

# Chapter 15

## Revisiting Sub–topic Retrieval in the ImageCLEF 2009 Photo Retrieval Task

Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose

**Abstract** Ranking documents according to the Probability Ranking Principle has been theoretically shown to guarantee optimal retrieval effectiveness in tasks such as ad hoc document retrieval. This ranking strategy assumes independence among document relevance assessments. This assumption, however, often does not hold, for example in the scenarios where redundancy in retrieved documents is of major concern, as it is the case in the sub–topic retrieval task. In this chapter, we propose a new ranking strategy for sub–topic retrieval that builds upon the interdependent document relevance and topic–oriented models. With respect to the topic–oriented model, we investigate both static and dynamic clustering techniques, aiming to group topically similar documents. Evidence from clusters is then combined with information about document dependencies to form a new document ranking. We compare and contrast the proposed method against state–of–the–art approaches, such as Maximal Marginal Relevance, Portfolio Theory for Information Retrieval, and standard cluster–based diversification strategies. The empirical investigation is performed on the ImageCLEF 2009 Photo Retrieval collection, where images are assessed with respect to sub–topics of a more general query topic. The experimental results show that our approaches outperform the state–of–the–art strategies with respect to a number of diversity measures.

---

Teerapong Leelanupab  
University of Glasgow, Glasgow, G12 8RZ, United Kingdom e-mail: [kimm@dcs.gla.ac.uk](mailto:kimm@dcs.gla.ac.uk)

Guido Zuccon  
University of Glasgow, Glasgow, G12 8RZ, United Kingdom e-mail: [guido@dcs.gla.ac.uk](mailto:guido@dcs.gla.ac.uk)

Joemon M. Jose  
University of Glasgow, Glasgow, G12 8RZ, United Kingdom e-mail: [jj@dcs.gla.ac.uk](mailto:jj@dcs.gla.ac.uk)

## 15.1 Introduction

Information Retrieval (IR) deals with finding documents relevant to a user's information need, usually expressed in the form of a query (van Rijsbergen, 1979). Documents are usually ranked and presented to the users according to the Probability Ranking Principle (PRP), that is, in decreasing order of the document's probability of relevance (Robertson, 1977). This ranking strategy is well accepted in IR, and can be justified using utility theory (Gordon and Lenk, 1999a). However, in particular scenarios, ranking documents according to the PRP does not provide an optimal ranking for the user's information need (Gordon and Lenk, 1999b). For example, this happens when redundant documents are of major concern, or when a broad view about the query topic is needed, thus aiming to retrieve all its possible sub-topics. In this situation, the PRP does not provide a satisfying ranking because it does not account for interdependent document relevance due to the assumption of independence between assessments of document relevance.

A number of recent approaches attempt to overcome PRP's limitations (Carbonell and Goldstein, 1998; Wang and Zhu, 2009; Zuccon and Azzopardi, 2010). The suggested approaches were tested on a novel retrieval task, called sub-topic retrieval (Zhai et al, 2003). In this task, documents are assessed with respect to the number of sub-topics. Interdependent document relevance is introduced in the evaluation measures, which reward strategies that retrieve all the relevant sub-topics at early ranks, while penalising unnecessary redundancy. This means promoting novelty and diversity within the ranking. The need for diversity within document rankings has been motivated by several empirical studies (Agichtein et al, 2006; Eisenberg and Berry, 2007). Addressing diversity issues allows retrieval systems to cope with poorly specified or ambiguous queries, maximizing the chance of retrieving relevant documents, and also to avoid excessive redundancy, providing complete coverage of sub-topics in the result list.

From the current strategies for sub-topic retrieval, two common patterns can be observed with respect to the modality used to achieve ranking diversification:

**Interdependent document relevance paradigm:** Relationships between documents are taken into account when ranking. Strategies maximise, at each rank position, a function that mixes relevance estimates and document relationships. This approach is followed by heuristics and strategies such as Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998), interpolating document relevance and diversity estimation, and Portfolio Theory (PT) (Wang and Zhu, 2009), combining relevance estimations and document correlations with respect to the previous ranked documents.

**Topic-oriented paradigm:** Relationships between retrieved documents are used to model sub-topics regardless of document relevance. Documents are thus characterised with respect to the sub-topics they cover using techniques such as clustering (Deselaers et al, 2009; van Leuken et al, 2009), classification (Huang et al, 1998), LDA (Blei et al, 2003), probabilistic latent semantic indexing (pLSI) (Hofmann, 1999), or relevance models (Lavrenko and Croft, 2001; Carterette and

Chandar, 2009). In our study, we are only interested in unsupervised clustering techniques although other techniques might be used. When considering cluster–based diversification methods, each cluster is assumed to represent a sub–topic of the information need, and thus diversification is achieved by retrieving documents belonging to different clusters.

Intuitively, the ranking strategies belonging to the first paradigm do not explicitly identify the sub–topic to be covered. Inter–document relationships, often based on statistical features of the documents, are used when accounting for novelty and diversity. Even though the number of relevant sub–topics is of importance when evaluating retrieval strategies, there is no explicit estimation of the number of sub–topics. In order to maximise the number of sub–topics, the retrieval systems based on this paradigm mainly rely on retrieving relevant documents that contain low redundant information, i.e. they are different enough to each other that they might refer to different sub–topics.

By contrast, clustering–based diversification methods, interleaving documents belonging to different clusters, explicitly cover all possible (or identified) sub–topics in top rankings. Nevertheless, this paradigm lacks an explicit model of relevance: how to combine relevance and the information regarding sub–topic clusters is still an open challenge. Documents that are selected after the clustering process to build the ranking can contain lots of new information, but they might be non–relevant; or even contain relevant information, but this might be redundant in the context of the document ranking. Consequently, ranking documents by either paradigm might produce unsatisfactory results.

