Inf Retrieval
DOI 10.1007/s10791-011-9180-x

On the role of novelty for search result diversification

Rodrygo L. T. Santos · Craig Macdonald · Iadh Ounis

Received: 3 June 2011 / Accepted: 22 November 2011

Abstract Re-ranking the search results in order to promote novel ones has traditionally been regarded as an intuitive diversification strategy. In this paper, we challenge this common intuition and thoroughly investigate the actual role of novelty for search result diversification, based upon the framework provided by the diversity task of the TREC 2009 and 2010 Web tracks. Our results show that existing diversification approaches based solely on novelty cannot consistently improve over a standard, non-diversified baseline ranking. Moreover, when deployed as an additional component by the current state-of-the-art diversification approaches, our results show that novelty does not bring significant improvements, while adding considerable efficiency overheads. Finally, through a comprehensive analysis with simulated rankings of various quality, we demonstrate that, although inherently limited by the performance of the initial ranking, novelty plays a role at breaking the tie between similarly diverse results.

Keywords Web search · Relevance · Diversity

1 Introduction

The assumption that a query unambiguously defines the user's information need does not always hold in a Web search scenario (Spärck-Jones et al. 2007; Sanderson 2008). Typical user queries bear some degree of ambiguity (Song et al. 2009). While truly ambiguous queries (e.g., '*office*') are open to different interpretations (e.g., '*business room*', '*software suite*', '*tv show*'), queries with a clearly defined interpretation (e.g., '*the office tv show*') might still be open to different aspects of this interpretation (e.g., '*schedule*', '*episode*

R. L. T. Santos (✉) · C. Macdonald · I. Ounis
School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK
e-mail: rodrygo@dcs.gla.ac.uk

C. Macdonald
e-mail: craigm@dcs.gla.ac.uk

I. Ounis
e-mail: ounis@dcs.gla.ac.uk

Published online: 21 December 2011

guide', '*cast*') (Clarke et al. 2008). An effective approach for tackling ambiguous queries is to diversify the search results, so as to maximise the chance that different users will find at least one relevant result to their particular need (Agrawal et al. 2009).

A classical diversification strategy consists in comparing the retrieved results to one another, in order to promote *novelty* in the ranking (Carbonell and Goldstein 1998; Zhai et al. 2003; Wang and Zhu 2009; Rafiei et al. 2010). In particular, novelty-based diversification approaches *implicitly* assume that different results will cover different aspects of the query, and hence should be promoted in the ranking. While classical approaches deploy novelty as their sole ranking strategy, the state-of-the-art approaches deploy a hybrid strategy. In particular, the latter approaches seek to promote not only novel search results, but also results with a high *coverage*[1] of the aspects underlying the initial query (Agrawal et al. 2009; Carterette and Chandar 2009; Santos et al. 2010a, c). This strategy is enabled by an *explicit* representation of the query aspects, in contrast to the implicit aspect representation adopted by the existing novelty-based approaches.

Unfortunately, the prevalence of different aspect representations has precluded a direct comparison between coverage and novelty as diversification strategies. As a result, it remains unclear whether the striking difference in performance commonly observed between coverage and novelty-based approaches is due to their underlying aspect representation (explicit vs. implicit) or to their diversification strategy (coverage vs. novelty). It is also unclear how much novelty actually contributes to the effectiveness of the current state-of-the-art approaches, while penalising their efficiency—differently from novelty, the coverage of a search result is estimated independently of other results. Although intuitive, novelty has yet to be shown effective for diversifying Web search results. In particular, existing evidence of the effectiveness of novelty as a diversification strategy is based on either qualitative studies (Carbonell and Goldstein 1998) or on curated corpora, such as Wikipedia (Rafiei et al. 2010) or newswire (Wang and Zhu 2009).

In this paper, we challenge the common view of novelty as an intuitive diversification strategy. To this end, we thoroughly investigate the role of this strategy in light of both classical as well as state-of-the-art diversification approaches in the literature. To enable our investigation, we adapt two existing novelty-based diversification approaches to leverage explicit query aspect representations. Likewise, we produce coverage-only versions of two state-of-the-art approaches that deploy a hybrid of coverage and novelty strategies. By doing so, we bridge the gap between the diversification approaches in the literature and enable their evaluation in terms of the aspect representation and the diversification strategy dimensions. As a result, we provide the first comprehensive account of the role of novelty as a ranking strategy for diversifying Web search results.

Using the evaluation framework provided by the diversity task of the TREC 2009 and 2010 Web tracks (Clarke et al. 2009, 2010), we empirically show that novelty cannot consistently improve over a standard, non-diversified baseline ranking. When leveraging explicit aspect representations (including a 'ground-truth' aspect representation), we show that novelty-based approaches can be improved, but are still not significantly more effective than a non-diversified ranking. On the diversification strategy dimension, we find that novelty does not contribute significantly to the coverage-based strategy deployed by the current state-of-the-art, suggesting that the efficiency overhead added by promoting novelty does not pay off. Finally, through a scrutinous analysis based on simulated

[1] Clarke et al. (2008) refer to this concept as '*diversity*'. We call it '*coverage*' to avoid any confusion with the task name.

rankings of various quality, we demonstrate that, under special conditions, novelty can still play a role at breaking the tie between results with similar coverage.

In summary, the major contributions of this paper are:

1. A unifying framework to enable the direct comparison of existing diversification approaches across the aspect representation and diversification strategy dimensions;
2. A thorough investigation of the impact of different aspect representations and diversification strategies for search result diversification;
3. A comprehensive analysis of the role of novelty as a diversification strategy, under a range of empirical and simulated relevance scenarios.

The remainder of this paper is organised as follows. In Sect. 2, we provide background on search result diversification and on representative approaches from the literature. In Sect. 3, we describe the methodology that supports our investigations. Our experimental setup is detailed in Sect. 4, while Sects. 5 and 6 discuss the results of our investigation, based on empirical and simulated experiments, respectively. Finally, in Sect. 7, we present our concluding remarks and directions for future research.

2 Background and related work

In 1964, Goffman (1964) pointed out that *'the relationship between a document and a query is necessary but not sufficient to determine relevance'*. Later, in 1991, Gordon and Lenk (1991) discussed two assumptions underlying the probability ranking principle (Cooper 1971; Robertson 1977), namely, that relevance is determined with certainty, and that documents are judged relevant or not independently of one another. Since then, several approaches have been proposed to overcome these limiting assumptions. Among these, *search result diversification* tackles the uncertainty of relevance estimates, primarily resulting from query ambiguity, by promoting documents with maximum *coverage* of the possible aspects underlying a query. Additionally, it accounts for the dependent relevance of documents by promoting those documents with maximum *novelty* with respect to the already selected ones.

