:** In $Glasgow - run - 5$ we combined visual and textual features to re-rank the results retrieved using the Okapi schema, aiming at diversifying the document rankings. In particular, during the re-ranking process, we used factor analysis, bi-clustering and successively from each bi-cluster we selected an image based both on its relevance and communality. Although the results from the proposed approach are not completely satisfying (especially for cluster recall at ranks higher than 30), there might be some chances to improve performances. A possible solution would be to employ different criteria for selecting an image from a bi-cluster. In particular, the information from re-ranked images can be exploited to select the next best image from a bi-cluster.

Acknowledgments

The authors are grateful to the ImageClefPhoto 2009 organizers, and in particular to Monica L. Paramita, for the additional evaluation of the initial rankings (baselines) and for the results from the corrected $Glasgow - run - 2$.

References

[1] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proc. 21st Int. ACM SIGIR Conf. on Research and Development in IR*, pages 335–336, New York, NY, USA, 1998. ACM.

[2] C. L.A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08: Proc. 31st Int. ACM SIGIR Conf. on Research and Development in IR*, pages 659–666, New York, NY, USA, 2008. ACM.

[3] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *CIVR '09: Proc. 8th ACM Int. Conf. on Image and Video Ret.*, Santorini, Greece, July 2009. ACM.

[4] M. Eisenberg and C. Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *JASIS*, 39(5):293–300, 1988.

[5] M. Ferecatu and H. Sahbi. Telecom paristech at imageclefphoto 2008: Bi-modal text and image retrieval with diversity enhancement. In *Working Notes for the CLEF 2008 workshop*, 2008.

[6] M. Halvey, P. Punitha, D. Hannah, R. Villa, F. Hopfgartner, A. Goyal, and J. M. Jose. Diversity, assortment, dissimilarity, variety: A study of diversity measures using low level features for video retrieval. In *ECIR '09: Proc. 31st Eu. Conf. on Research on Advances in IR*, pages 126–137, Berlin, Heidelberg, Germany, 2009. Springer-Verlag.

[7] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res*, 13(4):703–716, April 2003.

[8] L. Lebart. *Multivariate Descriptive Statistical Analysis (Probability & Mathematical Statistics)*. John Wiley & Sons Inc, May 1984.

[9] T. Leelanupab, F. Hopfgartner, and J. M. Jose. User centred evaluation of a recommendation based image browsing system. In *IICAI '09: Proc. 4th Indian Int. Conf. on AI*, December 2009. to appear.

[10] S.C. Madeira and A.L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45, Jan.-March 2004.

[11] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *SIGIR '06: Proc. 29th Int. ACM SIGIR Conf. on Research and Development in IR*, pages 691–692, New York, NY, USA, 2006. ACM.

[12] S. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at trec-4. In *NIST Special Publ. 500-236: The 4th Text REtrieval Conf. (TREC-4)*, pages 73–96, 1995.

[13] S. E. Robertson. *The probability ranking principle in IR*, pages 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[14] T. Urruty, F. Hopfgartner, Hannah D., D. Elliott, and J. M. Jose. Supporting aspect-based video browsing - analysis of a user study. In *CIVR '09: Proc. 8th ACM Int. Conf. on Image and Video Ret.*, Santorini, Greece, July 2009. ACM.

[15] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR '09: Proc. 32nd Int. ACM SIGIR Conf. on Research and Development in IR*, pages 115–122, New York, NY, USA, 2009. ACM.

[16] C.X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03: Proc. 26th Int. ACM SIGIR Conf. on Research and Development in IR*, pages 10–17, New York, NY, USA, 2003. ACM.

[17] G. Zuccon, L. Azzopardi, and C.J. van Rijsbergen. The quantum probability ranking principle. In *ICTIR '09: Proc. 2nd Int. Conf. on the Theory of IR*, 2009. to appear.

[18] G. Zuccon, L. A. Azzopardi, and C. J. Rijsbergen. Semantic spaces: Measuring the distance between different subspaces. In *QI '09: Proc. 3rd Int. Symp. on Quantum Interaction*, pages 225–236, Berlin, Heidelberg, 2009. Springer-Verlag.