Explicit Search Result Diversification through Sub-Queries

Rodrygo L.T. Santos, Jie Peng, Craig Macdonald, and Iadh Ounis

Department of Computing Science
University of Glasgow, G12 8QQ, UK
{rodrygo,pj,craigm,ounis}@dcs.gla.ac.uk

Abstract. Queries submitted to a retrieval system are often ambiguous. In such a situation, a sensible strategy is to diversify the ranking of results to be retrieved, in the hope that users will find at least one of these results to be relevant to their information need. In this paper, we introduce xQuAD, a novel framework for search result diversification that builds such a diversified ranking by explicitly accounting for the relationship between documents retrieved for the original query and the possible aspects underlying this query, in the form of sub-queries. We evaluate the effectiveness of xQuAD using a standard TREC collection. The results show that our framework markedly outperforms state-of-the-art diversification approaches under a simulated best-case scenario. Moreover, we show that its effectiveness can be further improved by estimating the relative importance of each identified sub-query. Finally, we show that our framework can still outperform the simulated best-case scenario of the state-of-the-art diversification approaches using sub-queries automatically derived from the baseline document ranking itself.

1 Introduction

Information needs are inherently underspecified in the form of queries; certain queries are particularly more ambiguous (e.g., java, jaguar) than others, but even those that are seemingly well-defined may have multiple interpretations depending on the context in which they are issued or on their underlying intent [1].

In the situation where a clear interpretation of the user's information need cannot be easily determined, it might be too risky to just assume a single, plausible one (e.g., the most popular) and to focus on retrieving results that satisfy that particular *aspect*. Instead, a more sensible approach is to provide a diverse ranking of results covering as many aspects of the original query as possible, in the hope that the users will find at least one of these results to be relevant to their original information need. In fact, modern Web search engines usually offer suggested interpretations of the original query based on previous user interactions, so as to help the users further refine their original queries.

Diversifying search results typically involves a departure from the traditional assumption of document relevance independence when ranking documents for a given query [2]. Considered as a whole, the relevance of a document ranking for a given query should depend not only on the individual ranked documents, but also on how they relate to each other. For example, it is questionable whether users will find a given document relevant to their information need after examining other similar documents [3]. The general problem of minimising the redundancy among the retrieved documents—or, conversely, of maximising their coverage with respect to different aspects of the original query—is NP-hard [4]. Most previous works on search result diversification are based on a greedy approximation to this problem [5]. In common, these works attempt to reduce the redundancy among the retrieved documents by comparing them with respect to their content or their estimated relevance to the original query. By doing so, they implicitly assume that similar documents will cover similar aspects underlying the query.

On the other hand, as queries often carry some ambiguity, the broader topic represented by a given query can be usually decomposed into distinct sub-topics. This, in turn, motivates an alternative approach to search result diversification, centred on explicitly modelling the possibly several aspects underlying a query. In this paper, we introduce a new framework for search result diversification that exploits this intuition in order to maximise the aspects covered in a document ranking by comparing the retrieved documents with respect to their estimated relevance to each of these aspects. In particular, we uncover different aspects underlying the original query in the form of *sub-queries*, which are then used as a central element for comparing a given pair of documents based on how well they satisfy each sub-query. By doing so, we can take into account both the diversity of aspects covered by a single document, as well as its novelty in face of the aspects already covered by other retrieved documents. Moreover, the relative importance of each identified sub-query can be directly incorporated within our framework, so as to bias the diversification process towards those sub-queries likely to represent more plausible aspects of the initial query.

We compare our proposed framework to both classical as well as some recent search result diversification approaches using a standard TREC collection with relevance assessments at the sub-topic level. Our results show that the explicit account of the possible aspects underlying the original query as sub-queries can produce substantial improvements over both implicit and explicit state-of-the-art diversification approaches. The remainder of this paper is organised as follows. Section 2 provides an overview of recent works on search result diversification. Section 3 describes the major components of our proposed diversification framework, built around the concept of sub-queries. Section 4 details our experimental settings, while Section 5 discusses our main findings. Lastly, Section 6 presents our conclusions and directions for future work.