The diversification approaches in the literature can be classified according to two complementary dimensions: *aspect representation* and *diversification strategy* (Santos et al. 2010c). The aspect representation determines how a document is described in light of the several aspects underlying a query. In particular, an *implicit* representation describes a document regardless of the query aspects, based on features intrinsic to the document (e.g., the terms it contains). In turn, an *explicit* representation describes how well a document covers the query aspects, where each aspect can be itself represented in a variety of ways. For instance, different aspects can represent different query classes according to a pre-defined taxonomy (Agrawal et al. 2009) or different topics covered by the retrieved documents (Carterette and Chandar 2009). More generally, different aspects can represent multiple information needs underlying the query, e.g., as different query reformulations (Radlinski and Dumais 2006; Santos et al. 2010a).

Complementarily to the aspect representation, the diversification strategy determines how a diversification approach achieves the goal of satisfying different aspects of a query. *Coverage*-based approaches achieve this goal by directly estimating how well each document covers each aspect of the query, regardless of the other retrieved documents. Alternative estimates of coverage depend on the adopted aspect representation and include classification confidence (Agrawal et al. 2009), topicality (Carterette and Chandar 2009),

Table 1 An overview of representative search result diversification approaches in the literature, organised in terms of two dimensions: diversification strategy and query aspect representation

		Query Aspect Representation	
		Implicit	Explicit
Diversification Strategy	Novelty	Carbonell and Goldstein (1998) Zhai et al (2003) Chen and Karger (2006) Wang and Zhu (2009) Rafiei et al (2010)	*Our methodology to enable a cross-dimension analysis*
	Coverage	N/A	Radlinski and Dumais (2006) Carterette and Chandar (2009)
	Coverage +Novelty	N/A	Agrawal et al (2009) Santos et al (2010a,c)

and relevance (Santos et al. 2010a, c). A different diversification strategy exploits the relationships among the retrieved documents. In particular, *novelty*-based approaches directly compare the retrieved documents to one another, in order to promote those that convey novel information (i.e., information not conveyed by the other retrieved documents). Existing approaches differ mostly in how they identify novel information. For instance, novelty can be estimated based on content dissimilarity (Carbonell and Goldstein 1998), divergence (Zhai et al. 2003), conditioned relevance (Chen and Karger 2006), or relevance score correlation (Rafiei et al. 2010; Wang and Zhu 2009).

Table 1 organises the most representative diversification approaches in the literature according to the aspect representation and diversification strategy dimensions. In particular, coverage (Carterette and Chandar 2009) and hybrid (i.e., coverage + novelty) approaches (Santos et al. 2010a, c) have been shown to substantially outperform pure novelty-based ones. On the other hand, to the best of our knowledge, novelty has only been tested on qualitative studies (Carbonell and Goldstein 1998) or on curated corpora such as Wikipedia (Rafiei et al. 2010) or newswire (Wang and Zhu 2009), with its effectiveness in a Web search result diversification scenario yet to be proven. Moreover, while hybrid approaches constitute the current state-of-the-art (Clarke et al. 2009; Clarke et al. 2010), it is unclear how much of their effectiveness comes from also promoting novelty. To address these questions, Sect. 3 describes our research methodology. The results of our thorough experimentation are discussed in Sects. 5 and 6 and unveil the role of novelty as a diversification strategy.

3 Bridging the gap

The objectives of search result diversification are two-fold: (1) to maximise the number of query aspects covered in the ranking, and (2) to avoid excessive redundancy among the covered aspects. Finding a subset of the retrieved documents with maximum coverage (or, similarly, minimum redundancy) with respect to the query aspects is an instance of the Maximum Coverage Problem[2] (Hochbaum 1997), and is therefore NP-hard (Agrawal et al.

[2] The Maximum Coverage Problem can be stated as: *given n sets and a number Ω, select at most Ω sets so that the maximum number of elements is covered*. In the context of search result diversification, a *document* represents a set, and a *query aspect* represents an element.

2009). Most of the diversification approaches in the literature deploy a greedy approximation algorithm for this problem. From an initial ranking $\mathcal{R}$, this algorithm builds a ranking $\mathcal{S}$, by iteratively selecting a document d^* such that:

$$d^* = \arg\max_{d \in \mathcal{R} \setminus \mathcal{S}} score(d, q, \mathcal{A}, \mathcal{S}), \tag{1}$$

where $score(d, q, \mathcal{A}, \mathcal{S})$ is typically computed as a trade-off between the estimated *relevance* of d given the query q, and the *diversity* of d given some representation of the aspects $\mathcal{A}$ underlying q and the documents in $\mathcal{S}$, which were selected in the previous iterations of the algorithm (Santos et al. 2010b).

Although having the same goal of producing a diverse ranking, coverage and novelty-based approaches implement the above objective function in different ways. While purely coverage-based approaches typically ignore the set of already selected documents $\mathcal{S}$, existing novelty-based approaches ignore the set of query aspects $\mathcal{A}$. In practice, this renders coverage and novelty, as implemented by existing approaches, not directly comparable. In this section, we describe our methodology to bridge the gap between these approaches and enable their direct comparison. Besides evaluating novelty in contrast to and in combination with coverage, our goal is to isolate these strategies from their underlying aspect representation, so as to provide a controlled setting for our investigations. To this end, in Sect. 3.1, we propose adaptations of two implicit novelty-based diversification approaches to leverage explicit aspect representations. Additionally, in Sect. 3.2, we deconstruct two explicit hybrid approaches to deploy a coverage-based strategy only.

3.1 Explicit novelty-based diversification

Existing novelty-based diversification approaches rely on an implicit aspect representation to estimate the diversity of a document with respect to the other retrieved documents (Carbonell and Goldstein 1998; Zhai et al. 2003; Wang and Zhu 2009). As a result, these approaches compare documents purely on the basis of their content, rather than based on how they satisfy different query aspects. Moreover, the resulting document representation (e.g., in the term-frequency space of a given corpus) is usually high-dimensional, which negatively impacts both the effectiveness and the efficiency of these approaches (Manning et al. 2008). To counter these limitations and—more importantly for this work—to enable a direct comparison of existing diversification approaches across both the aspect representation and the diversification strategy dimensions, we propose to leverage explicit aspect representations for estimating novelty. Besides providing a more expressive account of the relationship between documents and the aspects they cover, this representation also has a considerable impact on efficiency, since the feature space is reduced from the size of the corpus vocabulary (millions) to the number of aspects underlying a query (around a dozen).

Given a query q with a set of aspects $\mathcal{A}$, with $|\mathcal{A}| = k$, we explicitly represent each retrieved document $d \in \mathcal{R}$ as a k-dimensional vector $\boldsymbol{d}$ over $\mathcal{A}$. In particular, the mth dimension of the vector $\boldsymbol{d}$ is defined as:

$$\boldsymbol{d}_m = f(d, a_m), \tag{2}$$

where the function f estimates how well the document d satisfies the aspect $a_m \in \mathcal{A}$. As we will show in Sect. 4.3, different measures of the document-aspect association can be used, depending on how the aspects are identified, e.g., based on reformulations mined from a query log or on categories derived from a classification taxonomy.

