

# Technical Report: A Study of Ranking Paradigms and Their Integrations for Subtopic Retrieval

Teerapong Leelanupab, Guido Zuccon and Joemon M. Jose

School of Computing Science, University of Glasgow  
Glasgow, G12 8RZ, United Kingdom  
{kimm,guido,jj}@dcs.gla.ac.uk

**Abstract.** In this paper, we consider the problem of document ranking in a non-traditional retrieval task, called *subtopic retrieval*. This task involves promoting relevant documents that cover many subtopics of a query at early ranks, providing thus diversity within the ranking. In the past years, several approaches have been proposed to diversify retrieval results. These approaches can be classified into two main paradigms, depending upon how the ranks of documents are revised for promoting diversity. In the first approach subtopic diversification is achieved implicitly, by choosing documents that are different from each other, while in the second approach this is done explicitly, by estimating the subtopics covered by documents. Within this context, we compare methods belonging to the two paradigms. Furthermore, we investigate possible strategies for integrating the two paradigms with the aim of formulating a new ranking method for subtopic retrieval. We conduct a number of experiments to empirically validate and contrast the state-of-the-art approaches as well as instantiations of our integration approach. The results show that the integration approach outperforms state-of-the-art strategies with respect to a number of measures.

**Keywords:** Subtopic Retrieval, Subtopic Awareness, Interdependence Document Relevance, Diversity

## 1 Introduction

Presenting redundant information in a ranking is undesirable as users have to endure examining the same information repeatedly. This is the case for example in topic distillation, where a user wishes to find only a few high quality documents, rather than every relevant document [17]. In some contexts the user requires a broad view of a search topic, for instance because his information need is unclear or vague. In these situations, a retrieval system should provide a document ranking that covers several aspects (or subtopics) that the user might be interested in [22]. In real search scenarios, a document might be non-relevant if the user has already examined other documents containing similar information [4]. If this is the case, the utility of a document does depend upon which documents have been ranked in previous positions.

Although there is a clear need to account for the influence of previously ranked documents, traditional ranking approaches rely on the assumption that the relevance of a document is independent to other documents. This assumption is on the basis of the probability ranking principle (PRP) [16], where documents are ranked exclusively according to their probability of being relevant to a query. It is then likely that the results retrieved by the PRP address only a particular aspect of the information need [18]. In real search scenarios, however, the independent relevance assumption often does not hold and consequently ranking approaches that rely on it, such as the PRP, provide a suboptimal document ranking [9].

Many efforts have been devoted to overcome the limitations of the independent relevance assumption in document ranking. In parallel, several approaches have been devised so as to produce a document ranking that covers many different subtopics of the information need. These approaches can be thought of as two faces of the same coin: generally, diversifying a document ranking implies exploiting document dependencies, and vice versa when accounting for document dependencies (at relevance level) diversification can be achieved. Two different patterns can be recognised from the approaches suggested in the literature in order to achieve ranking diversification:

- **Interdependent document relevance paradigm.** When ranking documents, relationships between documents are considered by promoting documents that differ from each other. These approaches maximise, at each rank position, a function that depends upon both relevance estimates and documents relationships. The intuition underlying this is that novelty and diversity are achieved by ranking relevant documents containing information that has not yet been ranked. A similarity function is usually employed to estimate the novelty of a document (the less a document is similar to the ones already ranked, the more it carries novel information). Examples of heuristic or theoretically driven approaches that implement this paradigm are maximal marginal relevance (MMR) [2], which interpolates document relevance and documents relationships; portfolio theory (PT) [21], which combines relevance estimates and document correlations; and the quantum probability ranking principle [24], which implicitly captures dependencies between documents through quantum interference.
- **Subtopics aware paradigm.** The need of (subtopic) diversity can be achieved by estimating and modelling subtopics and then selecting documents within them. Regardless of document relevance, relationships between documents are employed to estimate subtopics. Many techniques can be applied to discriminate documents with respect to the possible subtopics they cover: examples are clustering [19], classification [11], latent Dirichlet allocation (LDA) [1], probabilistic latent semantic analysis (PLSA) [10], and relevance models [3]. Afterwards, result diversification is achieved by interleaving in a ranking the documents that belong to different estimated subtopics. Several criteria can be applied to select documents after the evidence about the estimated subtopics is obtained.

In this paper, we intend to determine which paradigm, and in turns which approach, performs best in the subtopic retrieval task. Furthermore, we investigate whether a new ranking approach can be devised so that we can integrate the merits of the two ranking paradigms, regardless of the choices of the similarity estimation function, the document dependency function, and the subtopic modelling algorithm. The intuition underlying the integration approach is as follows: if subtopics are estimated in a way that do not corresponds to the user’s common perception of subtopics, an interdependent document ranking strategy could assist in correctly ranking documents after the subtopic evidences are given. Possible subtopics are thus explicitly modelled; diversity among ranked documents is promoted and information overlapping (redundancy) is limited by selecting documents belonging to different estimated subtopics.