2 Background and Related Work

The problem of diversifying search results can be stated as:

> Given a query q, retrieve a ranking of documents $R(q)$ with maximum relevance with respect to q and minimum redundancy with respect to its coverage of the possible aspects underlying q.

In its general form, this problem can be reduced from the maximum coverage problem [6], which makes it NP-hard [4]. Most previous approaches to search result diversification are based on a greedy approximation to this problem, namely, the so-called maximal marginal relevance (MMR) method [5]. The general idea of MMR is to iteratively re-rank an initial set of documents retrieved for a given query by selecting, at each iteration, the document not yet selected with the highest estimated relevance to the query and highest dissimilarity to the already selected documents. The various approaches based on MMR differ mostly by how the similarity between documents is computed. For example, Carbonell and Goldstein [5] suggested using a content-based similarity function, e.g., the cosine distance between the vectors representing the retrieved documents. Zhai and Lafferty [7] proposed to model relevance and novelty within the language modelling framework. They devised six different methods, based on either the KL divergence measure or a simple mixture model. More recently, Wang and Zhu [8] employed the correlation between documents as a measure of their similarity.

In common, all of these approaches consider the possible aspects associated to the original query only in an implicit way, namely, by directly comparing the retrieved documents against one another, under the assumption that similar documents will cover similar aspects. By demoting similar documents in the ranking, these approaches aim to reduce the ranking overall redundancy. An alternative approach is to explicitly consider the different aspects associated to a query by directly modelling these aspects. For instance, Agrawal et al. [4] investigated the diversification problem by employing a taxonomy for both queries and documents. In their work, documents retrieved for a query are considered similar if they are confidently classified into one or more common categories covered by the query. By doing so, documents covering well-represented categories in the ranking are penalised, as they would bring little novelty in face of the already selected documents. A related approach was investigated by Radlinski and Dumais [9]. In order to compose a diverse ranking, they proposed to filter the results retrieved for a given query so as to limit the number of those satisfying each of the aspects of this query, represented as different query reformulations, as obtained from a large query log from a commercial search engine.

Our approach also considers the possible aspects associated to a query explicitly. Differently from the approaches of Agrawal et al. [4] and Radlinski and Dumais [9], however, we do not rely on a predefined taxonomy or on classification schemes, nor do we reserve predetermined shares of the final ranking for results answering each of the identified aspects associated to a query. Instead, we uncover different aspects underlying a query as sub-queries, and estimate the similarity between any two documents based on their estimated relevance to common sub-queries. Moreover, we make use of the associations between documents and the identified sub-queries in order to perform a richer re-ranking of the results retrieved for the original query. By doing so, we can take into account not only the estimated relevance of a document to the original query, but also the relative importance of the different aspects underlying this query, their coverage in the ranking, and how well the given document satisfies each of them.

3 The xQuAD Diversification Framework

In this section, we describe our novel framework for search result diversification, centred around the concept of sub-queries. The *eXplicit Query Aspect Diversification* (xQuAD) framework is inspired by the greedy approximation approach to the general diversification problem, which is at the heart of most of the previous works on search result diversification, as described in Section 2. Differently from these approaches, however, our framework performs an explicit diversification of the documents retrieved for a given query, by exploiting the relationship between these documents and the aspects uncovered from this query. In particular, we aim to promote a diverse ranking of documents according to the following four components: *aspect importance*, based on the relevance of each identified aspect with respect to the initial query; *document coverage*, based on the estimated relevance of a given document to multiple aspects; *document novelty*, based on the estimated relevance of the document to aspects not well represented among the already selected documents; and *document relevance*, based on the estimated relevance of the document to the initial query.