Regardless of the particular mechanism used to identify the aspects of a query, an explicit representation of documents with respect to these aspects can be seamlessly integrated into existing novelty-based diversification approaches. In particular, to enable our analysis in Sects. 5 and 6, we derive explicit versions of two well-known novelty-based approaches in the literature, namely, Maximal Marginal Relevance (MMR, Carbonell and Goldstein 1998) and Mean-Variance Analysis (MVA, Wang and Zhu 2009).

MMR (Carbonell and Goldstein 1998) instantiates the scoring function in Eq. 1 by estimating the similarity between $d \in \mathcal{R} \setminus \mathcal{S}$ and its most dissimilar document $d_j \in \mathcal{S}$. Likewise, we devise xMMR (Explicit Maximal Marginal Relevance) to estimate novelty over explicit representations of the retrieved documents:

$$score_{\text{xMMR}}(d, q, \mathcal{A}, \mathcal{S}) = \lambda \text{sim}_1(d, q) - (1 - \lambda) \max_{\boldsymbol{d}_j \in \mathcal{S}} \text{sim}_2(\boldsymbol{d}, \boldsymbol{d}_j), \tag{3}$$

where $\text{sim}_1(d,q)$ and $\text{sim}_2(\boldsymbol{d}, \boldsymbol{d}_j)$ estimate the relevance of d to the query q and its similarity to the documents already in $\mathcal{S}$, respectively. A balance between relevance (i.e., sim_1) and redundancy (i.e., $\max\text{sim}_2$, the opposite of novelty) is achieved through an appropriate setting of λ, as will be described in Sect. 4.5. In our experiments, $\text{sim}_1(d,q)$ is estimated by a standard retrieval model. Following Carbonell and Goldstein (1998), we compute $\text{sim}_2(\boldsymbol{d}, \boldsymbol{d}_j)$ as the cosine between explicit representations of $\boldsymbol{d}$ and $\boldsymbol{d}_j$ over the aspects $\mathcal{A}$.

Analogously to MMR, MVA (Wang and Zhu 2009) instantiates Eq. 1 by trading off relevance and redundancy. However, instead of computing the similarity between documents, MVA estimates the redundancy of a document based on how its relevance scores correlate to those of the other documents. Accordingly, we devise xMVA (Explicit Mean-Variance Analysis) to estimate these correlations based on how well the documents satisfy the explicitly represented query aspects. The objective function of xMVA is defined according to the following equation:

$$score_{\text{xMVA}}(d, q, \mathcal{A}, \mathcal{S}) = \mu_{(d)} - b w_i \sigma^2_{(d)} - 2b \sum_{d_j \in \mathcal{S}} w_j \sigma_{(d_j)} \sigma_{(d)} \rho_{(\boldsymbol{d}, \boldsymbol{d}_j)}, \tag{4}$$

where $\mu_{(d)}$ and $\sigma^2_{(d)}$ represent the mean and variance of the relevance estimates associated to document d, respectively, while the summation component estimates the redundancy of document d in light of the documents in $\mathcal{S}$. In particular, documents are compared in terms of their correlation $\rho_{(\boldsymbol{d}, \boldsymbol{d}_j)}$. A balance between relevance, variance, and redundancy is achieved through the parameter b. Following Wang and Zhu (2009), $\mu_{(d)}$ is estimated by a standard retrieval model, with relevance scores normalised to yield a probability distribution. Additionally, $\sigma_{(d)}$ is set as a constant for all documents. In our experiments, both σ and b are set through training, as will be described in Sect. 4.5. Finally, $\rho_{(\boldsymbol{d}, \boldsymbol{d}_j)}$ is estimated as the Pearson's correlation between explicit representations of $\boldsymbol{d}$ and $\boldsymbol{d}_j$ over the aspects $\mathcal{A}$.

3.2 Explicit coverage-based diversification

Besides making coverage and novelty directly comparable by introducing explicit novelty-based diversification approaches (i.e., xMMR and xMVA), we want to be able to assess the effectiveness of novelty when combined with coverage. To this end, we deconstruct two state-of-the-art diversification approaches, IA-Select (Agrawal et al. 2009) and xQuAD (Santos et al. 2010a), which deploy a hybrid of coverage and novelty. Our goal is to produce directly comparable versions of these approaches, which should deploy coverage as their only strategy.

IA-Select (Agrawal et al. 2009) was originally proposed to diversify the search results according to a predefined taxonomy, such as the one provided by the Open Directory Project (ODP). Its objective function is defined as:

$$score_{\text{IA-Select}}(d, q, \mathcal{A}, \mathcal{S}) = \sum_{a_m \in \mathcal{A}} u(a_m|q, \mathcal{S})v(d|q, a_m), \tag{5}$$

where the function u estimates the *marginal utility* of the query aspect a_m given the query q and the documents already selected in $\mathcal{S}$, and the function v estimates the coverage of d with respect to q and a_m. The marginal utility u incorporates both the relative importance of the aspect a_m in light of all aspects $\mathcal{A}$ of the query q, as well as the current utility of a_m, in light of the aspects already covered by the documents in $\mathcal{S}$. In practice, the function u emulates a novelty component, by estimating how much the already selected documents satisfy each aspect of the query. To produce a coverage-only version of IA-Select, we assume that the query aspects do not lose their utility even if they are already covered by the documents in $\mathcal{S}$. In practice, this is achieved simply by dropping the term $\mathcal{S}$ in Eq. 5:

$$score_{\text{IA-Select}^*}(d, q, \mathcal{A}, \mathcal{S}) = \sum_{a_m \in \mathcal{A}} u(a_m|q)v(d|q, a_m). \tag{6}$$

To emphasise its difference from the standard IA-Select in Eq. 5, we call this coverage-only version IA-Select*.

Different from IA-Select, xQuAD (Santos et al. 2010a) implements the objective function in Eq. 1 as a mixture of probabilities:

$$score_{\text{xQuAD}}(d, q, \mathcal{A}, \mathcal{S}) = (1 - \lambda)\mathrm{P}_R(d|q) + \lambda\mathrm{P}_D(d, \bar{\mathcal{S}}|q), \tag{7}$$

where $\mathrm{P}_R(d|q)$ denotes the probability of d being relevant given the query q and $\mathrm{P}_D(d, \bar{\mathcal{S}}|q)$ denotes the probability of d but none of the documents already selected in $\mathcal{S}$ being diverse given q. These two probabilities are mixed using the parameter λ, which implements a trade-off between promoting relevant and diverse documents (Santos et al. 2010b). By marginalising over the possible aspects of q, the probability $\mathrm{P}_D(d, \bar{\mathcal{S}}|q)$ can be further broken down as:

$$\mathrm{P}_D(d, \bar{\mathcal{S}}|q) = \sum_{a_m \in \mathcal{A}} \mathrm{P}_D(a_m|q)\mathrm{P}_D(d|q, a_m)\mathrm{P}_D(\bar{\mathcal{S}}|q, a_m), \tag{8}$$