Documents can cover several sub–topics (clusters), and these might even to some extent overlap. For example, the topic ‘Victoria’ can contain a set of documents regarding people (Queen Victoria or Victoria Beckham) and places (the state in Australia or the memorial in London) from a topical point of view. A document, clustered into the sub–topic ‘Victoria Beckham’, can contain information about her appearance in the Victoria state, Australia. Motivated by these considerations, we aim to alleviate the deficiencies of the two paradigms by combining their merits together. To the best of our knowledge, no empirical study has been performed comparing and integrating these two ranking paradigms in the context of sub–topic retrieval.

In this chapter, we propose a novel ranking strategy which explicitly models possible sub–topics and promotes diversity among documents in a ranking. Our strategy enables the development of a variety of algorithms for integrating statistical similarity, diversity structure, conveyed by clusters, and document dependencies. The paradigm relies on the cluster hypothesis (van Rijsbergen, 1979; Hearst and Pedersen, 1996), which assumes that relevant documents tend to be more similar to each other than non–relevant documents. When clustering documents, topically coherent groups of documents are formed by encoding possible aspects (i.e. sub–topics) of a more general topic. Relevance and diversity evidence is then merged in a ranking function (e.g. MMR), which selects documents from different clusters. The result is a document ranking that covers all the identified sub–topics, conveying at the same time relevant and novel information. This ranking approach provides insights for integrating two ranking paradigms for sub–topic retrieval: this

can generally be applied to any method for estimating sub–topic classes (e.g. LDA, pLSI, relevance models) and for considering document dependencies (e.g. QPRP, PT, MMR). This study aims to investigate the performance gained from the ranking strategy produced by the integration of two ranking models, not to propose a new specific ranking model.

The contributions of this chapter are to:

- analyse and discuss the current state–of–the–art methods for diversity–aware information retrieval;
- investigate a new ranking strategy which is able to model sub–topics by means of clustering techniques and promote diversity among documents in a document ranking;
- conduct an empirical study comparing state–of–the–art ranking models for sub–topic retrieval (e.g. MMR, PT, static and dynamic clustering) against the integration models we introduce based on MMR and clustering;
- discuss the results of this study, and show that our proposed solutions outperform state–of–the–art strategies.

The chapter is structured as follows. In Section 15.2 we frame the sub–topic retrieval problem and present existing strategies for encoding novelty and diversity in document rankings. Next, we illustrate our approach based on clustering that considers sub–topic evidence when ranking using an MMR–inspired approach (Section 15.3). In Section 15.4, we present the methodology of our empirical investigation, which employs the imageCLEF 2009 Photo collection (Paramita et al, 2009), an image collection suited for the sub–topic retrieval task. The results obtained in the empirical investigation are illustrated and discussed in Section 15.5, while Section 15.6 concludes the paper stating the major contributions of our work in the light of the results obtained on the ImageClef 2009 Photo Retrieval Task collection together with lines of future investigation.

## 15.2 Background and Related Work

### 15.2.1 Sub–topic Retrieval

Conventional IR systems employ the PRP to rank documents with respect to the user’s queries. Systems based on the PRP ignore interdependencies among documents ranked in the search results. These systems are generally appropriate when there are very few relevant documents and high–recall is required. An example of this situation is topic distillation in Web search, where a typical user wishes to find very few relevant core websites rather than every relevant Web page (Soboroff, 2004).

The assumption of independence in document relevance assessments that accompanied IR evaluation since the adoption of the Cranfield paradigm and on which the PRP is based, have recently been questioned. This generated a spate of work, not

only on ranking functions that account for interdependent document relevance, such as MMR (Carbonell and Goldstein, 1998), PT (Wang and Zhu, 2009), QPRP (Zuccon and Azzopardi, 2010), but also with respect to test collections, evaluation measures, and retrieval tasks (Zhai et al, 2003; Clarke et al, 2008; Paramita et al, 2009). In fact, the relaxation of the independence assumption requires test collections to encode information about relevance dependencies between documents and measures are developed so as to account for such relationships. Research on novelty and diversity document ranking flowered from these needs, and a new retrieval task, called sub–topic document retrieval (or alternatively, diversity retrieval, novelty and diversity retrieval, facets retrieval), has been introduced. A number of collections has been realised for this task, including the one based on the Text REtrieval Conference (TREC) 6, 7, 8 interactive track (Zhai et al, 2003) and ImageCLEF 2008 and 2009 photo retrieval task collections (Sanderson, 2008; Paramita et al, 2009). In these collections, query topics induce sub–topics (also called facets, nuggets, aspects, intentions): each document contains zero or more sub–topics of a query topic, and one sub–topic can be contained in one or more documents. Note that if a document contains at least one query sub–topic, then it is relevant. No assumptions are made about the extent of the relevance of a document, i.e. grade of relevance: even if a document covers more sub–topics than another, the former yet might not be more relevant than the latter. Specifically, in all the collections produced for this retrieval task until today, relevance is treated as a binary feature: a document is either relevant or not, although it contains one or more sub–topics.

The aim of the task is to rank documents such that all the sub–topics associated with a query topic are covered as early in the ranking as possible, and sub–topics are covered with the least redundancy possible. Thus, the requirement that document rankings should cover all the sub–topics is greater than that documents should be relevant, since a document that covers a sub–topic is also relevant, but a list of relevant documents covers just one sub–topic. As a matter of fact, however, pure relevance ranking is unsuitable in this task.

This task resembles real situations. Often, in fact, there are an enormous number of potentially relevant documents containing largely similar content, resulting in partially or nearly duplicate information being presented within the document ranking. Secondly, in a large number of cases users pose a query for which the result set contains very broad topics related to multiple search facets, or has however multiple distinct interpretations. The query ‘London’ represents an example of a broad query. This might refer to ‘London weather’, ‘London transport’, ‘London people’, ‘London travel’, etc. The query ‘Chelsea’ represents an example of an ambiguous query that might be interpreted as ‘Chelsea Clinton’, ‘Chelsea football club’, or ‘Chelsea area in London’ etc. These are examples of situations where IR systems have to provide a document ranking that minimises redundant information and covers all the possible search facets (sub–topics).