To the best of our knowledge, no empirical study has been performed comparing and integrating the two ranking paradigms in the context of subtopic retrieval. The empirical results we present in this study show that our integration approach improves the retrieval effectiveness (measured by  $\alpha$ -NDCG@10) of about 19.12% on three test collections when subtopics are appropriately estimated according to user’s judgements.

The rest of the paper is organised as follows. First, we introduce in Section 2 the task of subtopic retrieval and outline examples of approaches belonging to the two different ranking paradigms. Subsequently, we describe our integration approach in the Section 3. Section 4 describes the empirical study we perform in this investigation, while the results from the study are discussed in Section 5. The paper concludes in Section 6, where we summarise our contributions and suggest directions of future work.

## 2 Related work

### 2.1 Subtopic retrieval

The need of accounting for document dependencies, and thus ultimately for diversity, when ranking documents was already recognised by Goffman in the 60s [8]. In his work, Goffman pointed out that the query-document relationship is not sufficient to determine the “relevance” of a document when relevance is defined as a measure of information. Instead, the relevance measure should include the relationships between a document and the documents ranked at previous positions.

A traditional ranking criterion used in information retrieval, the PRP, discards the dependencies between assessments of document relevance. If this route is followed, the ranking that is generated might be suboptimal for particular user’s needs [9]. However, conventional evaluation measures (e.g. precision, recall) and retrieval tasks (e.g. ad-hoc retrieval) ignore the fact that a relevant piece of information is retrieved more than once or that retrieved documents belong to only one of a number of possible subtopics of the information need. The subtopic retrieval task attempts to remedy this fallacy by the introduc-

tion of different evaluation contexts and measures [4, 22], which in turns require different ranking approaches to achieve ranking optimality.

## 2.2 Beyond Independent Relevance

In the following we examine two popular examples of ranking approaches for subtopic retrieval based on the interdependent document relevance paradigm.

**Maximal marginal relevance (MMR)** is an intuitive technique for addressing diversity between documents [2]. Using a tuneable parameter, this ranking approach balances the relevance of a candidate document to 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<sup>1</sup>:

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

where  $I$  is the set of documents retrieved by the traditional ranking method, i.e. PRP;  $J$  is the set of documents that have been already ranked, i.e.  $x_j$ ; and  $x_i$  is a candidate document in  $I \setminus J$ , which is the set of documents that have not been ranked yet. 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 between documents. A value of the parameter  $\lambda$  greater than 0.5 assigns more importance to the similarity between document and query, rather than to the novelty/diversity of the document with respect to the ones ranked previously. Conversely, when  $\lambda < 0.5$ , novelty/diversity is favoured over relevance.

**Portfolio theory (PT)** suggests that ranking strategies should rank documents also considering the risk associated with ranking selecting specific documents [21]. 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 while minimising variance. The resultant objective function that is maximised by PT is as follows:

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

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  $\rho_{x_i, x_j}$  is the correlation between document  $x_i$  and document  $x_j$ .

<sup>1</sup> Note that the ranking formula that we report and use in our work is a modification of the formula originally proposed in [2]. However, the behaviour of the approach and the outcome of the ranking process is equivalent in both versions.

In summary, MMR and PT have a similar underlying additive schema for combining relevance and diversity. A common component of their ranking functions is the estimation of the probabilities of relevance. Both methods then balance the relevance estimation using a second component, which in turns captures the degree of diversity between the candidate document and the ranking. Other approaches that implement, to some extent, the interdependent document relevance paradigms have been proposed: see for example the seminal work of Goffman [8] and the recent work of Zuccon et al. [24]. In this paper we focus just on MMR and PT and we empirically compare them to the alternative paradigm for subtopic retrieval and our integration proposal.

### 2.3 Subtopic Aware Paradigm for Diversity

In the following we revise a number of examples belonging to the subtopics aware paradigm. These approaches have an explicitly indication of which subtopics are covered by each document. The underlying intuition is that once the subtopics have been modelled and the documents that cover these subtopics are identified, a ranking strategy can be devised so that it selects documents that belong to different classes of subtopics. Several techniques can be employed to produce or estimate a hypothetical partition of the retrieved documents according to the subtopics they might cover. For example, in [3] Carterette and Chandar use LDA and Lavrenko’s relevance models [13] for estimating the presence of subtopics within documents. Alternative techniques that can be employed to this end are probabilistic latent semantic analysis (PLSA) [10] and clustering (e.g. K-mean clustering [14]). In [6, 23] subtopics are estimated from the retrieved documents using clustering: presenting results that belong to different clusters is meant to guarantee the novelty of subtopics in the document ranking. However, information redundancy and document relevance are ignored in the document selection process. Regardless of the specific technique employed to estimate subtopics, a document ranking that exploits such explicit evidence can be formulated in various ways. In the following paragraphs we examine two approaches that follow the subtopic aware paradigm by exploiting evidence drawn from clusters of documents. Common to both approaches is the assumption that each cluster contains documents that address the same subtopic, and thus documents can be divided into classes on the basis of the subtopic (or subtopics) they cover.