where $\mathrm{P}_D(a_m|q)$ denotes the importance of the aspect a_m given the query q, $\mathrm{P}_D(d|q, a_m)$ denotes the coverage of d given q and a_m, and $\mathrm{P}_D(\bar{\mathcal{S}}|q, a_m)$ denotes the novelty of any document satisfying a_m, based on the probability that none of the documents in $\mathcal{S}$ satisfy this aspect. Analogously to our adaptation of IA-Select, we introduce a coverage-only version of xQuAD by assuming that all query aspects retain their utility, regardless of the documents previously selected in $\mathcal{S}$. In practice, this is achieved simply by dropping the probability of novelty $\mathrm{P}_D(\bar{\mathcal{S}}|q, a_m)$, which produces xQuAD*:

$$\begin{aligned} score_{\text{xQuAD}^*}(d, q, \mathcal{A}, \mathcal{S}) = {} & (1 - \lambda)\mathrm{P}_R(d|q) \\ & + \lambda \sum_{a_m \in \mathcal{A}} \mathrm{P}_D(a_m|q)\mathrm{P}_D(d|q, a_m). \end{aligned} \tag{9}$$

Note that, without a novelty component, the coverage-only objective functions of both IA-Select* (Eq. 6) and xQuAD* (Eq. 9) no longer require an iterative, greedy diversification strategy. In practice, for an initial ranking of n documents, we reduce the cost of estimating Eq. 1 from $O(n)$ to $O(1)$. In Sects. 5 and 6, we evaluate all these approaches, in order to investigate the role of novelty when deployed in isolation, as well as when combined with coverage in a hybrid strategy.

4 Experimental setup

In this section, we describe the setup that supports our investigations in Sects. 5 and 6. These investigations aim to answer the following questions:

1. Is novelty an effective diversification strategy, and can it be improved with an explicit aspect representation?
2. How does an explicit novelty strategy perform in contrast to and in combination with a coverage strategy?
3. What is the role of novelty as a diversification strategy?

We address the first two research questions in Sect. 5. To answer the first question, we fix the diversification strategy dimension to novelty, in order to evaluate the impact of different aspect representations. Conversely, to tackle the second question, we fix the aspect representation dimension to different explicit representations and measure the effectiveness of novelty in contrast to and in combination with coverage. Finally, to provide further insights into the role of novelty as a search result diversification strategy, in Sect. 6, we answer the third question, by thoroughly evaluating this strategy with simulated rankings of various quality. The remainder of this section describes the experimental setup that supports all these investigations.

4.1 Collection and topics

Our investigations are conducted within the standard experimentation framework of the diversity task of the TREC 2009 and 2010 Web tracks (Clarke et al. 2009, 2010), henceforth referred to as WT09 and WT10 tasks, respectively. These tasks provide a total of 98 queries (50 for WT09, 48 for WT10), sampled from the query log of a commercial search engine. For each query, TREC assessors identified multiple sub-topics, representing different aspects of the initial query, with relevance assessments conducted at the sub-topic level. As the document corpus, we use the category-B subset of the TREC ClueWeb09 corpus (henceforth ClueWeb09 B), as used in the WT09 and WT10 tasks. In our experiments, this 50-million Web document corpus is indexed using Terrier,[3] with Porter's stemmer and standard stopword removal.

4.2 Retrieval approaches

To verify the consistency of our results, we experiment with several retrieval approaches under a uniform setting. As an ad hoc retrieval approach, which does not perform diversification, we use the Divergence From Randomness DPH model (Amati et al. 2007). Besides being effective, DPH is a parameter-free probabilistic model, and hence requires no training. On top of DPH, we experiment with diversification approaches representative of the novelty and coverage strategies. In particular, these approaches directly leverage the scores produced by DPH as their underlying 'relevance' estimation, as discussed in Sect. 3. As novelty-based approaches, we use MMR (Carbonell and Goldstein 1998) and MVA

[3] http://terrier.org

(Wang and Zhu 2009), as well as their explicit variants, xMMR and xMVA, introduced in Sect. 3.1. As coverage-based approaches, we consider our variants IA-Select* and xQuAD*, from Sect. 3.2. Their standard versions, namely, IA-Select (Agrawal et al. 2009) and xQuAD (Santos et al. 2010a), are used as hybrid approaches, and are representative of the state-of-the-art. Indeed, an instance of xQuAD attained the top performance in the diversity task of the TREC 2009 and 2010 Web tracks (cat. B) (Clarke et al. 2009; Clarke et al. 2010). Following Zhai et al. (2003), to cope with the quadratic pairwise comparisons performed by novelty-based approaches, both novelty, coverage, and hybrid approaches are applied to diversify the top 100 documents retrieved by DPH.

4.3 Aspect representations

To analyse the impact of different aspect representations, we compare a traditional implicit representation of documents in the space of the terms in the ClueWeb09 B corpus to four explicit aspect representations, described in the remainder of this section. Additionally, Table 2 summarises these explicit representations in terms of the average number of aspects identified for the WT09 and WT10 queries. For keyword-based aspect representations (i.e., WS, WR, and WT in Table 2), we also show the average length (in tokens) of each query aspect, and the average overlap between each aspect and the initial query, measured as the fraction of unique query terms covered by the aspect.

Our first explicit aspect representation (DZ in Table 2) was proposed by Agrawal et al. (2009), and corresponds to the 15 top-level categories from the Open Directory Project (ODP): adult, arts, business, computers, games, health, home, news, recreation, reference, regional, science, shopping, society, and sports. In particular, each document is represented as a k-dimensional vector, with each dimension corresponding to the probability that the document belongs to a category. Following Agrawal et al. (2009), this probability is estimated by the cosine between the document and the centroid representing the category, according to a Rocchio classifier (Manning et al. 2008). To obtain a centroid for each category, we randomly select 3,000 documents from the ClueWeb09 B corpus that belong exclusively to this category in ODP.

Our second and third aspect representations were proposed by Santos et al. (2010a). In particular, for each of the WT09 and WT10 queries, we obtain two sets of query reformulations from a commercial search engine: suggested queries (WS, displayed in the search engine's search box) and related queries (WR, displayed alongside the search engine's results). For each set with k aspects, we represent a document as a k-dimensional vector, with each dimension (i.e., the function f in Eq. 2) corresponding to the estimated relevance of the document to a different reformulation. To ensure this estimation is consistent with the one produced for the initial query, both are given by DPH.