Clarke et al (2008) identify the precise distinction between the concepts of *novelty* and *diversity* in information retrieval. Novelty is the need to avoid *redundancy* in search results, while diversity is the need to resolve queries’ *ambiguity*. A popular approach for dealing with the redundancy problem is to provide diverse results in

response to a search adopting an explicit ranking function, which usually requires a tuning of a user parameter. For example, MMR (Carbonell and Goldstein, 1998) and the Harmonic measure (Smyth and McClave, 2001) combine similarity and novelty in a unique ranking strategy. On the other hand, a traditional approach for coping with poorly specified or ambiguous queries relies on promoting diversity. This is motivated by the fact that the chances to retrieve relevant results can be maximised if results containing information from different query interpretations are presented within the document ranking.

### 15.2.2 The Probability Ranking Principle

The PRP (Robertson, 1977) is a well accepted ranking strategy that suggests presenting documents according to decreasing probability of document relevance to the user's information need; and the relevance of one document is considered independent from the rest of the documents. Formally, given a query  $q$ , if  $P(x_i)$  is the relevance estimation for document  $x_i$  and  $S(x_i, q)$  is the similarity function employed for producing such estimation, then the PRP suggests to present at rank  $J + 1$  a document  $d$  such that:

$$PRP_{J+1} \equiv \operatorname{argmax}_{x_i \in I \setminus J} [p(x_i)] \approx \operatorname{argmax}_{x_i \in I \setminus J} [S(x_i, q)] \quad (15.1)$$

where  $I$  is the set of results retrieved by the IR system;  $J$  is the set formed by the documents ranked until iteration  $J$ ;  $x_i$  is a candidate document in  $I \setminus J$ , which is the set of documents that have not been ranked yet.

In the PRP, a document's relevance judgements are assumed independent and thus no relationship between documents is explicitly modelled in the ranking function. This is a known limitation of the PRP and, although it does not affect the optimality of the ranking principle for tasks such as ad hoc retrieval, it is the cause of the sub-optimality of the PRP in particular scenarios, such as sub-topical retrieval.

### 15.2.3 Beyond Independent Relevance

**Maximal Marginal Relevance:** Several techniques have been proposed for sub-topical retrieval. A simple method to address diversity between documents is that of MMR in set-based IR (Carbonell and Goldstein, 1998). Using a tuneable parameter, this ranking method balances the relevance between a candidate document and a query, e.g. the probability of relevance, and the diversity between the candidate document and all the documents ranked at previous positions. The ranking is linearly produced by maximising relevance and diversity scores at each rank. The MMR strategy is characterised by the following ranking function:

$$MMR_{J+1} \equiv \operatorname{argmax}_{x_i \in I \setminus J} [\lambda S(x_i, q) + (1 - \lambda) \max_{x_j \in J} D(x_i, x_j)] \quad (15.2)$$

where  $x_j$  is a document in  $J$ , i.e. the set of documents that have been ranked already. The function  $S(x_i, q)$  is a normalised similarity metric used for document retrieval, such as the cosine similarity, whereas  $D(x_i, x_j)$  is a diversity metric. A value of the parameter  $\lambda$  greater than 0.5 assigns more importance to the similarity between document and query rather than to novelty/diversity. Conversely, when  $\lambda < 0.5$ , novelty/diversity is favoured over relevance.

In our work, when operationalising MMR, we modify how the diversity function impacts on the ranking: we substitute the function  $\max$  with  $\operatorname{avg}$ , which returns the average dissimilarity value between all pairs of  $x_i$  and  $x_j$ , instead of their largest value. To compute the dissimilarity between documents, we resort to estimating their similarity and then we revert this estimation. Specifically, the cosine function is used as a similarity measure between documents' term vectors obtained using the BM25 weighting schema. Since its similarity values range between  $-1$  and  $1$ , we can estimate the dissimilarity by the following formula:

$$\operatorname{avg}_{x_j \in J} D(x_i, x_j) = \frac{\sum_{j=1}^J (-S(x_i, x_j))}{J} \quad (15.3)$$

In Figure 15.1a we depict the document selection procedure suggested by MMR. In the figure, we simulate the possible clusters of documents that identify the sub–topics covered by those documents. Documents inserted in the ranking following the MMR strategy might come from the same group of sub–topics (i.e.  $x_1$  and  $x_3$ ), colliding with what is required in the sub–topic retrieval task.

**Portfolio Theory:** Wang and Zhu (2009) suggested to rank documents according to a paradigm proposed to select stocks in the financial market, PT. In the IR scenario diversification is achieved using PT by reducing the risk associated with document ranking. The intuition underlying PT is that the ideal ranking order is the one that balances the relevance of a document against the level of its risk or uncertainty (i.e. variance). Thus, when ranking documents, relevance should be maximised whilst minimising variance. The objective function that PT optimises is:

$$PT_{J+1} \equiv \operatorname{argmax}_{x_i \in I \setminus J} \left( p(x_i) - bw_{x_i} \delta_{x_i}^2 - 2b \sum_{x_k \in J} w_{x_k} \delta_{x_i} \delta_{x_k} \rho_{x_i, x_k} \right) \quad (15.4)$$

where  $b$  represents the risk propensity of the user,  $\delta_{x_i}^2$  is the variance associated to the probability estimation of document  $x_i$ ,  $w_{x_i}$  is a weight expressing the importance of the rank position, and  $\rho_{x_i, x_k}$  is the correlation between document  $x_i$  and document  $x_k$ .