Aspect importance is discussed in Section 3.2. The document coverage and novelty components follow from the intuitive definition of an ideal diverse ranking, which should provide a broad coverage of the aspects underlying the initial query, while reducing its overall redundancy with respect to aspects already well covered. Just as for aspect importance, the effectiveness of these components depend on the quality of the aspects uncovered from the initial query in the form of sub-queries, as discussed in Section 3.1. As for the last component, namely, document relevance, we argue that it can help cope with the uncertainty associated with the relevance estimations for multiple sub-queries. Indeed, it provides a common basis for comparing documents retrieved for different sub-queries, as the relevance scores based on these sub-queries may not be comparable.

Our proposed framework integrates all these components into Algorithm 1. The algorithm takes as input the initial query q, the set $R(q)$ of documents retrieved for q, a set $Q(q)$ of sub-queries q_i derived from q, a scoring function $r(d, q)$ that estimates the relevance of a document d to a query q (analogously, $r(d, q_i)$ estimates the relevance of d to the sub-query q_i), and a sub-query importance estimator $i_X(q_i, q)$ (see Section 3.2). Additionally, it has two parameters: the number τ of top-ranked results from $R(q)$ to be returned, and the weight ω, used for balancing the influence of the relevance and diversity estimations.

The algorithm constructs a result set $S(q)$ by iteratively selecting a document d^* which contributes the most relevant and novel information among the remaining documents from the initial ranking, $R(q)$. The core of the algorithm is the computation of $r(d, q, Q(q))$ (lines 3-5), which combines the relevance score of d with respect to the query q, and a diversity score, computed as a summation over each of the sub-queries $q_i \in Q(q)$ that are satisfied by this document. In particular, the contribution of a given sub-query q_i to the diversity of document d takes into account: (1) the relative importance $i_X(q_i, q)$ of q_i in light of the query q, (2) the estimated relevance of d to q_i, $r(d, q_i)$, and (3) a measure of the novelty of any document satisfying q_i. The latter is given by $m(q_i)^{-1}$, with

```
**xQuAD**$[q, R(q), Q(q), r, i_X, \tau, \omega]$
1  $S(q) \leftarrow \emptyset$
2  while $|S(q)| < \tau$ do
3      for $d \in R(q)$ do
4          $r(d, q, Q(q)) \leftarrow r(d, q) \times \left(\sum_{q_i \in Q(q)} i_X(q_i, q) r(d, q_i)/m(q_i)\right)^{\omega}$
5      end for
6      $d^* \leftarrow \arg\max_d \; r(d, q, Q(q))$
7      for $q_i \in Q(q)$ do
8          $m(q_i) \leftarrow m(q_i) + r(d^*, q_i)$
9      end for
10     $R(q) \leftarrow R(q) \setminus \{d^*\}$
11     $S(q) \leftarrow S(q) \cup \{d^*\}$
12 end while
13 return $S(q)$
```

Alg. 1: The xQuAD framework.

$m(q_i)$ defined as the "mass" of information satisfying q_i that is already included in the final ranking, $S(q)$. After the top scored document d^* is selected at the end of each iteration (line 6), the information mass $m(q_i)$ is updated to account for the selection of this document from all the sub-queries it satisfies (lines 7-9). The selected document is then removed from $R(q)$ (line 10) and included in the final document ranking, $S(q)$ (line 11). At the end of the process, $S(q)$ is the final diverse ranking to be presented to the user (line 13).

3.1 Uncovering Query Aspects

An important component of our proposed diversification framework is the generation of sub-queries, in the form of keyword-based representations of the possible aspects underlying the initial query. Several techniques can be used for this purpose. For instance, we could mine a query log for common reformulations of the initial query [9], or use a large external corpus, such as Wikipedia, in order to obtain possible disambiguation terms [10]. Alternatively, sub-queries can be generated from the target collection itself, e.g., by uncovering the most salient phrases from the top retrieved results for a given query [11].

To validate our approach, we use a test collection with relevance assessments at the level of sub-topics. These sub-topics can be seen as a simulation of ground-truth sub-queries, as discussed in Section 4. This allows us to investigate the full potential of our approach, by focusing on how to effectively exploit sub-queries within the xQuAD diversification framework. Additionally, we propose a technique inspired by traditional text clustering, in order to generate sub-queries from the baseline ranking. Given a ranking of documents retrieved for the original query, a clustering algorithm is applied to partition this ranking into a predefined number of clusters. In our experiments, we use the k-means algorithm [12]. From each generated cluster, we select the most informative terms as a sub-query, using the Bo1 information-theoretic query expansion model from the Divergence From Randomness (DFR) framework [13].