Table 2 Statistics of the explicit query aspect representations used in this paper

$\mathcal{A}$	Aspects per query		Aspect length		Query overlap	
	WT09	WT10	WT09	WT10	WT09	WT10
DZ	15.00	15.00	N/A	N/A	N/A	N/A
WS	9.18	9.82	3.37	3.49	0.98	0.98
WR	19.90	19.50	2.40	2.37	0.50	0.55
WT	4.86	4.34	9.25	8.37	0.55	0.46

Finally, as a 'ground-truth' aspect representation (WT), we represent the retrieved documents in the space of the sub-topics identified by TREC assessors for each of the WT09 and WT10 queries (Clarke et al. 2009, 2010). In particular, these sub-topics provide a reference performance for the other explicit aspect representations used in our investigation. Analogously to using query reformulations from a commercial search engine, the retrieved documents are represented as k-dimensional vectors, with each dimension denoting the estimated relevance of a document to a TREC sub-topic, once again according to DPH. Additionally, the availability of relevance assessments for these 'ground-truth' aspects enables the evaluation of coverage and novelty using diversity estimates of various simulated quality, as we will show in Sect. 6.

4.4 Evaluation metrics

To evaluate the various approaches investigated in this paper, we use the two primary metrics in the diversity task of the TREC 2010 Web track (Clarke et al. 2010): ERR-IA and α-nDCG. The Intent-Aware Expected Reciprocal Rank (ERR-IA) metric (Chapelle et al. 2009) implements a cascade user model (Craswell et al. 2008), which penalises redundancy across multiple query aspects, by assuming that users will stop examining the result list once they find relevant information. The α-normalised Discounted Cumulative Gain (α-nDCG) metric (Clarke et al. 2008) extends the traditional nDCG (Järvelin and Kekäläinen 2002), with a parameter α that controls how much redundancy should be penalised. This tunable parameter is particularly suited for our investigation, as it allows the evaluation of novelty in an extreme scenario ($\alpha = 1$), which models a user with no tolerance to redundancy (Clarke et al. 2008). Both ERR-IA and α-nDCG have been shown to reward rankings that achieve a balance of coverage and novelty (Clarke et al. 2011). Moreover, α-nDCG has been shown to possess a discriminative power at least as high as that of the traditional nDCG (Sakai and Song 2011). Following the standard TREC setting, unless otherwise noted, both metrics are reported at rank cutoff 20 (Clarke et al. 2010). It is worth noting, however, that the observed trends are consistent across different rank cutoffs up to 100.

4.5 Training procedure

Most approaches in our evaluation require some parameter tuning. The exceptions are DPH, IA-Select (Agrawal et al. 2009), and IA-Select*, which are parameter-free. In order to train the parameters of the other approaches (i.e., MMR (Carbonell and Goldstein 1998) and xMMR's λ, MVA (Wang and Zhu 2009) and xMVA's b and σ, and xQuAD* and xQuAD's λ (Santos et al. 2010a), we use the WT09 and WT10 topics as training and test sets, in a cross-year fashion—i.e., we train on WT09 and test on WT10, and vice versa. All parameters are optimised through simulated annealing (Kirkpatrick et al. 1983), to maximise ERR-IA@100 on the training topics. To ensure our conclusions are not limited by the available training data, besides reporting our results on the test topics, we also report the training performance of all approaches.

5 Empirical evaluation

In this section, we address our first two research questions through an empirical evaluation within the framework provided by the TREC 2009 and 2010 Web tracks (Clarke et al. 2009, 2010). In particular, Sect. 5.1 covers our first question, to assess the effectiveness of novelty-based approaches across implicit and explicit aspect representations. Sections 5.2 and 5.3 address our second research question, by further investigating how novelty performs in contrast to and in combination with coverage across multiple aspect representations.

5.1 Implicit versus explicit novelty

To answer our first question, we contrast novelty-based diversification approaches based on implicit and explicit aspect representations. In particular, we aim to investigate not only whether existing approaches can be improved with a more refined aspect representation, but also whether any of these representations can improve over a standard, non-diversified baseline. Table 3 shows the training and test diversification performances of MMR and MVA (as implicit novelty-based approaches), as well as their explicit counterparts (xMMR and xMVA, respectively) in terms of ERR-IA and α-nDCG. The latter approaches are deployed with the four explicit representations described in Sect. 4.3: ODP categories (DZ) (Agrawal et al. 2009), suggested (WS) and related (WR) Web search queries (Santos et al. 2010a), and the official TREC Web track sub-topics (WT) (Clarke et al. 2009, 2010). The performance of DPH is provided as a non-diversified baseline. The best performance per approach is highlighted in bold. Significance is verified using the Wilcoxon signed-rank test. The symbols ▲ (▼) and △ (∇) denote a significant increase (decrease) at the $p < 0.01$ and $p < 0.05$ levels, respectively, while = denotes no significant difference. A first instance of these symbols denotes the significance of each approach compared to DPH. A second instance, for all variants of xMMR and xMVA, denotes significance with respect to MMR or MVA, respectively.

From Table 3, we first observe that both MMR and MVA show at best marginal yet not significant improvements over the non-diversified ranking produced by DPH, even under training. Indeed, the largest observed improvement is only +3% (MVA's α-nDCG on the WT09 topics). These results corroborate our initial observations in this paper, regarding the lack of empirical validation of novelty-based approaches for diversifying Web search results. Answering our first research question, these results show that novelty is generally an innefective diversification strategy for Web search.

With respect to the different aspect representations, we observe that both xMMR and xMVA can improve over their implicit counterparts in most settings. Under the test scenario, these improvements can be significant, particularly for xMMR using related Web queries (WR) on the WT09 topics (ERR-IA only), and the 'ground-truth' (WT) sub-topics on the WT10 topics, and for xMVA using ODP categories (DZ) on WT09 and WT10 (the latter for α-nDCG only). This completes the investigation of our first research question, by showing that an explicit aspect representation can help improve novelty-based diversification. Nevertheless, only xMMR using the 'ground-truth' aspect representation is able to significantly improve over the non-diversified DPH ranking, which suggests that an explicit representation per se cannot guarantee an effective performance for novelty-based approaches.

5.2 Explicit coverage versus explicit novelty

The observations in Sect. 5.1 suggest an inherent limitation of novelty as a diversification strategy, regardless of any particular aspect representation. To address our second research question, we first contrast the effectiveness of novelty and coverage-based approaches using the same representations. To this end, in Table 4, we compare the diversification

Table 3 Diversification performance (@20) of novelty-based approaches for implicit (MMR and MVA) and explicit (xMMR and xMVA) aspect representations