**Interpolated approach.** This approach is directly connected with the cluster hypothesis<sup>2</sup>, and it prescribes that the relevance estimation of a document should be interpolated with the information obtained by clusters [12]. Formally, the retrieval score of a candidate document  $x_i$  is calculated as:

$$\hat{p}(x_i, q) = \lambda p(x_i, q) + (1 - \lambda) \sum_{c_j \in C} p(c_j, q) p(x_i, c_j) \quad (3)$$

where  $c_j$  is a cluster of documents in  $C$ , i.e. the set of document clusters modelled by topic modelling approaches;  $\lambda$  is a hyper-parameter that controls the

<sup>2</sup> Relevant documents tend to be more similar to each other than non-relevant documents [20].

balance between the probability of relevance and the probability of the document belonging to a cluster. In the context of our paper, we assume that  $p(a, b)$  is a similarity function between the objects<sup>3</sup>  $a$  and  $b$ . Note that when  $\lambda = 0$ , the ranking function of Eq. 3 returns documents within the cluster with highest similarity to the query, i.e. the cluster with higher  $p(c_j, q)$ . In the following we indicate this approach with **Interp**(.).

**Cluster representative approach.** This approach aims to cover the whole set of subtopics at early ranks at least with one representative document. For example, in [19] three clustering methods were employed and only the representative documents from visually formed clusters were presented to the users with the aim of facilitating faster browsing and retrieval. Similarly, in [7] the document ranking is formed by selecting documents from clusters in a round-robin fashion, i.e. assigning an order to the clusters and selecting a representative document cyclically through all clusters. The same approach might be applied to different algorithms that model subtopics, i.e. K-Mean, EM, and DBSCAN clustering, LDA, PLSA, and relevance models. What differentiates each instantiation of the approach is the function used to select cluster representatives. For example, in [7] cluster representatives are selected according to the order documents are added to clusters. An alternative approach is suggested by Deselaers et al. [6] where cluster representatives are selected according to their relevance to the query. In our empirical study we opt to investigate this latest solution, that we denote in the following with **Repre<sub>PRP</sub>**(.).

### 3 Integration Approach

In the interdependent document relevance paradigm, subtopic coverage is implicitly achieved by considering both document relevance and a measure of similarity/diversity between documents, where the latter measure indicates the dependency of documents. Nevertheless, since there is no explicit knowledge or model of the subtopics contained in the documents, subtopics coverage is hardly addressed although it is a main criterion for assessing ranking quality in the subtopic retrieval task.

In the subtopic aware paradigm, subtopics that a document covers are explicitly identified. However, document relevance is commonly ignored and the novelty of a ranking relies exclusively on the quality of the subtopic estimation techniques employed. Furthermore, these techniques might not be able to precisely model subtopics as they are perceived by users. Therefore there might be, in practice, subtopic redundancy within the ranking formed using this paradigm.

In this section we consider whether the two paradigms we have exposed so far can be integrated in order to form a family of new approaches for subtopic retrieval. Additionally, we hypothesise that subtopic redundancy due to falsely modelling subtopics in the subtopic aware paradigm can be alleviated by measuring document dependency in the interdependent document relevance paradigm.

---

<sup>3</sup> These can be queries, documents, or clusters.

To this end, we suggest to exploit document dependencies when selecting representatives from subtopic classes (e.g. clusters), obtained employing any of the approaches belonging to the subtopic aware paradigm. 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 documents with associated estimations of probability of relevance. Thereafter, the set of retrieved documents is partitioned into classes, for example according to clustering or LDA. The assumption at this stage is that a class corresponds to a subtopic of the information need and thus a class contains all the documents that address a common subtopic. When producing a ranking, we impose that each class has to be represented by a document in the ranking at least once. Specifically, we first rank the subtopic classes according to the average relevance of the documents contained in each class. Given a query  $q$  and a class  $c_k$ , average class relevance is defined as  $S_{avg}(c_k, q) = \frac{1}{|I_k|} \sum_{x_i \in I_k} s(x_i, q)$ , where  $I_k$  is the set of documents belonging to  $c_k$ ,  $X = \{x_1, \dots, x_n\}$  is the initial set of retrieved documents and  $s(x, q)$  is the estimated relevance of document  $x$  with respect to query  $q$ . Average class relevance is employed to arrange in a decreasing order the subtopic classes. Thereafter, a round-robin approach that follows the order suggested by average class relevance is used so as to select individual documents within the subtopic classes.