3.2 Estimating Aspect Relative Importance

The importance of the different aspects underlying a given query should ultimately reflect the interests of the user population—i.e., information consumers—with respect to each of these aspects [3], e.g., based on the popularity of each corresponding sub-query in a query log. In the absence of such data, an alternative is to rely on sub-query importance as conveyed by information producers. To do so, we propose four different aspect importance estimators, which implement the $i_X(q_i, q)$ component in the xQuAD framework, as presented in Algorithm 1. The first one, $i_U(q_i, q)$, considers a uniform distribution of importance:

$$i_U(q_i, q) = \frac{1}{|Q(q)|}, \tag{1}$$

where $|Q(q)|$ is the total number of identified sub-queries. As a more refined estimator, we introduce $i_N(q_i, q)$, which considers the total number of results retrieved from the target collection for a particular sub-query q_i as an indication of the importance of q_i:

$$i_N(q_i, q) = \frac{n(q_i)}{\sum_{q_j \in Q(q)} n(q_j)}, \tag{2}$$

where $n(q_i)$ is the total number of results retrieved for q_i, and $Q(q)$ is the set of all sub-queries derived from the initial query q. Alternatively, inspired by resource selection techniques in distributed information retrieval [14], we devise richer importance estimators by considering the top retrieved documents for each sub-query as a sample from the resource represented by all the documents associated to that particular sub-query in the whole collection.

In particular, in this work, we investigate two effective resource selection algorithms as estimators of sub-query importance: Relevant Document Distribution Estimation (ReDDE) [15], and Central Rank-based Collection Selection (CRCS) [16]. The ReDDE algorithm estimates the number of relevant documents contained in a given resource based on the estimated size of this resource and on the number of its documents that are ranked above a certain threshold in a centralised ranking comprising samples from all resources. We devise a ReDDE-inspired sub-query importance estimator $i_R(q_i, q)$ as:

$$i_R(q_i, q) = \sum_{d | r(d, q_i) > 0} r(d, q) \times r(d, q_i) \times n(q_i), \tag{3}$$

where $r(d, q)$ is the estimated relevance of a document d with respect to the query q. Analogously, $r(d, q_i)$ estimates the relevance of d to the sub-query q_i. As above, $n(q_i)$ is the total number of results associated with q_i.

Besides $i_U(q_i, q)$, $i_N(q_i, q)$, and $i_R(q_i, q)$, we propose another way of estimating the relative importance of different sub-queries, inspired by the CRCS algorithm. CRCS ranks resources according to their estimated sizes, differing from other approaches—including ReDDE—by also considering the position (or rank) of each of the sampled documents in the centralised ranking of resource descriptions. The idea is that a document ranked higher should convey more importance of its resource than a document appearing towards the bottom of the ranking. Inspired by CRCS, we devise the $i_C(q_i, q)$ importance estimator as:

$$i_C(q_i, q) = \frac{n(q_i)}{\max_{q_j \in Q(q)} n(q_j)} \times \frac{1}{\hat{n}(q_i)} \sum_{d|r(d,q_i)>0} \tau - j(d, q), \qquad (4)$$

where $n(q_i)$ is as above, $\hat{n}(q_i)$ corresponds to the number of results associated to the sub-query q_i that are among the top τ ranked results for the query q, with $j(d, q)$ giving the ranking position of the document d with respect to q.

4 Experimental Setup

In this section, we describe our experimental setup, in order to support the investigation of the following research questions:

1. Is the explicit account of the aspects underlying a given query an effective approach for diversifying the results retrieved for this query?
2. Is our proposed framework an effective diversification approach when compared to other explicit diversification approaches?
3. Can we further improve the effectiveness of our framework by taking into account the relative importance of individual sub-queries?
4. Can we effectively derive sub-queries from the baseline ranking itself?