In summary, intuitively MMR and PT have a similar underlying schema for combining relevance and diversity. One common component of their ranking functions is the estimation of the probabilities of relevance. In both methods, the relevance estimation is balanced by a second component, which captures the degree of diversity

between the candidate document and the ranking. In the empirical study we present in Section 15.4, we implemented both strategies and compared them to the novel paradigm we propose.

## 15.3 Document Clustering and Inter-Cluster Document Selection

### 15.3.1 Re-examining Document Clustering Techniques

It has been hypothesised that ‘closely associated documents tend to be more relevant to the same requests’ (van Rijsbergen, 1979). Similarly, in our work we hypothesise that clusters obtained considering the documents relevant to a user’s information need have the potential to represent different sub-topics of the more general topic the user is interested in. Thus, clustering using unsupervised learning models extracts meaningful and representative information from documents, that can be used to model sub-topical diversity. The set of documents contained in each cluster is assumed to resemble what users perceive as a sub-topic. We then believe that incorporating evidence drawn from clusters of similar documents can enhance the performance of IR systems in the sub-topic retrieval task.

Although clustering can group documents containing similar contents, we do not intend to use clustering in isolation, in particular when selecting documents from such clusters. We hypothesise that clustering techniques combined with suitable criteria for document selection can improve the performances of systems in the sub-topic retrieval task. Regardless of the clustering technique used, strategies following the cluster-based paradigm can be characterised by how documents are selected from the clusters in order to output the final document ranking. In the following, two common approaches are revised.

The first approach, which is directly inspired by the cluster hypothesis, attempts to retrieve documents that are more similar to each other at higher ranks. Kurland and Lee (2004) propose a method, called the *interpolation algorithm*, to compute a retrieval score by mixing the evidence obtained from clusters together with document relevance estimations. The retrieval score of a candidate document  $d_i$  given this approach is calculated as:

$$\hat{p}(x_i, q) = \lambda p(x_i, q) + (1 - \lambda) \sum_{t \in X} p(c_j, q) p(x_i, c_j) \quad (15.5)$$

where  $\lambda$  indicates the degree of emphasis on individual document information. In our study, we assume that  $p(a, b)$  is the similarity between objects  $a$  and  $b$ <sup>1</sup>. Note that setting  $\lambda = 0$  returns documents within the cluster with highest similarity to the query, i.e. the cluster with the highest  $p(c_j, q)$ . We refer to this approach as **Interp(.)**.

---

<sup>1</sup> These can be queries, documents, or clusters.

In the second approach we assume that each cluster represents a different sub–topic. Thus, to cover the whole set of sub–topics all the clusters have to be chosen at early ranks. In (van Leuken et al, 2009), three clustering methods with a dynamic weighting function are exploited to visually diversify image search results. Only representatives of visual–based clusters are presented to the users with the aim to facilitate faster browsing and retrieval. A similar work has been pursued in (Ferecatu and Sahbi, 2008), where the ranking is formed by selecting documents from clusters in a round–robin fashion, i.e. assigning an order to the cluster and selecting a document when cycling through all clusters. Cluster representatives are selected according to the order of the documents and are added to clusters<sup>2</sup>. The same approach may be applied to different clustering algorithms, i.e. k–means, Expectation–Maximisation (EM), Density–Based Spatial Clustering of Applications with Noise (DBSCAN).

Once sub–topical clusters are formed, several approaches can be employed to select a cluster representative. Deselaers et al (2009) propose selecting within each cluster the document that is most similar to the query, whereas Zhao and Glotin (2009) suggest choosing the document with the lowest rank within each cluster of the top retrieved results. In (Leelanupab et al, 2009; Urruty et al, 2009), the medoid<sup>3</sup> is assumed to be the best cluster representative. Finally, Halvey et al (2009) propose selecting the document that is most similar to other members of the selected cluster. In summary, these approaches ensure that documents are retrieved from all the clusters, and thus from all the sub–topics if these are correctly captured by the clustering process. However, document relevance and redundancy are not accounted for after clustering. Furthermore, despite the selection of documents from clusters according to their probability of relevance, documents at early ranks can contain duplicate information.

As a result, the top ranked documents might still be similar or highly correlated to each other. For example, as shown in Figure 15.1b, the distances of documents  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  selected using the clusters’ medoids are constant and far away from the query  $q$ , in particular  $x_3$  and  $x_4$ . Furthermore, if the closest documents to the query were to be selected, then the result will be documents that are close to each other, in particular when the query lays in the centre of the document space, which is surrounded by the clusters.

### 15.3.2 Clustering for Sub–topic Retrieval

As we have illustrated in the previous section, no current *cluster–based* method for sub–topic retrieval addresses novelty and relevance at the same time. Motivated by this consideration, this chapter investigates the effect of integrating intra–list dependence ranking models, i.e. ranking strategies that account for dependencies amongst ranked documents, and topic–oriented/cluster–based models, i.e. strategies that di-

<sup>2</sup> This is possible because the clustering algorithm in (Ferecatu and Sahbi, 2008) builds clusters iteratively by first selecting the centre of a cluster and then gathering its members.

<sup>3</sup> The document closest to the centroid of the cluster.