To select a specific document within each subtopic class, we employ an intra-list dependency-based approach, and thus integrate the two different subtopic retrieval paradigms into a common family of approaches. For example, if at this stage a MMR-like function is used, then the following objective function should be maximised:

$$J_j = J_{j-1} \cup \underset{x_{k,n} \in X_k \setminus J}{\operatorname{argmax}} [\lambda S(x_{k,n}, q) + (1 - \lambda) \underset{x_j \in J}{\operatorname{avg}} D(x_{k,n}, x_j)] \quad (4)$$

where  $X_k = \{x_{k,1}, x_{k,2}, x_{k,3}, \dots, x_{k,n}\}$  is the set of retrieved documents that belong to the subtopic class  $c_k$  and  $J$  is the set of documents that has been already ranked. Of course, other approaches, such as PT or the quantum probability ranking principle [24], can be used at this stage.

## 4 Empirical study

In the following we present the experimental methodology of the empirical study we perform in this paper. The objectives of our empirical investigation are:

1. to compare different state-of-the-art approaches based on the two ranking paradigms presented in Section 2. Specifically, which paradigm delivers the best document ranking for subtopic retrieval?
2. to investigate and validate the integration approach we outlined in Section 3. Specifically, we aim to answer the question: does considering at the same time interdependent document relevance and subtopic awareness improve performances in the subtopic retrieval task?

In order to answer these questions, we test state-of-the-art approaches belonging to both paradigms and our integration approach on a number of test collections. In particular, we use the ImageCLEF 2009 Photo Retrieval<sup>4</sup> [15], the TREC ClueWeb 2009 (limited to part B) [5], and the TREC 6,7,8 interactive [22] collections.

Textual information have been indexed using Lemur<sup>5</sup>, which served also as platform for developing the ranking approaches using the C++ API. We removed standard stop-words [20] and applied Porter stemming to both documents and queries. Queries are extracted from the titles of the TREC and CLEF topics.

Okapi BM25 has been used to estimate document relevance given a query; these estimates have been directly employed to produce the PRP run in our experiments. The same weighting schema has been used to produce the relevance estimates and the document term vectors that are employed by some of the re-ranking strategies to compute similarity (e.g. in MMR) or correlation (e.g. in PT). This is consistent with previous works [21]. We experiment with several ranking lengths, i.e. 100, 200, 500, and 1000, but in this paper we report results for ranking up to 100 documents long for space matters<sup>6</sup>.

The MMR approach has been instantiated as discussed in Section 2, where we employed the BM25 score as similarity function between document and query, and the opposite of the cosine similarity between documents as a measure of dissimilarity. Furthermore we varied the value of  $\lambda$  in the range  $[0,1]$  with steps of 0.1. When testing PT, we explored values of  $b$  in the range<sup>7</sup>  $[-9, 9]$ ; we treat the variance of a document as a parameter that is constant with respect to all the documents, similarly to [21]. We experimented with variance values  $\delta^2$  ranging from  $10^{-9}$  to  $10^{-1}$ , and selected the ones that achieve the best performances in combination with the values of  $b$  through a grid search of the parameter space. Correlation between documents is computed by the Pearson's correlation between the term vectors representing documents.

Regarding the runs based on the subtopic aware paradigm, we adopt three techniques to model subtopics: K-mean clustering, PLSA and LDA, although alternative strategies may be suitable. For each query, the number of clusters/classes required by the techniques has been set according to the subtopic relevance judgements for that query. When techniques like LDA and PLSA are used, we obtain an indication of the probability that a subtopic is covered by a document. Because in our study we do not consider overlapping classes of subtopics, we assign to each document only one subtopic: i.e. the subtopic that has been estimated as the most likely for that document. After the classes or clusters are formed, documents are ranked according to the approaches we illustrated in Sections 2.3 and 3, specifically:

---

<sup>4</sup> This collection consists of images with associated text captions. We discard the image features, and just consider the text captions.

<sup>5</sup> <http://www.lemurproject.org/>

<sup>6</sup> The results obtained with different ranking depths present similar results and will be posted in the author website.

<sup>7</sup> Note that when  $b = 0$  the ranking of PT is equivalent to the one of PRP.

- **Interp(.)**: selects documents that maximise the interpolation algorithm for cluster-based retrieval;
- **Repre<sub>PRP</sub>(.)** : selects documents with the highest probability of relevance in the given classes/subtopics;
- **Integr<sub>MMR</sub>(.)**: selects documents according to MMR, as an example of strategy based on the interdependent document relevance paradigm.

Interp(.) requires to build a vector which represents the cluster/class in order to compute  $sim(c, q)$ ,  $sim(c, d)$ , and the distance to the centre of the cluster/class. To this aim we create cluster’s centroid vector: for a cluster  $c_k$  the cluster representative vector is expressed by  $\mathbf{c}_k = (\bar{w}_{1,k}, \bar{w}_{2,k}, \dots, \bar{w}_{t,k})$ , where  $\bar{w}_{t,k}$  is the average of the term weights of all the documents within cluster  $c_k$ . Cosine similarity is used to evaluate the similarity of clusters against query and document.