In the following, we detail the test collection, topics, and metrics used in our evaluation, as well as the diversification approaches to which ours is compared, including the procedure for training their parameters. The Terrier Information Retrieval platform[1] [17] is used for both indexing and retrieval.

4.1 Collection and Topics

In our experiments, we index the Financial Times of London 1991-1994, a standard test collection with 210,158 news articles. In particular, this collection was used in a diversity-oriented task investigated under the standard experimentation paradigm of the Text REtrieval Conference (TREC), as part of the TREC Interactive track in TREC-6, TREC-7, and TREC-8 [18]. The investigated task, then called "aspect retrieval", involved finding documents covering as many different aspects of a given query as possible. As part of this evaluation campaign, a total of 20 topics were adapted from the corresponding years of the TREC Adhoc track. Each topic includes from 7 to 56 sub-topics, as identified by TREC assessors, with relevance assessments provided at the sub-topic level. Figure 1 illustrates one of such topics, 353i, along with some of its identified aspects.

[1] `http://www.terrier.org`

```
<top>                                      353i-1 mining prospection
<num> Number: 353i                         353i-2 oil resources
<title> Antarctic exploration              353i-3 rhodium exploration
<desc>                                     353i-4 ozone hole / upper atmosphere
  Identify systematic explorations and     353i-5 greenhouse effect
  scientific investigations of Antarctica, 353i-6 measuring chemicals in the atmosphere
  current or planned.                      353i-7 analysis of toxic wast
</top>                                      ...
```

Fig. 1: TREC-7 Interactive track, topic 353i, and corresponding sub-topics.

In the example, several sub-topics were identified by the assessors for topic 353i, as shown on the right-hand side of Figure 1. In order to test the full benefit of our explicit diversification framework, we follow Zhai et al. [7] and derive ground-truth sub-queries based on the official aspects associated to each of the TREC Interactive track topics. As discussed in Section 3, this experimental design choice allows us to simulate a best-possible sub-query generation mechanism in order to focus our attention to evaluating the diversification framework itself.

4.2 Retrieval Baselines

We evaluate the effectiveness of our framework at diversifying the rankings produced by two effective document ranking approaches as retrieval baselines, namely, BM25 [19] and the DPH hypergeometric, parameter-free model from the DFR framework [20]. On top of the initial ranking produced by either of these baselines, we compare our framework to several other diversification approaches, namely, the previously described approaches of Carbonell and Goldstein [5], Radlinski and Dumais [9], and Agrawal et al. [4]. In particular, as the last two make use of external resources or judgements, such as query logs or a classification taxonomy, which are not available for the test collection at hand, we simulate their best-case scenario, by considering the ground-truth sub-topics provided by the collection as input to their proposed diversification models.

4.3 Evaluation Metrics

Our analysis is based on two evaluation metrics that reward diversity, namely, α-normalised discounted cumulative gain (α-NDCG) [3], and intent-aware mean average precision (MAP-IA) [4], reported at two different cutoffs: 10 and 100. α-NDCG balances relevance and diversity through the tuning parameter α. The larger its value, the more diversity is rewarded. In the opposite end, when $\alpha = 0$, this metric is equivalent to the normal NDCG [21]. Following Wang and Zhu [8], we use $\alpha = 0.5$, in order to give equal weights to either of these dimensions.

Differently from other evaluation metrics in the literature, MAP-IA also takes into account how well a given document satisfies each aspect underlying the initial query, as well as the relative importance of each aspect, as given by a ground-truth importance distribution. In our evaluation, we devise two variants of MAP-IA. The first of these is a uniform variant, u-MAP-IA, which considers all aspects underlying a query as equally important, so as to provide a fair ground for the approaches that do not take the aspect importance into account. The second proposed variant, i-MAP-IA, estimates an ideal importance distribution over aspects as the ratio of relevant documents that cover each aspect when compared to all documents judged relevant for the initial query, as given by the provided ground-truth relevance assessments. Note, however, that although our framework can take the relative importance of different aspects into account, it relies on different estimation mechanisms, as proposed in Section 3.2.