		WT09		WT10	
		ERR-IA	α-nDCG	ERR-IA	α-nDCG
Training	DPH	0.1430	0.2426	0.1952	0.2977
	+MMR	$0.1430^{=}$	$0.2422^{=}$	$0.1989^{=}$	$0.2921^{=}$
	+xMMR$_{\mathrm{DZ}}$	$0.1431^{==}$	$0.2426^{==}$	$0.2005^{==}$	$0.3034^{==}$
	+xMMR$_{\mathrm{WS}}$	$0.1435^{==}$	$0.2433^{==}$	$0.2071^{\vartriangle=}$	$0.3106^{=\vartriangle}$
	+xMMR$_{\mathrm{WR}}$	$\mathbf{0.1520}^{==}$	$\mathbf{0.2641}^{==}$	$0.2117^{\vartriangle=}$	$0.3235^{\blacktriangle\blacktriangle}$
	+xMMR$_{\mathrm{WT}}$	$0.1486^{==}$	$0.2538^{==}$	$\mathbf{0.2236}^{\vartriangle\vartriangle}$	$\mathbf{0.3369}^{\vartriangle\blacktriangle}$
	+MVA	$0.1444^{=}$	$0.2508^{=}$	$0.2072^{=}$	$0.3070^{=}$
	+xMVA$_{\mathrm{DZ}}$	$0.1441^{==}$	$0.2483^{==}$	$0.1955^{==}$	$0.2979^{==}$
	+xMVA$_{\mathrm{WS}}$	$0.1457^{==}$	$0.2467^{==}$	$0.1804^{==}$	$0.2617^{\triangledown\triangledown}$
	+xMVA$_{\mathrm{WR}}$	$0.1367^{==}$	$0.2329^{==}$	$0.1926^{==}$	$0.2737^{=\triangledown}$
	+xMVA$_{\mathrm{WT}}$	$\mathbf{0.1589}^{==}$	$\mathbf{0.2629}^{==}$	$\mathbf{0.2238}^{==}$	$\mathbf{0.3208}^{==}$
Test	DPH	0.1430	0.2426	0.1952	0.2977
	+MMR	$0.1377^{=}$	$0.2434^{=}$	$0.1951^{=}$	$0.2979^{=}$
	+xMMR$_{\mathrm{DZ}}$	$0.1402^{==}$	$0.2405^{==}$	$0.1954^{==}$	$0.2980^{==}$
	+xMMR$_{\mathrm{WS}}$	$0.1429^{==}$	$0.2460^{==}$	$0.1969^{==}$	$0.2984^{==}$
	+xMMR$_{\mathrm{WR}}$	$\mathbf{0.1477}^{=\vartriangle}$	$\mathbf{0.2551}^{==}$	$0.2009^{==}$	$0.3087^{==}$
	+xMMR$_{\mathrm{WT}}$	$0.1431^{==}$	$0.2508^{==}$	$\mathbf{0.2183}^{\vartriangle\vartriangle}$	$\mathbf{0.3310}^{\vartriangle\vartriangle}$
	+MVA	$0.1301^{=}$	$0.2164^{\triangledown}$	$0.1908^{=}$	$0.2841^{\triangledown}$
	+xMVA$_{\mathrm{DZ}}$	$0.1442^{=\vartriangle}$	$0.2488^{=\blacktriangle}$	$0.1952^{==}$	$0.2968^{=\vartriangle}$
	+xMVA$_{\mathrm{WS}}$	$0.1047^{\triangledown=}$	$0.1743^{\triangledown=}$	$\mathbf{0.1969}^{==}$	$\mathbf{0.2975}^{==}$
	+xMVA$_{\mathrm{WR}}$	$0.1362^{==}$	$0.2318^{==}$	$0.1924^{==}$	$0.2734^{\triangledown=}$
	+xMVA$_{\mathrm{WT}}$	$\mathbf{0.1496}^{==}$	$\mathbf{0.2539}^{==}$	$0.1903^{==}$	$0.2939^{==}$

performance of xMMR and xMVA (novelty-based) to that of IA-Select* and xQuAD* (coverage-based) across the four explicit aspect representations considered in this work. Two instances of the previously introduced significance symbols denote whether IA-Select* and xQuAD* differ significantly from xMMR and xMVA, respectively.

From Table 4, we observe that both coverage-based approaches substantially outperform the novelty-based ones in almost all settings, often significantly. The only exception is IA-Select* using the DZ aspect representation, which slightly underperforms on the WT10 topics, yet not significantly. This might be due to the overall lower performance of the DZ aspect representation compared to the other considered representations. Nevertheless, xQuAD* still outperforms both xMMR and xMVA in this scenario. Considering the other aspect representations, both xMMR and xMVA are significantly outperformed when using the WR representation on both WT09 and WT10 topics, and the WT representation on the WT10 topics. Additionally, on the WT09 topics, significant improvements over xMVA are observed when using the WS representation, and over xMMR when using the WT representation. This answers our second research question, by showing that, whenever the underlying aspect representation is held fixed, coverage provides an often significantly superior diversification strategy compared to novelty.

5.3 Explicit coverage versus explicit coverage + novelty

The results in Sect. 5.2 show that novelty cannot improve against a pure coverage-based strategy. To complete the investigation of our second research question, we investigate

Table 4 Diversification performance (@20) of coverage (IA-Select* and xQuAD*) and novelty-based (xMMR and xMVA) approaches for different explicit aspect representations