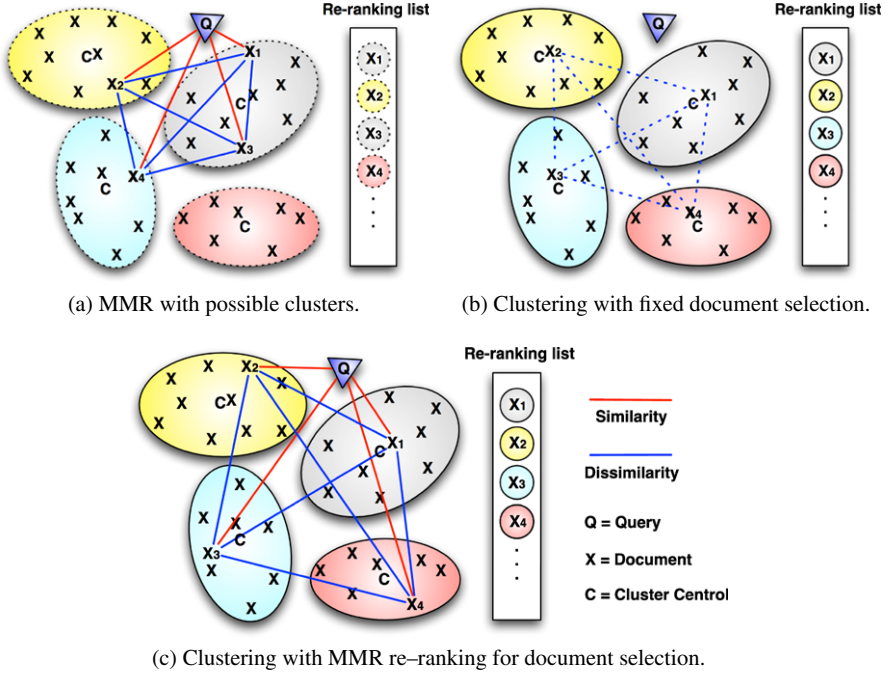


Fig. 15.1: Re-ranking methods for promoting diversity.

vide documents into sub-topic classes and consider these when ranking. Specifically, we propose two simple strategies that follow this idea, and evaluate them in the context of sub-topic retrieval. In particular, document dependencies can be exploited during the selection of representatives from sub-topic classes, obtained by employing any of the latter models. Figure 15.1c depicts the result of this approach, in which documents  $x_1, x_2, x_3$ , and  $x_4$  are selected according to particular sub-topics, thus addressing diversity in the document ranking. We do not focus on the retrieval and relevance estimation, but we assume to have a reliable function that is able to provide an initial set of relevant documents. We suggest clustering these documents and then ranking the clusters according to the average relevance of the documents contained in each cluster. Given a query  $q$  and a cluster  $c_k$ , the average cluster relevance is defined as:

$$S_{avg}(c_k, q) = \frac{1}{I_k} \sum_{i=1}^{I_k} s(x_{k,i}, q) \tag{15.6}$$

where  $I_k$  is the number of documents in  $c_k$  and  $X = \{x_1, \dots, x_n\}$  is the initial set of relevant documents. Average cluster relevance is employed for ordering the clusters;

---

**Algorithm 15.1** Intra-list dependency re-ranking (using MMR) on the evidence gathered from clusters.

---

**Require:**  $q$ , a user query

**Require:**  $C = \{c_1, c_2, c_3, \dots, c_k\}$ , set of clusters  $k$  ranked according to average cluster relevance  $S_{avg}(c_k, q)$

**Require:**  $X_k = \{x_{k,1}, x_{k,2}, x_{k,3}, \dots, x_{k,n}\}$ , set of retrieved documents  $x$  within cluster  $c_k$

**Require:**  $j = 0$ , where  $j$  is the number of documents that has been already ranked

**Require:**  $maxDocs$ , the maximum number of retrieved documents

$J_0 = \{\}$

**while**  $j \leq maxDocs$  **do**

**if**  $j = 0$  **then**

$J_0 = \operatorname{argmax}_{x_{k,n} \in X_k \setminus J} [S(x_{k,n}, q)]$

**else**

$J_j = J_{j-1} \cup \operatorname{argmax}_{x_{k,n} \in X_k \setminus J} [\lambda S(x_{k,n}, q) + (1 - \lambda) \operatorname{avg}_{x_j \in J} D(x_{k,n}, x_j)]$

**end if**

$j = j + 1; k = k + 1$

**if**  $k \geq j$  **then**

$k = 0$

**end if**

**end while**

**return**  $J_j = \{x_1, x_2, x_3, \dots, x_j\}$ , a set of re-ranked documents to present to the user

---

then a round–robin approach that follows the order suggested by average cluster relevance is used in order to select individual documents within the clusters. To select a specific document within each cluster, we have to employ an intra–list dependency–based model: in our empirical study we choose to use MMR, for its simple formulation. In this step, alternative ranking functions may be employed. The complete algorithm is outlined in Algorithm 15.1: this is the same algorithm that has been implemented to produce the results reported in our empirical investigation.

## 15.4 Empirical Study

To empirically validate our approach and contrast it to state–of–the–art ranking strategies for sub–topic retrieval, we adopted the ImageCLEF 2009 photo retrieval collection (Paramita et al, 2009). This collection is composed of almost 500,000 images from the Belga News Agency. A text caption is associated with each image; the average length of the textual captions is 36 terms, while the total numbers of unique terms in the collection is over 260,000. Textual captions have been indexed using Terrier<sup>4</sup>, which also served as a platform for developing the ranking strategies using Java. Before indexing the captions, we removed standard stop–words (van Rijsbergen, 1979) and applied Porter stemming. Low–level descriptors were not considered as this year’s topics focus on topical, rather than visual, diversity. Moreover, the goal

<sup>4</sup> <http://ir.dcs.gla.ac.uk/terrier/>

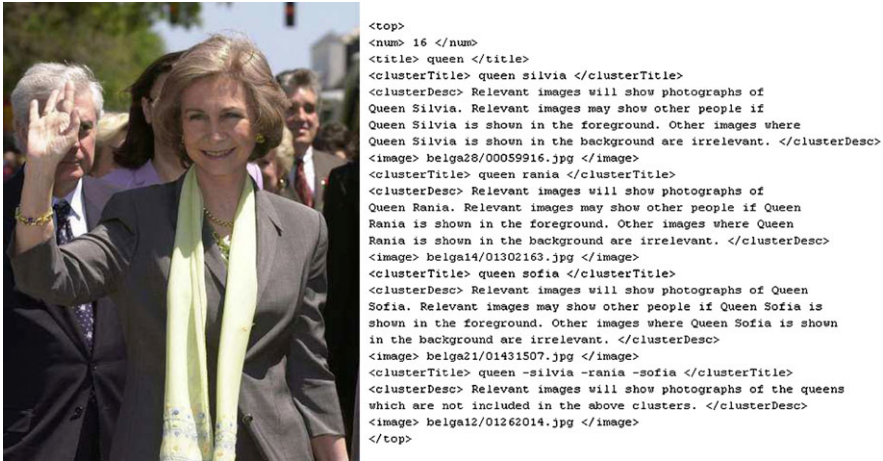


Fig. 15.2: Example of ImageCLEF 2009 database entry with sub-topic image (left) and query topic format (right).

