When Two Is Better Than One: 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. A document might be non-relevant if the user has already examined other documents containing similar information [3]. The utility of a document thus depends upon which documents have been ranked in previous positions. 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 covering several subtopics that the user might be interested in [16].

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, e.g. the probability

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 162–172, 2010.

ranking principle (PRP) [13], where documents are ranked exclusively according to their probability of being relevant to a query. 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 [7].

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) [1], which interpolates document relevance and documents relationships; and portfolio theory (PT) [15], which combines relevance estimates and document correlations.
- **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 [5], classification [9], latent Dirichlet allocation (LDA) [2], and probabilistic latent semantic analysis (PLSA) [8]. Afterwards, result diversification is achieved by interleaving in a ranking the documents belonging to different estimated subtopics. Several criteria can be applied to select documents after the evidence of 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 and diversity among ranked documents is promoted. 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.

2 Related Work

2.1 Beyond Independent Relevance

In this paper, we examine just two popular examples of ranking approaches for subtopic retrieval based on the interdependent document relevance paradigm, e.g. MMR [1] and PT [15]. Both approaches have a similar underlying assumption, which combines and maximises the estimated document relevance and diversity during ranking process. For instance, MMR method attempts to maximise marginal similarity between documents and query, and dissimilarity between candidate documents and all documents ranked at previous positions. To rank a document at rank $j+1$, the MMR strategy is characterised by the following ranking function[1]: $\text{argmax}[\lambda S(x_i, q) + (1-\lambda)\,\text{avg}\, D(x_i, x_j)]$, where x_i is a candidate document that has been retrieved by a traditional ranking method but has not been ranked yet; x_j is a document that has been already ranked; and λ is a tuneable parameter that assigns importance to either similarity or novelty/diversity. The function $S(x_i, q)$ is a normalised similarity function used for document retrieval, whereas $D(x_i, x_j)$ is a normalised diversity metric between documents, such as the cosine similarity. For further details of PT approach, we refer to the paper [15].

2.2 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 [2] Carterette and Chandar use LDA to estimate the presence of subtopics within documents. Alternative techniques that can be employed to this end are PLSA [8] and clustering (e.g. K-mean clustering). In [5] 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

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

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[2], and it prescribes that the relevance estimation of a document should be interpolated with the information obtained by clusters [10]. Formally, the retrieval score of a candidate document x_i is calculated as: $\hat{p}(x_i, q) = \lambda p(x_i, q) + (1 - \lambda) \sum_{c_j \in C} p(c_j, q) p(x_i, c_j)$, where c_j is a cluster of documents in C, i.e. the set of document clusters modelled by topic modelling approaches; λ[3] is a hyper-parameter that controls the 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[4] a and b. 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 [6] 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 [6] cluster representatives are selected according to the order in which documents are added to clusters. An alternative approach is suggested by Deselaers et al. [5] where cluster representatives are selected according to their relevance to the query. In our empirical study we opt to investigate Deselaers's solution, that we denote in the following with $\mathbf{Repre}_{PRP}\mathbf{(.)}$.

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

[2] Relevant documents tend to be more similar to each other than non-relevant documents [14].

[3] Note that when $\lambda = 0$, the ranking function returns documents within the cluster with highest similarity to the query, i.e. the cluster with higher $p(c_j, q)$.

[4] These can be queries, documents, or clusters.

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 the subtopic redundancy can be alleviated by measuring dependencies between documents after imprecisely estimating subtopics. To this end, we suggest to exploit the 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, $X = \{x_1, ..., x_n\}$ 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 the subtopic classes in a decreasing order. Thereafter, a round-robin approach that follows the order suggested by average class relevance is used 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:

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

where $X_k = \{x_{k,1}, x_{k,2}, x_{k,3}, ..., x_{k,n}\}$ is the set of retrieved documents belonging 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, 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 in a high level regardless of particular techniques of ranking paradigms used. 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[5] [12], and the TREC ClueWeb 2009 (limited to part B) [4]. A broader empirical investigation, we refer the interested reader to our extended technical report [11], which includes results based on the TREC 6,7,8 interactive collection [16].