		WT09		WT10	
		ERR-IA	α-nDCG	ERR-IA	α-nDCG
Training	DPH	0.1430	0.2426	0.1952	0.2977
	+xMMR$_{\text{DZ}}$	0.1431	0.2426	0.2005	0.3034
	+xMVA$_{\text{DZ}}$	0.1441	0.2483	0.1955	0.2979
	+IA-Select$^*_{\text{DZ}}$	$0.1576^{==}$	$0.2528^{==}$	$0.1805^{==}$	$0.2743^{==}$
	+xQuAD$^*_{\text{DZ}}$	$\mathbf{0.1872}^{\triangle=}$	$\mathbf{0.2879}^{==}$	$\mathbf{0.2326}^{=\triangle}$	$\mathbf{0.3310}^{==}$
	+xMMR$_{\text{WS}}$	0.1435	0.2433	0.2071	0.3106
	+xMVA$_{\text{WS}}$	0.1457	0.2467	0.1804	0.2617
	+IA-Select$^*_{\text{WS}}$	$0.1626^{==}$	$0.2657^{==}$	$0.2407^{==}$	$0.3343^{=\blacktriangle}$
	+xQuAD$^*_{\text{WS}}$	$\mathbf{0.1785}^{==}$	$\mathbf{0.2815}^{==}$	$\mathbf{0.2465}^{=\triangle}$	$\mathbf{0.3417}^{=\blacktriangle}$
	+xMMR$_{\text{WR}}$	0.1520	0.2641	0.2117	0.3235
	+xMVA$_{\text{WR}}$	0.1367	0.2329	0.1926	0.2737
	+IA-Select$^*_{\text{WR}}$	$0.2008^{\triangle\blacktriangle}$	$0.3105^{=\blacktriangle}$	$0.2786^{\blacktriangle\blacktriangle}$	$0.3934^{\blacktriangle\blacktriangle}$
	+xQuAD$^*_{\text{WR}}$	$\mathbf{0.2024}^{\triangle\blacktriangle}$	$\mathbf{0.3126}^{\triangle\blacktriangle}$	$\mathbf{0.2870}^{\blacktriangle\blacktriangle}$	$\mathbf{0.4003}^{\blacktriangle\blacktriangle}$
	+xMMR$_{\text{WT}}$	0.1486	0.2538	0.2236	0.3369
	+xMVA$_{\text{WT}}$	0.1589	0.2629	0.2238	0.3208
	+IA-Select$^*_{\text{WT}}$	$0.1901^{==}$	$0.2968^{==}$	$0.2735^{\triangle=}$	$0.3925^{\triangle\triangle}$
	+xQuAD$^*_{\text{WT}}$	$\mathbf{0.1903}^{==}$	$\mathbf{0.2969}^{==}$	$\mathbf{0.2874}^{\triangle\blacktriangle}$	$\mathbf{0.4004}^{\blacktriangle\blacktriangle}$
Test	DPH	0.1430	0.2426	0.1952	0.2977
	+xMMR$_{\text{DZ}}$	0.1402	0.2405	0.1954	0.2980
	+xMVA$_{\text{DZ}}$	0.1442	0.2488	0.1952	0.2968
	+IA-Select$^*_{\text{DZ}}$	$0.1576^{==}$	$0.2528^{==}$	$0.1805^{==}$	$0.2743^{==}$
	+xQuAD$^*_{\text{DZ}}$	$\mathbf{0.1731}^{==}$	$\mathbf{0.2726}^{==}$	$\mathbf{0.2210}^{==}$	$\mathbf{0.3223}^{==}$
	+xMMR$_{\text{WS}}$	0.1429	0.2460	0.1969	0.2984
	+xMVA$_{\text{WS}}$	0.1047	0.1743	0.1969	0.2975
	+IA-Select$^*_{\text{WS}}$	$0.1626^{=\triangle}$	$0.2657^{=\blacktriangle}$	$\mathbf{0.2407}^{==}$	$0.3343^{==}$
	+xQuAD$^*_{\text{WS}}$	$\mathbf{0.1724}^{=\triangle}$	$\mathbf{0.2758}^{=\blacktriangle}$	$0.2391^{==}$	$\mathbf{0.3363}^{==}$
	+xMMR$_{\text{WR}}$	0.1477	0.2551	0.2009	0.3087
	+xMVA$_{\text{WR}}$	0.1362	0.2318	0.1924	0.2734
	+IA-Select$^*_{\text{WR}}$	$\mathbf{0.2008}^{\triangle\blacktriangle}$	$\mathbf{0.3105}^{\triangle\blacktriangle}$	$0.2786^{\blacktriangle\blacktriangle}$	$0.3934^{\blacktriangle\blacktriangle}$
	+xQuAD$^*_{\text{WR}}$	$0.1923^{=\blacktriangle}$	$0.3039^{=\blacktriangle}$	$\mathbf{0.2835}^{\blacktriangle\blacktriangle}$	$\mathbf{0.3990}^{\blacktriangle\blacktriangle}$
	+xMMR$_{\text{WT}}$	0.1431	0.2508	0.2183	0.3310
	+xMVA$_{\text{WT}}$	0.1496	0.2539	0.1903	0.2939
	+IA-Select$^*_{\text{WT}}$	$\mathbf{0.1901}^{\triangle=}$	$\mathbf{0.2968}^{==}$	$0.2735^{\triangle\blacktriangle}$	$0.3925^{\blacktriangle\blacktriangle}$
	+xQuAD$^*_{\text{WT}}$	$0.1811^{\triangle=}$	$0.2888^{\triangle=}$	$\mathbf{0.2751}^{\triangle\blacktriangle}$	$\mathbf{0.3938}^{\blacktriangle\blacktriangle}$

whether novelty can be effective in combination with coverage. To address this, Table 5 shows the diversification performance of IA-Select and xQuAD, which deploy hybrid diversification strategies, compared to their coverage-only versions, IA-Select* and xQuAD*, respectively. The previously described symbols are used to denote significant improvements between hybrid and coverage-only versions.

From Table 5, we note that neither IA-Select nor xQuAD can consistently improve upon their coverage-only versions. Indeed, no significant improvement is observed across the entire table. Recalling our second question, this surprising result shows that novelty does not significantly contribute to the effectiveness of the state-of-the-art diversification approaches in the literature. Along with the other results in this section, it raises further questions regarding the role of novelty as a diversification strategy, and the conditions (if any) under which this strategy could be effective. We investigate these questions in the next section. A full breakdown analysis of the results in Tables 3, 4, and 5 is provided in Appendix A.

6 Simulated evaluation

The results in Sect. 5 show that novelty performs ineffectively in comparison to and in combination with coverage, and even when compared to a standard, non-diversified ad hoc retrieval baseline. What remains unknown is why this is the case. Hence, in this section, we address our third and last research question, by further investigating the role of novelty as a search result diversification strategy. In particular, our ultimate goal is to identify the conditions (if any) under which novelty could be deployed effectively.

To this end, we perform two complementary simulations. Section 6.1 analyses the impact of simulated relevance and diversity estimates on the effectiveness of novelty-based diversification. Section 6.2 investigates how novelty is affected by non-relevant documents. As the results of both simulations lead to identical conclusions on both WT09 and WT10 settings, for brevity, we only present and discuss the latter.

6.1 Relevance versus diversity

Building upon the view of search result diversification as a trade-off between promoting relevance or diversity (Santos et al. 2010b), we analyse the diversification performance of novelty-based, coverage-based, and hybrid approaches over a range of simulated relevance and diversity estimation performances. The first scenario (*simulated relevance*) simulates the application of these approaches over baseline rankings of various quality. The second scenario (*simulated diversity*) has different interpretations for different approaches. For coverage-based approaches, it represents a refined estimation of how well a document covers different query aspects (e.g., the probability $P_D(d|q, a_m)$ in Eqs. 8 and 9). For explicit novelty-based approaches, it equates to a refined document representation in the space of the considered aspects (see Eq. 2), which allows for an improved identification of novel documents.

Following Turpin and Scholer (2006), we produce a range of relevance estimation performances by simulating re-rankings of the top 1000 results retrieved by DPH for each of the WT10 queries. In particular, each re-ranking seeks a different target *query* average precision (AP), by iteratively swapping randomly chosen pairs of relevant and irrelevant documents. For this simulation, we use the relevance assessments for the ad hoc task of the TREC 2010 Web track (Clarke et al. 2010).[4] A similar procedure is used to simulate diversity estimates. For this simulation, we use the TREC Web track sub-topics as an aspect representation. As described in Sect. 4.3, this is the only available aspect representation with relevance assessments (i.e., those from the diversity task of the TREC 2010 Web track). Based on these ‘ground-truth’ aspects and their corresponding relevance assessments, our simulation iteratively re-ranks the top 1000 results retrieved by DPH for a given query with respect to each sub-topic of this query, until a target *aspect* AP performance is achieved.

[4] The ad hoc and diversity tasks share the same queries.