Repre<sub>PRP</sub>(.) does not require parameter tuning. On the contrary, when instantiating Interp(.) and Integr<sub>MMR</sub>(.), we varied their hyper-parameter in the range [0,1] and select the value that obtained the best performances. The combinations of the subtopic estimation algorithms and the document selection criteria form in total nine experimental instantiations that we tested in our empirical study, i.e. Interp(K-Mean), Repre<sub>PRP</sub>(K-Mean), Integr<sub>MMR</sub>(K-Mean), Interp(PLSA), Repre<sub>PRP</sub>(PLSA), Integr<sub>MMR</sub>(PLSA), Interp(LDA), Repre<sub>PRP</sub>(LDA), and Integr<sub>MMR</sub>(LDA).

In addition to the use of subtopic estimation techniques, we investigate the situation where subtopic coverage evidence is drawn from the relevance judgements. We assume that a document can cover only one subtopic: although this assumption is limitative (and not true), it is adequate in the context of our study<sup>8</sup>. Documents that have been judged as belonging to only one subtopic are assigned to a specific cluster that represents the subtopic. These documents are then used to construct clusters’ centroid vectors in order to represent the clusters. Afterwards, Euclidean distance is used to assign to a cluster those documents that have been judged to cover two or more subtopics, and the cluster representative is updated. The documents that have not been judged are assigned to clusters using the same procedure. Instantiations of the approaches based on this subtopic evidence (denoted by “**Ideal Subtopics**” ) are an indication of the upper bound performances each approach can achieve.

## 5 Experimental Results

The results obtained in our empirical investigation are reported in Tables 1, 2, 3 for ImageCLEF 2009, TREC ClueWeb 2009, and TREC 6,7,8 collections respectively. Results are evaluated using  $\alpha$ -NDCG [4], S-recall and S-MRR [22]; regarding the parametrisation of some approaches, we report here only the best results of each ranking strategy with respect to  $\alpha$ -NDCG@10. Parameter values are shown underneath the methods. The results obtained employing *Ideal*

<sup>8</sup> Further work will be directed towards a methodology for generating subtopic clusters/classes where this assumption is relaxed.

		Models	$\alpha$ -NDCG@10	S-R@10	S-R@20	S-MRR 25%	S-MRR 50%
		<b>PRP</b>	0.4550	0.5330	0.6235	0.7589	0.5221
		<b>MMR</b> ( $\lambda = 0.7$ )	0.4830 (+6.15%)	<b>0.6651</b> (+24.80%)	<b>0.7315</b> (+17.33%)	0.7297 (-3.85%)	0.5041 (-3.44%)
		<b>PT</b> ( $b = 4, \delta^2 = 10^{-1}$ )	0.4450* (-2.20%)	0.5648* (+5.97%)	0.6636* (+6.44%)	0.7307 (-3.72%)	0.4916 (-5.84%)
Subtopic Estimation	KMean	<b>Interp</b> ( $\lambda = 1.0$ )	0.4550 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		<b>Repre<sub>PRP</sub></b>	0.4660 (+2.42%)	0.5701* (+6.97%)	0.6573* (+5.43%)	0.7503 (-1.13%)	0.5173 (-0.92%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 0.9$ )	0.4860 <sup>†</sup> (+6.81%)	0.6256 <sup>†</sup> (+17.39%)	0.6910* (+10.83%)	0.7588 (-0.01%)	0.4985 (-4.53%)
	PLSA	<b>Interp</b> ( $\lambda = 1.0$ )	0.4550 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		<b>Repre<sub>PRP</sub></b>	0.4730 (+3.96%)	0.5766* (+8.19%)	0.6805* (+9.15%)	0.7608 (+0.25%)	0.5361 (+2.69%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 0.9$ )	0.4950 <sup>†</sup> (+8.79%)	0.6520 <sup>†</sup> (+22.33%)	0.7179 (+15.14%)	0.7743 (+2.03%)	0.4865 (-6.81%)
	LDA	<b>Interp</b> ( $\lambda = 1.0$ )	0.4550 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		<b>Repre<sub>PRP</sub></b>	0.4740 (+4.18%)	0.5683* (+6.62%)	0.6637* (+6.45%)	<b>0.8104*<sup>†</sup></b> (+6.79%)	<b>0.5406</b> (+3.55%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 0.9$ )	<b>0.5020<sup>†</sup></b> (+10.33%)	0.6236* <sup>†</sup> (+17.01%)	0.6842* (+9.74%)	0.7973 (+5.06%)	0.5223 (+0.04%)
	Ideal Subtopics	<b>Interp</b> ( $\lambda = 1.0$ )	0.4550 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		<b>Repre<sub>PRP</sub></b>	0.5700* <sup>†</sup> (+25.27%)	0.7901* <sup>†</sup> (+48.24%)	0.8066* <sup>†</sup> (+29.37%)	0.7440 (-1.97%)	0.5544 (+6.18%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 0.9$ )	0.6080* <sup>†</sup> (+33.63%)	0.8066* <sup>†</sup> (+51.33%)	0.8066* <sup>†</sup> (+29.37%)	0.8183* <sup>†</sup> (+7.83%)	0.6241* <sup>†</sup> (+19.54%)

**Table 1.** Retrieval performances on the *ImageCLEF 2009 (Photo Retrieval)* collection with % of improvement over PRP. Parametric runs are tuned w.r.t.  $\alpha$ -NDCG@10. Statistical significances at 0.05 level against MMR, and PT are indicated by \* and <sup>†</sup> respectively.