Textual information have been indexed using Lemur[6], which served also as platform for developing the ranking approaches using the C++ API. We removed standard stop-words [14] 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 [15]. 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.

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 λ in the range [0,1] with steps of 0.1. When testing PT that requires two setting parameters, we explored values of b in the range[7] [-9, 9]; we treat the variance of a document as a parameter that is constant with respect to all the documents, similarly to [15]. We experimented with variance values δ^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

[5] This collection consists of images with associated text captions. We discard the image features, and just consider the text captions.

[6] `http://www.lemurproject.org/`

[7] Note that when $b = 0$ the ranking of PT is equivalent to the one of PRP.

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 2.2 and 3, specifically:

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

Interp(.) requires to build a vector representing 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 $(\bar{w}_{1,k}, \bar{w}_{2,k}, ..., \bar{w}_{t,k})$, 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 a query and documents.

Repre$_{PRP}$(.) does not require parameter tuning. On the contrary, when instantiating Interp(.) and Integr$_{MMR}$(.), 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, such as Interp(K-Mean), Repre$_{PRP}$(PLSA), Integr$_{MMR}$(LDA) etc.

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[8]. 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

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

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 for ImageCLEF 2009 and TREC ClueWeb 2009 collections respectively. Results are evaluated using α-NDCG [3], S-recall and S-MRR [16]; regarding the parametrisation of some approaches, we report here only the best results of each ranking strategy with respect to α-NDCG@10. Parameter values are shown underneath the methods. The results obtained employing *Ideal Subtopics* represent the upper bound each technique can achieve. When statistical significant differences (according to t-test, with $p < 0.05$) against MMR and PT are individuated, we report them with * and † respectively.

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 α-NDCG@10 and

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

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

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

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

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[9] α-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 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 α-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$_{PRP}$(.) 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$_{PRP}$(.) approach is shown to outperform instantiations

[9] Note that parameters have been tuned according to this measure.

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 $\text{Repre}_{PRP}(.)$. 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.

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. Two 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 provide 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 collection. 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.

References

1. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR 1998, pp. 335–336 (1998)
2. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: CIKM 2009, pp. 1287–1296 (2009)
3. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR 2008, pp. 659–666 (2008)
4. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 Web Track. In: Proc. of TREC 2009 (2009)
5. Deselaers, T., Gass, T., Dreuw, P., Ney, H.: Jointly optimising relevance and diversity in image retrieval. In: CIVR 2009, pp. 1–8 (2009)
6. Ferecatu, M., Sahbi, H.: TELECOM ParisTech at ImageCLEFphoto 2008: Bimodal text and image retrieval with diversity enhancement. In: Working Notes for the CLEF 2008 workshop (2008)
7. Gordon, M.D., Lenk, P.: When is the probability ranking principle suboptimal. JASIS 43, 1–14 (1999)

8. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR 1999, pp. 50–57 (1999)
9. Huang, J., Kumar, S.R., Zabih, R.: An automatic hierarchical image classification scheme. In: MM 1998, pp. 219–228 (1998)
10. Kurland, O., Lee, L.: Corpus structure, language models, and ad hoc information retrieval. In: SIGIR 2004, pp. 194–201 (2004)
11. Leelanupab, T., Zuccon, G., Jose, J.M.: Technical report: A study of ranking paradigms and their integrations for subtopic retrieval. Technical report, School of Computing Science, University of Glasgow (2010)
12. Paramita, M.L., Sanderson, M., Clough, P.: Developing a test collection to support diversity analysis. In: Proc. of Redundancy, Diversity, and IDR workshop SIGIR 2009, pp. 39–45 (2009)
13. Robertson, S.E.: The probability ranking principle in IR. J. of Doc. 33, 294–304 (1977)
14. van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworth (1979)
15. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: SIGIR 2009, pp. 115–122 (2009)
16. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: SIGIR 2003, pp. 10–17 (2003)