of our study is to determine whether the integration of two diversity-aware ranking models is valid, and thus there is no major concern in ruling out visual features from the empirical investigation. Query topics have been similarly processed to the text captions. We used the set of 50 available topics, consisting of topic titles, cluster titles, cluster descriptions, and image examples. We employ only the topic titles, so as to simulate the situation where a user posts a broad or ambiguous query. Finally, we used the sub-topic judgements associated with each topic that are provided with the collection; an example of topic and image with related sub-topics is shown in Figure 15.2.

Okapi BM25 has been used to estimate document relevance given a query: its parameters have been set to standard values (Robertson et al, 1995). The same weighting schema has been used to produce document term vectors that are subsequently employed by re-ranking strategies to compute similarity/correlation. In preliminary results we have observed that this approach returns higher precision values, compared with alternative strategies, e.g. TF-IDF weighting. We experiment with several ranking lengths, i.e. 100, 200, 500, and 1,000, meaning that all the documents retrieved at ranks lower than these thresholds are discarded. In this chapter we report results for ranking up to 500 documents. Other ranking thresholds have shown similar results, and are not reported here.

Once estimates of document relevance are obtained using Okapi BM25, we produce an initial document ranking according to the probability ranking principle, i.e. we order documents with respect to decreasing probability of relevance. We denote this run with **PRP**, and it represents the baseline for every re-ranking strategy. Fur-

thermore, the initial document ranking obtained using the PRP is used as input of the re–ranking functions. The runs obtained by implementing the maximal marginal relevance heuristic and the portfolio theory approach are denoted with **MMR** and **PT**, and they represent the state–of–the–art strategies for interdependent document relevance in the context of this investigation.

For MMR, we investigated the effect on retrieval performances of the hyper–parameter  $\lambda$  by varying it in the range <sup>5</sup> [0,1). The ranking function of MMR has been instantiated as discussed in Section 15.2.3.

When testing PT, we set  $b$ , the risk propensity parameter, as ranging from  $\pm 1$  to  $\pm 9$ ; we treat the variance of a document as a parameter that is constant with respect to all the documents, similarly to (Wang and Zhu, 2009). We experiment with variance values  $\delta^2$  ranging from  $10^{-1}$  to  $10^{-9}$ , and select the ones that achieve best performances in combination with the values of  $b$  through a grid search of the parameter space. The correlation between textual captions is computed as Pearson’s correlation between the term vectors associated to the textual captions themselves.

Regarding the runs based on the topic–oriented paradigm, we adopt two different static and dynamic clustering algorithms: **k-means** and expectation maximization (**EM**), although alternative strategies may be suitable. For each query, the number of clusters required in k-means was set according to sub–topic relevance judgements for that query. In contrast, we allow the EM algorithm to determine the optimal number of clusters using cross validation, and set the minimum expected number of clusters using the sub–topic relevance judgements. After clusters are formed, documents are ranked according to techniques for selecting documents within clusters as illustrated in Section 15.3, specifically:

**Interp(.)**: selects documents that maximise the interpolation algorithm for cluster–based retrieval;

**PRP(.)** : selects documents with the highest probability of relevance in the given clusters;

**Mediod(.)**: selects the medoids of the given clusters as cluster representatives;

**MMR(.)** : selects documents according to maximal marginal relevance, as an example of a strategy based on an interdependent document relevance model.

Techniques that implement PRP(.) and Medoid(.) do not require any parameter tuning, whereas when instantiating Interp(.) and MMR(.), we varied their hyper–parameters in the range [0,1), and selected the value that obtained the best performance. In total, the combination of clustering algorithms and document selection criteria forms eight experimental runs that we tested in this study, i.e. Interp(k-means), PRP(k-Means), Medoid(k-means), MMR(k-means), Interp(EM), PRP(EM), Medoid(EM), and MMR(EM).

We employed three measures to assess the effectiveness of different ranking strategies in sub–topic retrieval. The first measure, called  $\alpha$ –**NDCG**, extends the normalised discounted cumulative gain to the case of the sub–topic retrieval task; the parameter  $\alpha$  ranges between 0 and 1 and directly accommodates novelty and

---

<sup>5</sup> We excluded the value  $\lambda = 1$ , since MMR’s ranking function would be equivalent to that of PRP.

diversity (Clarke et al, 2008). We set  $\alpha = 0.5$ , as it is common practice (Clarke et al, 2009b): intuitively, this means that novelty and relevance are equally accounted for in the measure. Novelty and rank biased precision (NRBP) (Clarke et al, 2009a) integrate nDCG, rank-biased precision (RBP) and intention aware measures in order to model the decreasing interest of the user examining documents at late rank positions. Sub-topic recall (S-R) (Zhai et al, 2003) monitors sub-topic coverage in the document ranking.

## 15.5 Results

In Table 15.1 we report the results obtained by the instantiations of the ranking strategies considered in our empirical investigation and evaluated them using  $\alpha$ -NDCG@10,  $\alpha$ -NDCG@20, NRBP, and S-recall. Due to the presence of varying parameters that require tuning, we only report the best results obtained by each strategy with respect to  $\alpha$ -NDCG@10. Percentage improvements over the PRP of each re-ranking strategy are reported in the table. The instantiations of the approaches we propose in this study, i.e. MMR(k-means) and MMR(EM), underlined in the table, provide an example of the results the integration approach, based either on static or dynamic clustering, and MMR, can achieve. Statistical significance against MMR and PT using a t-test is calculated for each of the eight runs reported in the lower part of Table 15.1, and it is indicated with \*, w.r.t. MMR, and †, w.r.t. PT.