4.4 Training Settings

To train the interpolation parameter of our framework, as well as the one used by the approach of Carbonell and Goldstein [5], we perform a 5-fold cross validation over the 20 topics, optimising for u-MAP-IA. The approaches of Radlinski and Dumais [9] and Agrawal et al. [4] do not require training under their simulated best-case scenario. As for our proposed clustering-based query expansion technique to uncover sub-queries from the baseline ranking, we apply the k-means algorithm on the top 1000 retrieved documents. In particular, we use $k = 20$ (the average number of sub-topics per considered topic), and extract the 10 most informative terms from each generated cluster as a sub-query.

5 Experimental Evaluation

In this section, we evaluate our framework with respect to the research questions stated in Section 4. Table 1 shows the performance of xQuAD and several baseline approaches in terms of α-NDCG, u-MAP-IA, and i-MAP-IA. In Table 1, MMR stands for the maximal marginal relevance method of Carbonell and Goldstein [5], whereas the simulated approaches of Agrawal et al. [4] and Radlinski and Dumais [9] using the official TREC Interactive track sub-topics as input are referred to as IA-Select and QFilter, respectively. It is worth noting that although IA-Select can also consider the relative importance of the different aspects underlying the initial query, our simulated version does not take this information into account, as it is not trivial to derive an analogy for their classification scheme in this case, without relying on relevance assessments. Nonetheless, to provide for a fairer comparison, we report the performance of xQuAD using the uniform aspect importance estimator given by Equation (1). This variant of our framework is denoted xQuAD$_U$. All approaches are applied over the top 1000 documents retrieved by the underlying baseline ranking. Significance with respect to this ranking is given by the Wilcoxon signed-rank test. In particular, the superscript symbols $\blacktriangle$ ($\blacktriangledown$) and $\vartriangle$ ($\triangledown$) denote a significant increase (decrease) at the $p < 0.01$ and $p < 0.05$ levels, respectively. A second such symbol (subscript) denotes significance with respect to the strongest among the considered baseline diversification approaches.

From Table 1, recalling our first and second research questions, we observe that xQuAD$_U$ markedly outperforms all other diversification approaches across all settings, except for the u-MAP-IA metric, when DPH is used as the baseline ranking. This attests the effectiveness of our proposed framework when compared to both the implicit diversification performed by MMR and the explicit diversification provided by IA-Select and QFilter. Moreover, xQuAD$_U$ is the only

Table 1: Comparative performance with a uniform aspect importance estimator.

	α-NDCG		u-MAP-IA		i-MAP-IA	
	@10	@100	@10	@100	@10	@100
BM25	0.4505	0.5308	0.2286	0.1710	0.1416	0.1969
+MMR	0.4364	0.5102	0.2289	0.1700	0.1380	0.1841
+IA-Select	0.3392	0.4141	0.1592	0.1141	0.0868	0.1271
+QFilter	0.4509	0.5200	0.2300	0.1856	0.1416	0.1934
+xQuAD$_U$	**0.5727**$^{\blacktriangle}_{\blacktriangle}$	**0.6120**$^{\vartriangle}_{\blacktriangle}$	**0.2760**	**0.2240**	**0.1825**	**0.2235**
DPH	0.4633	0.5476	0.2464	0.1827	0.1620	0.2134
+MMR	0.4087$^{\blacktriangledown}$	0.4273$^{\blacktriangledown}$	**0.2876**	**0.2422**	0.1479	0.1805
+IA-Select	0.3585	0.4340	0.1765	0.1318	0.1029	0.1403
+QFilter	0.4634	0.5342	0.2466	0.1947	0.1620	0.2103
+xQuAD$_U$	**0.5935**$^{\blacktriangle}_{\blacktriangle}$	**0.6151**$^{\vartriangle}_{\vartriangle}$	0.2871	0.2371	**0.1998**	**0.2424**

approach to consistently improve over the baseline document rankings across all settings, with significant gains in terms of α-NDCG at both cutoffs. Indeed, all other approaches perform generally worse than the baseline rankings. In particular, the performance of IA-Select is disappointing, given its simulation with the ground-truth sub-topics. Nevertheless, these differences are not significant.