*Subtopics* represent the upper bound each technique can achieve. When statistically significant differences (according to t-test, with  $p < 0.05$ ) against MMR and PT are individuated, we report them with \* and <sup>†</sup> respectively. In Table 3, the statistical significance analysis is not reported as the number of topics is very limited (just 20 topics) and thus calculating statistical significance does not convey meaningful information.

The results obtained on the ImageCLEF 2009 collection suggest that instantiations of the subtopic aware paradigm outperform instantiations of the interdependent document relevance paradigm, with respect to  $\alpha$ -NDCG@10 and when subtopics are estimated using LDA. Other subtopic estimation techniques (PLSA and clustering) obtain comparable results. However, the best results overall (at least when considering<sup>9</sup>  $\alpha$ -NDCG@10) are obtained by our integration paradigm using LDA for estimating subtopics. Thus integrating the two retrieval paradigms improves performances in the case of ImageCLEF 2009. The results obtained employing evidences derived from the ideal subtopics configuration in-

<sup>9</sup> Note that parameters have been tuned according to this measure.

Models		$\alpha$ -NDCG@10	S-R@10	S-R@20	S-MRR 25%	S-MRR 50%	
<b>PRP</b>		0.0680	0.1606	0.2719	0.1787	0.0953	
<b>MMR</b> ( $\lambda = 0.7$ )		0.1050 (+54.41%)	0.1664 (+3.65%)	0.2451 (-9.86%)	0.1741 (-2.58%)	0.0786 (-17.53%)	
<b>PT</b> ( $b = -5, \delta^2 = 10^{-4}$ )		0.1510 (+122.06%)	<b>0.2676*</b> (+66.64%)	<b>0.3486*</b> (+28.20%)	0.2179 (+21.90%)	0.1264 (+32.69%)	
Subtopic Estimation	KMean	<b>Interp</b> ( $\lambda = 0.2$ )	<b>0.1670*</b> (+145.59%)	0.1682 <sup>†</sup> (+4.77%)	0.2331 <sup>†</sup> (-14.27%)	<b>0.3411*</b> (+90.84%)	0.1367 (+43.44%)
		<b>Repre<sub>PRP</sub></b>	0.1030 <sup>†</sup> (+51.47%)	0.1819 <sup>†</sup> (+13.29%)	0.2466 <sup>†</sup> (-9.32%)	0.2077 (+16.21%)	0.1145 (+20.21%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 1.0$ )	0.12700 (+86.76%)	0.20191 (+25.74%)	0.26424 <sup>†</sup> (-2.82%)	0.29128 (+62.96%)	0.13653 (+43.31%)
	PLSA	<b>Interp</b> ( $\lambda = 0.3$ )	<b>0.1670*</b> (+145.59%)	0.1682 <sup>†</sup> (+4.77%)	0.2331 <sup>†</sup> (-14.27%)	<b>0.3411*</b> (+90.84%)	0.1367 (+43.44%)
		<b>Repre<sub>PRP</sub></b>	0.1160 (+70.59%)	0.1876 (+16.81%)	0.2858 (+5.10%)	0.2265 (+26.73%)	0.1120 (+17.55%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 1.0$ )	0.1440* (+111.76%)	0.2099 (+30.72%)	0.2926 (+7.62%)	0.3140* (+75.69%)	<b>0.1490*</b> (+56.41%)
	LDA	<b>Interp</b> ( $\lambda = 0.2$ )	<b>0.1670*</b> (+145.59%)	0.1682 <sup>†</sup> (+4.77%)	0.2331 <sup>†</sup> (-14.27%)	<b>0.3411*</b> (+90.84%)	0.1367 (+43.44%)
		<b>Repre<sub>PRP</sub></b>	0.1130 (+66.18%)	0.2047 (+27.46%)	0.2902 (+6.74%)	0.2134 (+19.40%)	0.0990 (+3.93%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 1.0$ )	0.1260 (+85.29%)	0.2149 (+33.84%)	0.2741 (+0.81%)	0.2333 (+30.51%)	0.1211 (+27.15%)
Ideal Subtopics	<b>Interp</b> ( $\lambda = 0.1$ )	0.1670* (+145.59%)	0.1682 <sup>†</sup> (+4.77%)	0.2331 <sup>†</sup> (-14.27%)	0.3411* (+90.84%)	0.1367 (+43.44%)	
	<b>Repre<sub>PRP</sub></b>	0.2000* (+194.12%)	0.3332* (+107.53%)	0.3872* (+42.42%)	0.2868* (+60.48%)	0.1780* (+86.85%)	
	<b>Integr<sub>MMR</sub></b> ( $\lambda = 0.1$ )	0.2330* (+242.65%)	0.3376* (+110.23%)	0.3774* (+38.81%)	0.4041* <sup>†</sup> (+126.09%)	0.1891* (+98.46%)	