The results suggest that the integration of either static or dynamic clustering techniques with interdependent document ranking approaches can improve the performance in sub-topic retrieval. In particular, the percentage improvements of MMR(k-means) and MMR(EM) are greater than the one obtained by their peers in all the evaluation measures, except in NRBP, for which however a consistent trend can not be extracted. Furthermore, it can be observed that even applying the integration paradigm on a simple lightweight clustering algorithm such as k-means, which can be executed in runtime, can increase the retrieval performance when compared to MMR or k-means alone.

Our empirical investigation also suggests that selecting clusters in a round-robin fashion when ranking, as is done in PRP(k-means), PRP(EM), or Medoid(k-means), Medoid(EM), outperforms other policies, such as the one implemented by Interp(.). This result is consistent for all the investigated measures. Note that ranking documents according to Interp(.) may result in documents from the same cluster being ranked consecutively: this might be the case when the probabilities of cluster relevance and of the document being in the specific cluster are high. In addition, the results suggest that the runs based on the topic-oriented paradigm produce better rankings than the ones based on the interdependent document relevance paradigm (i.e. MMR and PT) in terms of  $\alpha$ -NDCG@10, that has been used as an objective function for parameter tuning.

In summary, the results show that integrating the two paradigms for sub-topic retrieval based on interdependent document relevance and topic-oriented models can

Table 15.1: Sub–topic retrieval performances obtained in the ImageCLEF 2009 photo retrieval collection. Percentage improvements refer to the PRP baseline. Parameters are tuned with respect to  $\alpha$ –NDCG@10, in particular: MMR ( $\lambda = 0.6$ ), PT ( $b = 9$ ,  $\delta^2 = 10^{-3}$ ), Interp(k–means) ( $\lambda = 0.8$ ), MMR(k–means) ( $\lambda = 0.8$ ), Interp(EM) ( $\lambda = 0.9$ ), and MMR(EM) ( $\lambda = 0.7$ ). The best performance improvements are highlighted in bold, and statistical significance at 0.05 level, computed using a t–test, against MMR and PT are indicated by \* and † respectively.

Model	$\alpha$ -NDCG@10	$\alpha$ -NDCG@20	NRBP	S-R@10
<b>PRP</b>	0.425	0.467	0.270	0.542
<b>MMR</b>	0.484	0.516	0.288	0.661
	+13.88%	+10.49%	+6.67%	+21.90%
<b>PT</b>	0.470	0.511	0.287	0.629
	+10.59%	+9.42%	+6.30%	+16.09%
<b>Interp(K-Mean)</b>	0.448	0.475	0.302	0.524*†
	+5.41%	+1.71%	+11.85%	-3.32%
<b>Medoid(K-Mean)</b>	0.463	0.490	0.291	0.591*
	+8.94%	+4.93%	+7.78%	+8.93%
<b>PRP(K-Mean)</b>	0.486	0.515	0.309	0.617
	+14.35%	+10.28%	+14.44%	+13.87%
<b>MMR(K-Mean)</b>	0.491	0.520	0.302	0.655
	+15.53%	+11.35%	+11.85%	+20.83%
<b>Interp(EM)</b>	0.457	0.486	0.311	0.532*
	+7.53%	+4.07%	+15.19%	-1.84%
<b>Medoid(EM)</b>	0.497	0.524	<b>0.320</b> †	0.646
	+16.94%	+12.21%	+18.52%	+19.11%
<b>PRP(EM)</b>	0.502†	0.536	0.314	0.670
	+18.12%	+14.78%	+16.30%	+23.61%
<b>MMR(EM)</b>	<b>0.508</b> †	<b>0.539</b> †	0.311	<b>0.681</b> †
	+19.53%	+15.42%	+15.19%	+25.59%

deliver better performance than state–of–the–art ranking strategies. In three out of four measures, our best approach based on the integration paradigm, i.e. MMR(EM), outperforms state–of–the–art approaches with statistical significance against our instantiation of PT. Despite the integration–based strategies providing less performance increments than other re–ranking approaches with respect to NRBP, the difference is very limited and might be related to the settings of the parameters internal to NRBP. Furthermore, it is difficult to quantify how this small difference in NRBP affects the user.

## 15.6 Conclusions

Diversity with respect to the sub–topics of the query topic is a highly desired feature for generating satisfying search results, in particular when there is a large number of documents containing similar information, or when a user enters a very broad or am-

biguous query. Common diversity-based ranking approaches are devised on the basis of interdependent document relevance or topic-oriented paradigms. In this chapter, state-of-the-art strategies for diversity-aware IR are reviewed and discussed. We propose a new ranking approach, which incorporates two ranking paradigms with the aim to explicitly model sub-topics and reduce redundancy in document ranking simultaneously. An empirical investigation was conducted using the ImageCLEF 2009 photo retrieval task collection, where we contrast state-of-the-art approaches against two instantiations of the integration approach we propose. Maximal marginal relevance and portfolio theory for IR are examined as examples of the interdependent document relevance paradigm, whilst  $k$ -means and EM clustering methods are employed to explicitly model sub-topics. Various criteria for selecting documents within clusters are investigated in our study, and their performance is compared. The interpolation algorithm assumes that relevant documents tend to be more similar to each other; whereas the selection methods based on cluster representatives or PRP stem from the hypothesis that clusters can represent the sub-topics of a query. Parametric ranking functions are tuned with respect to  $\alpha$ -NDCG@10, which is used to measure retrieval effectiveness in the sub-topic retrieval tasks. We also evaluate the strategies in terms of NRBP and S-recall.