Next, we address our third research question, by assessing the impact of accounting for the relative importance of the different query aspects. Table 2 shows the performance of our framework using the different importance estimators introduced in Section 3.2. In particular, the subscript 'X' in xQuAD$_X$ reflects the use of the corresponding importance estimator $i_X(q_i, q)$, with $X \in \{U, N, R, C\}$ corresponding to Equations (1)-(4), respectively. Analogously to Table 1, a superscript symbol denotes statistical significance with respect to the baseline ranking, whereas a subscript symbol denotes significance with respect to xQuAD$_U$.

Table 2: Comparative performance using different aspect importance estimators.

	α-NDCG		u-MAP-IA		i-MAP-IA	
	@10	@100	@10	@100	@10	@100
BM25	0.4505	0.5308	0.2286	0.1710	0.1416	0.1969
+xQuAD$_U$	**0.5727**$^{\blacktriangle}$	0.6120$^{\vartriangle}$	0.2760	0.2240	0.1825	0.2235
+xQuAD$_N$	0.4856$^{\blacktriangle}_{\blacktriangledown}$	0.5666$^{\blacktriangle}$	0.2484	0.1919	0.1597	0.2142
+xQuAD$_R$	0.4796	0.5716	0.2715	0.2132	0.1728	0.2274
+xQuAD$_C$	0.5204	**0.6238**	**0.3815**$_{\vartriangle}$	**0.2622**	**0.1871**	**0.2387**
DPH	0.4633	0.5476	0.2464	0.1827	0.1620	0.2134
+xQuAD$_U$	**0.5935**$^{\blacktriangle}$	**0.6151**$^{\vartriangle}$	0.2871	0.2371	**0.1998**	**0.2424**
+xQuAD$_N$	0.4878$_{\blacktriangledown}$	0.5281$_{\triangledown}$	0.2649	0.2171	0.1720$_{\blacktriangledown}$	0.2213
+xQuAD$_R$	0.4695$_{\triangledown}$	0.5664	0.2684	0.2131$_{\triangledown}$	0.1696	0.2240
+xQuAD$_C$	0.4894	0.5812	**0.3099**	**0.2409**	0.1708$_{\triangledown}$	0.2270

From Table 2, we first note that the variants of our framework improve over the baseline rankings in all but one case (DPH+xQuAD$_N$, α-NDCG@100). As for our stated research question, these results show that further improvements can be attained by taking into account the estimated relative importance of the different aspects underlying a query. In particular, the results using the importance estimators inspired by resource selection techniques are promising, notably for the xQuAD$_C$ variant, which outperforms the uniform estimation variant according to all but the α-NDCG@10 metric over BM25, with gains of up to 38% in terms of u-MAP-IA. The improvements, however, are less marked when the DPH baseline is considered, in which case the variant using the uniform aspect importance estimator is surprisingly the best one. This suggests that the performance of the different variants can be highly influenced by the performance of the baseline ranking. Indeed, as the same retrieval technique is used to estimate the relevance of the retrieved documents with respect to each different aspect, it can directly impact the estimation of the importance of this aspect.

Lastly, we address the question of whether sub-queries can be effectively generated from the baseline ranking itself. This can be particularly useful in cases when additional resources, such as a query log or a taxonomy of categories over queries and documents, are not available. As discussed in Section 3.1, we propose a clustering-based query expansion technique, in an attempt to uncover terms representative of different aspects underlying a query from a clustering of the top retrieved results for this query. Table 3 shows the results of xQuAD$_U$ using sub-queries generated by the DFR Bo1 query expansion model.

Table 3: Performance using sub-queries generated from the target collection.