**Table 2.** Retrieval performances on the *TREC ClueWeb 2009* collection with % of improvement over PRP. Parametric runs are tuned w.r.t.  $\alpha$ -NDCG@10. Statistical significances at 0.05 level against MMR, and PT are indicated by \* and † respectively.

indicate how much each subtopic aware strategy would perform if subtopics were correctly identified. In this case, the integration approach performs the best.

In Table 2 we report the results from our investigation on TREC ClueWeb 2009. Approaches based on the subtopic aware paradigm only slightly outperform (with respect to  $\alpha$ -NDCG@10) approaches based on the interdependent document relevance. In particular, this is evident when the runs obtained by PT are compared against the runs obtained by Interp(.) and when the MMR runs are compared against the Repre<sub>PRP</sub>(.) runs. However, it can be noticed that the performances of the subtopic aware approaches do not highly vary when considering different subtopic estimation techniques. If the ideal subtopic estimation is considered, then the Repre<sub>PRP</sub>(.) approach is shown to outperform instantiations of the other state-of-the-art approaches. However, in this scenario our integration approach outperforms any other method, and gains up to the 16.5% over the Repre<sub>PRP</sub>(.). The performance difference between the approaches that use the estimated subtopic evidence and the ones that employ the ideal subtopic evidence suggests that subtopic estimation techniques fail to capture subtopics. This might be because of the more noisy nature of the ClueWeb collection with respect to the ImageCLEF collection.

		Models	$\alpha$ -NDCG@10	S-R@10	S-R@20	S-MRR 25%	S-MRR 50%
		<b>PRP</b>	0.4260	<b>0.3868</b>	<b>0.5319</b>	0.2877	<b>0.1618</b>
		MMR ( $\lambda = 1.0$ )	0.4260 (0.00%)	<b>0.3868</b> (0.00%)	<b>0.5319</b> (0.00%)	0.2877 (0.00%)	<b>0.1618</b> (0.00%)
		<b>PT</b> ( $b = -1, \delta^2 = 10^{-1}$ )	<b>0.4330</b> (+1.64%)	0.3735 (-3.44%)	0.4972 (-6.52%)	<b>0.3028</b> (+5.26%)	0.1643 (+1.58%)
Subtopic Estimation	K-means	<b>Interp</b> ( $\lambda = 1.0$ )	0.4260 (0.00%)	<b>0.3868</b> (0.00%)	<b>0.5319</b> (0.00%)	0.2877 (0.00%)	<b>0.1618</b> (0.00%)
		<b>Repre<sub>PRP</sub></b>	0.2380 (-44.13%)	0.2517 (-34.94%)	0.3483 (-34.52%)	0.1340 (-53.43%)	0.0692 (-57.24%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 1.0$ )	0.2380 (-44.13%)	0.2517 (-34.94%)	0.3483 (-34.52%)	0.1340 (-53.43%)	0.0692 (-57.24%)
	PLSA	<b>Interp</b> ( $\lambda = 1.0$ )	0.4260 (0.00%)	<b>0.3868</b> (0.00%)	<b>0.5319</b> (0.00%)	0.2877 (0.00%)	<b>0.1618</b> (0.00%)
		<b>Repre<sub>PRP</sub></b>	0.2580 (-39.44%)	0.3132 (-19.03%)	0.4090 (-23.11%)	0.1788 (-37.84%)	0.0688 (-57.47%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 0.6$ )	0.2630 (-38.26%)	0.3178 (-17.84%)	0.3953 (-25.68%)	0.1797 (-37.54%)	0.0657 (-59.40%)
	LDA	<b>Interp</b> ( $\lambda = 1.0$ )	0.4260 (0.00%)	<b>0.3868</b> (0.00%)	<b>0.5319</b> (0.00%)	0.2877 (0.00%)	<b>0.1618</b> (0.00%)
		<b>Repre<sub>PRP</sub></b>	0.2720 (-36.15%)	0.3078 (-20.44%)	0.4049 (-23.87%)	0.2043 (-28.99%)	0.1024 (-36.69%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 0.4$ )	0.2820 (-33.80%)	0.3111 (-19.57%)	0.3902 (-26.64%)	0.2163 (-24.82%)	0.0989 (-38.88%)
	Ideal Subtopics	<b>Interp</b> ( $\lambda = 1.0$ )	0.4260 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		<b>Repre<sub>PRP</sub></b>	0.5060 (+18.78%)	0.5664 (+46.41%)	0.6761 (+27.12%)	0.2898 (+0.74%)	0.1575 (-2.67%)
		<b>Integr<sub>MMR</sub></b> ( $\lambda = 1.0$ )	0.5080 (+19.25%)	0.5692 (+47.15%)	0.6793 (+27.72%)	0.2971 (+3.28%)	0.1565 (-3.28%)

**Table 3.** Retrieval performances on the *TREC 6,7,8 interactive* collection with % of improvement over PRP. Parametric runs are tuned w.r.t.  $\alpha$ -NDCG@10.