The results of our empirical investigation suggest that ranking strategies built upon the integration of MMR and EM clustering significantly outperform all other approaches. Furthermore, we show that the integration intuition can be ideally applied to any clustering algorithm. The comparison between interdependent document relevance and topic-oriented paradigms suggests that the cluster-based retrieval strategies perform better than the former in sub-topic retrieval. With respect to our study, the interpolation algorithm in cluster-based retrieval is not suitable for results diversification, while still being suitable for the ad hoc retrieval task (Kurland and Lee, 2004). The round-robin policy for selecting clusters performs consistently better than the interpolation strategy; furthermore, selecting documents within clusters using the PRP is better than doing so by using cluster representatives such as medoids.

In summary, the integration approach effectively improves diversity performance for sub-topic retrieval. Further investigation will be directed towards the empirical validation of our integration approach on other collections for sub-topic retrieval, such as TREC 6, 7, 8 interactive and ClueWeb 2009. Furthermore, image low-level features can be employed to refine the results from text clustering since they can enhance visual diversity in addition to topical diversity.

**Acknowledgements** The authors are grateful to Chales Clarke for advise on the implementation of the evaluation measures used in this work. Teerapong Leelanupab would like to thank the Royal Thai Government for providing financial support. Guido Zuccon is funded by the EPSRC Renaissance project (EP/F014384/1).

## References

- Agichtein E, Brill E, Dumais S (2006) Improving web search ranking by incorporating user behavior information. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp 19–26
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022
- Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, ACM press, pp 335–336
- Carterette B, Chandar P (2009) Probabilistic models of ranking novel documents for faceted topic retrieval. In: Proceeding of the 18th ACM conference on information and knowledge management, pp 1287–1296
- Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, pp 659–666
- Clarke CL, Kolla M, Vechtomova O (2009a) An effectiveness measure for ambiguous and under-specified queries. In: Proceedings of the 2nd International Conference on Theory of Information Retrieval, pp 188–199
- Clarke CLA, Craswell N, Soboroff I (2009b) Overview of the TREC 2009 Web Track. In: Proceedings of the Text REtrieval Conference (TREC–2009)
- Deselaers T, Gass T, Dreuw P, Ney H (2009) Jointly optimising relevance and diversity in image retrieval. In: Proceeding of the ACM International Conference on Image and Video Retrieval
- Eisenberg M, Berry C (2007) Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science and Technology* 39(5):293–300
- Ferecatu M, Sahbi H (2008) TELECOM ParisTech at ImageCLEFphoto 2008: Bi-modal text and image retrieval with diversity enhancement. In: Working Notes of CLEF 2008
- Gordon MD, Lenk P (1999a) A utility theoretic examination of the probability ranking principle in information retrieval. *Journal of the American Society for Information Science and Technology* 42(10):703–714
- Gordon MD, Lenk P (1999b) When is the probability ranking principle suboptimal. *Journal of the American Society for Information Science and Technology* 43(1):1–14
- Halvey M, Punitha P, Hannah D, Villa R, Hopfgartner F, Goyal A, Jose JM (2009) Diversity, assortment, dissimilarity, variety: A study of diversity measures using low level features for video retrieval. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, pp 126–137
- Hearst M, Pedersen J (1996) Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, pp 76–84
- Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pp 50–57
- Huang J, Kumar SR, Zabih R (1998) An automatic hierarchical image classification scheme. In: Proceedings of the sixth ACM international conference on Multimedia, pp 219–228
- Kurland O, Lee L (2004) Corpus structure, language models, and ad hoc information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, pp 194–201
- Lavrenko V, Croft WB (2001) Relevance based language models. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM press, pp 120–127

- Leelanupab T, Hopfgartner F, Jose JM (2009) User centred evaluation of a recommendation based image browsing system. In: Proceedings of the 4th Indian International Conference on Artificial Intelligence, pp 558–573
- van Leuken RH, Garcia L, Olivares X, van Zwol R (2009) Visual diversification of image search results. In: Proceedings of the 18th international conference on World Wide Web, pp 341–341
- Paramita ML, Sanderson M, Clough PD (2009) Developing a test collection to support diversity analysis. In: Proceedings of Redundancy, Diversity, and Interdependent Document Relevance workshop held at ACM SIGIR' 09
- van Rijsbergen CJ (1979) Information Retrieval, 2nd Ed. Butterworth
- Robertson SE (1977) The probability ranking principle in IR. *Journal of Documentation* 33:294–304
- Robertson SE, Walker S, Beaulieu MM, Gatford M (1995) Okapi at TREC 4. In: Proceedings of the 4th Text REtrieval Conference (TREC-4)
- Sanderson M (2008) Ambiguous queries: test collections need more sense. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp 499–506
- Smyth B, McClave P (2001) Similarity vs. diversity. In: Proceedings of the 4th International Conference on Case-Based Reasoning, pp 347–361
- Soboroff I (2004) On evaluating web search with very few relevant documents. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp 530–531
- Urruty T, Hopfgartner F, D. H, Elliott D, Jose JM (2009) Supporting aspect-based video browsing - analysis of a user study. In: Proceeding of the ACM International Conference on Image and Video Retrieval
- Wang J, Zhu J (2009) Portfolio theory of information retrieval. In: Proceedings of the 32nd annual international ACM SIGIR conference on Diversity, and Interdependent Document Relevance workshop, pp 115–122
- Zhai CX, Cohen WW, Lafferty J (2003) Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp 10–17
- Zhao ZQ, Glotin H (2009) Diversifying image retrieval by affinity propagation clustering on visual manifolds. *IEEE MultiMedia* 99(1)
- Zuccon G, Azzopardi L (2010) Using the quantum probability ranking principle to rank interdependent documents. In: Proceedings of the 32th European Conference on IR Research on Advances in Information Retrieval, pp 357–369