	α-NDCG		u-MAP-IA		i-MAP-IA	
	@10	@100	@10	@100	@10	@100
BM25	0.4505	**0.5308**	0.2286	0.1710	0.1416	**0.1969**
+xQuAD$_{U(Bo1)}$	**0.4509**	0.5193$^\blacktriangledown$	**0.2300**	**0.1742**	0.1416	0.1919$^\blacktriangledown$
DPH	0.4633	**0.5476**	0.2464	0.1827	0.1620	**0.2134**
+xQuAD$_{U(Bo1)}$	**0.4634**	0.5226$^\blacktriangledown$	**0.2466**	**0.1906**$^\vartriangle$	0.1620	0.2084$^\blacktriangledown$

From Table 3, as expected, we first observe that the obtained performances are much lower than those observed for the variants of our framework using the ground-truth sub-topics, as shown in Table 2. Nevertheless, they are comparable to the performances attained by our competing diversification approaches using the ground-truth sub-topics, as shown in Table 1. This suggests that investigating alternative methods for sub-query generation is a promising direction for further enhancing the performance of our framework.

6 Conclusions and Future Work

In this paper, we have proposed a novel framework for search result diversification. Given an initial set of documents retrieved for a query, the *eXplicit Query Aspect Diversification* (xQuAD) framework produces a diverse ranking by considering the relationship between the retrieved documents and the possible aspects underlying the query, explicitly modelled as sub-queries.

Using a standard test collection for the evaluation of diversity, we have shown that our framework is effective and can markedly outperform state-of-the-art diversification approaches, either implicit or explicit. Moreover, by estimating the relative importance of each of the identified aspects of a given query, we have shown that further improvements can be attained. Finally, we have proposed a clustering-based query expansion technique to demonstrate the feasibility of automatically generating sub-queries from the baseline document ranking itself.

Overall, our results attest the effectiveness of the xQuAD framework for search result diversification. As identifying meaningful sub-queries and estimating their relative importance are challenging problems in themselves, we plan to carry on our investigations in these directions, and also to evaluate our framework in a broader search scenario, such as the Web.

References

1. Spärck-Jones, K., Robertson, S.E., Sanderson, M.: Ambiguous requests: implications for retrieval tests, systems and theories. SIGIR Forum **41**(2) (2007) 8–17
2. Robertson, S.E.: The probability ranking principle in IR. Journal of Documentation **33**(4) (1977) 294–304
3. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR. (2008) 659–666
4. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM. (2009) 5–14
5. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. (1998) 335–336
6. Hochbaum, D.S., ed.: Approximation algorithms for NP-hard problems. PWS Publishing Co. (1997)
7. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: SIGIR. (2003) 10–17
8. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: SIGIR. (2009) 115–122
9. Radlinski, F., Dumais, S.: Improving personalized web search using result diversification. In: SIGIR. (2006) 691–692
10. Mihalcea, R.: Using Wikipedia for automatic word sense disambiguation. In: HLT-NAACL. (2007) 196–203
11. Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y., Ma, J.: Learning to cluster web search results. In: SIGIR. (2004) 210–217
12. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Berkeley SMSP. (1967) 281–297
13. Amati, G.: Probability models for information retrieval based on Divergence From Randomness. PhD thesis, University of Glasgow (2003)
14. Callan, J.: Distributed information retrieval. In Croft, W.B., ed.: Advances in Information Retrieval. Kluwer Academic Publishers (2000) 127–150
15. Si, L., Callan, J.: Relevant document distribution estimation method for resource selection. In: SIGIR. (2003) 298–305
16. Shokouhi, M.: Central-rank-based collection selection in uncooperative distributed information retrieval. In: ECIR. (2007) 160–172
17. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: a high performance and scalable information retrieval platform. In: SIGIR/OSIR. (2006)
18. Hersh, W., Over, P.: TREC-8 Interactive track report. In: TREC. (2000)
19. Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gull, A., Lau, M.: Okapi at TREC. In: TREC. (1992)
20. Amati, G., Ambrosi, E., Bianchi, M., Gaibisso, C., Gambosi, G.: FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog track. In: TREC. (2007)
21. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM TOIS **20**(4) (2002) 422–446