A similar consideration can be evidenced by the results obtained on the TREC 6,7,8 interactive collection, and reported in Table 3. Techniques for subtopic estimation seem to provide the wrong evidence to the subtopic aware approaches, and thus these approaches perform as well as or worse than the PRP baseline or the interdependent document relevance approaches. In particular note that the results of MMR and Interp(.) are obtained when their hyper-parameter  $\lambda$  is set to 1, that is, when their ranking formula is equivalent to the one of the PRP baseline. However, when subtopics are estimated from the relevance judgements, as in the case of the ideal subtopics technique, the  $\text{Repre}_{PRP}(\cdot)$  and  $\text{Integr}_{MMR}(\cdot)$  instantiations outperform any other approach.

## 6 Conclusions

The goal of this paper is to empirically compare state-of-the-art methods and an integration approach we propose for subtopic retrieval. Three test collections has been used to this aim. We find that overall approaches derived from the subtopic aware paradigm perform better (and in many cases significantly better) than approaches based on the interdependent document relevance paradigm. Amongst the techniques for estimating subtopics, LDA and PLSA has been shown to pro-

vide better evidences than K-mean clustering. However, all the techniques for estimating subtopics fail to some extent to provide high quality evidences in the case of the TREC ClueWeb 2009 and the TRE 6,7,8 interactive collections. This might be due to the noisy nature of the documents contained in the collections (web pages and newswire articles). The integration approach, that combines implicit and explicit approaches for ranking diversification, has been shown to outperform state-of-the-art approaches, in particular when subtopics are directly derived from the relevance judgements. Thus, the integration approach has the capability to improve subtopic retrieval performances when effective topic estimation is deployed. Further investigation will be directed towards the empirical validation of effective topic estimation techniques.

## 7 Acknowledgement

This work has been supported by the Royal Thai Government and the EPSRC Renaissance project (EP/F014384/1).

## References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. of Mach. Learning Res.*, 3:993–1022, 2003.
2. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, pages 335–336, 1998.
3. B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM '09*, pages 1287–1296, 2009.
4. C. L. 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*, pages 659–666, 2008.
5. C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In *Proc. of TREC-2009*, 2009.
6. T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *CIVR '09*, pages 1–8, 2009.
7. M. Ferecatu and H. Sahbi. TELECOM ParisTech at ImageCLEFphoto 2008: Bimodal text and image retrieval with diversity enhancement. In *Working Notes for the CLEF 2008 workshop*, 2008.
8. W. Goffman. An indirect method of information retrieval. *Inf. Stor. and Ret.*, 4(4):361 – 373, 1968.
9. M. D. Gordon and P. Lenk. When is the probability ranking principle suboptimal. *JASIS*, 43(1):1–14, 1999.
10. T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, pages 50–57, 1999.
11. J. Huang, S. R. Kumar, and R. Zabih. An automatic hierarchical image classification scheme. In *MM '98*, pages 219–228, 1998.
12. O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *SIGIR '04*, pages 194–201, 2004.

13. V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, pages 120–127, 2001.
14. J. B. MacQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of the 5th Berkeley Symp. on Math. Stat. and Prob.*, pages 281–297, 1967.
15. M. L. Paramita, M. Sanderson, and P. Clough. Developing a test collection to support diversity analysis. In *Proc. of Redundancy, Diversity, and IDR workshop - SIGIR' 09*, pages 39–45, 2009.
16. S. E. Robertson. The probability ranking principle in IR. *J. of Doc.*, 33(4):294–304, 1977.
17. I. Soboroff. On evaluating web search with very few relevant documents. In *SIGIR '04*, pages 530–531, 2004.
18. K. H. Stirling. On the limitations of document ranking algorithms in information retrieval. In *SIGIR '81*, number 1, pages 63–65. ACM, 1981.
19. R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *WWW '09*, pages 341–341, 2009.
20. C. J. van Rijsbergen. *Information Retrieval, 2nd Ed.* Butterworth, 1979.
21. J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR '09*, pages 115–122, 2009.
22. C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*, pages 10–17, 2003.
23. Z. Q. Zhao and H. Glotin. Diversifying image retrieval by affinity propagation clustering on visual manifolds. *IEEE MultiMedia*, 99(1), 2009.
24. G. Zuccon, L. Azzopardi, and K. van Rijsbergen. The quantum probability ranking principle for information retrieval. In *ICTIR '09*, pages 232–240, 2009.