Table 5 Diversification performance (@20) of coverage-based (IA-Select* and xQuAD*) and hybrid approaches (IA-Select and xQuAD) for different explicit aspect representations

		WT09		WT10	
		ERR-IA	α-nDCG	ERR-IA	α-nDCG
Training	DPH	0.1430	0.2426	0.1952	0.2977
	+IA-Select$^{*}_{\mathrm{DZ}}$	0.1576	0.2528	0.1805	0.2743
	+IA-Select$_{\mathrm{DZ}}$	**0.1593**$^{=}$	**0.2539**$^{=}$	**0.1883**$^{=}$	**0.2866**$^{=}$
	+xQuAD$^{*}_{\mathrm{DZ}}$	0.1872	**0.2879**	**0.2326**	**0.3310**
	+xQuAD$_{\mathrm{DZ}}$	**0.1876**$^{=}$	0.2878$^{=}$	0.2323$^{=}$	0.3276$^{=}$
	+IA-Select$^{*}_{\mathrm{WS}}$	0.1626	0.2657	0.2407	0.3343
	+IA-Select$_{\mathrm{WS}}$	**0.1768**$^{=}$	**0.2802**$^{=}$	**0.2421**$^{=}$	**0.3422**$^{=}$
	+xQuAD$^{*}_{\mathrm{WS}}$	**0.1785**	**0.2815**	**0.2465**	**0.3417**
	+xQuAD$_{\mathrm{WS}}$	**0.1785**$^{=}$	**0.2815**$^{=}$	0.2464$^{=}$	**0.3417**$^{=}$
	+IA-Select$^{*}_{\mathrm{WR}}$	**0.2008**	**0.3105**	0.2786	0.3934
	+IA-Select$_{\mathrm{WR}}$	0.1954$^{=}$	0.3088$^{=}$	**0.2793**$^{=}$	**0.3938**$^{=}$
	+xQuAD$^{*}_{\mathrm{WR}}$	**0.2024**	**0.3126**	0.2870	0.4003
	+xQuAD$_{\mathrm{WR}}$	**0.2024**$^{=}$	**0.3126**$^{=}$	**0.2873**$^{=}$	**0.4014**$^{=}$
	+IA-Select$^{*}_{\mathrm{WT}}$	**0.1901**	**0.2968**	0.2735	**0.3925**
	+IA-Select$_{\mathrm{WT}}$	0.1849$^{=}$	0.2927$^{=}$	**0.2738**$^{=}$	0.3899$^{=}$
	+xQuAD$^{*}_{\mathrm{WT}}$	**0.1903**	**0.2969**	**0.2874**	0.4004
	+xQuAD$_{\mathrm{WT}}$	0.1887$^{=}$	0.2964$^{=}$	**0.2874**$^{=}$	**0.4005**$^{=}$
Test	DPH	0.1430	0.2426	0.1952	0.2977
	+IA-Select$^{*}_{\mathrm{DZ}}$	0.1576	0.2528	0.1805	0.2743
	+IA-Select$_{\mathrm{DZ}}$	**0.1593**$^{=}$	**0.2539**$^{=}$	**0.1883**$^{=}$	**0.2866**$^{=}$
	+xQuAD$^{*}_{\mathrm{DZ}}$	**0.1731**	**0.2726**	**0.2210**	**0.3223**
	+xQuAD$_{\mathrm{DZ}}$	0.1725$^{=}$	0.2713$^{=}$	0.2149$^{=}$	0.3194$^{=}$
	+IA-Select$^{*}_{\mathrm{WS}}$	0.1626	0.2657	0.2407	0.3343
	+IA-Select$_{\mathrm{WS}}$	**0.1768**$^{=}$	**0.2802**$^{=}$	**0.2421**$^{=}$	**0.3422**$^{=}$
	+xQuAD$^{*}_{\mathrm{WS}}$	**0.1724**	**0.2758**	0.2391	0.3363
	+xQuAD$_{\mathrm{WS}}$	0.1709$^{=}$	0.2736$^{=}$	**0.2392**$^{=}$	**0.3371**$^{=}$
	+IA-Select$^{*}_{\mathrm{WR}}$	**0.2008**	**0.3105**	0.2786	0.3934
	+IA-Select$_{\mathrm{WR}}$	0.1954$^{=}$	0.3088$^{=}$	**0.2793**$^{=}$	**0.3938**$^{=}$
	+xQuAD$^{*}_{\mathrm{WR}}$	0.1923	0.3039	**0.2835**	**0.3990**
	+xQuAD$_{\mathrm{WR}}$	**0.2005**$^{=}$	**0.3105**$^{=}$	**0.2835**$^{=}$	**0.3990**$^{=}$
	+IA-Select$^{*}_{\mathrm{WT}}$	**0.1901**	**0.2968**	0.2735	**0.3925**
	+IA-Select$_{\mathrm{WT}}$	0.1849$^{=}$	0.2927$^{=}$	**0.2738**$^{=}$	0.3899$^{=}$
	+xQuAD$^{*}_{\mathrm{WT}}$	**0.1811**	0.2888	0.2751	0.3938
	+xQuAD$_{\mathrm{WT}}$	**0.1811**$^{=}$	**0.2889**$^{=}$	**0.2779**$^{=}$	**0.3949**$^{=}$

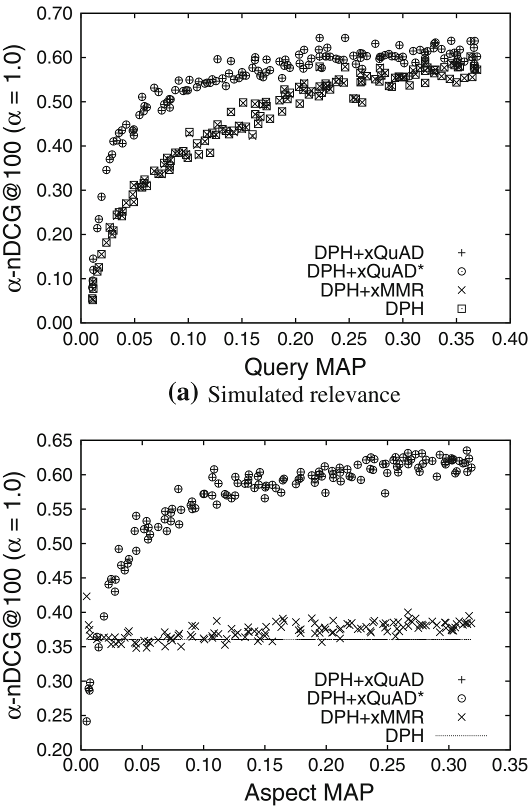
As target relevance (for queries) and diversity (for query aspects) estimation performances, we split the range of possible AP values (i.e., [0,1]) into 20 equally sized bins (i.e., each bin has size 0.05). Within the range of each bin, we randomly select 20 target AP values, making up a total of 400 simulated relevance and diversity estimation performances per query. To enable a comprehensive yet controlled analysis, we focus on xMMR, xQuAD*, and xQuAD as representative explicit novelty-based, coverage-based, and hybrid diversification approaches, respectively. These approaches are particularly suited for this analysis, as they directly implement the aforementioned trade-off between relevance and diversity, hence allowing a controlled experimentation, by varying these two components independently. To avoid any bias towards one of these components, all approaches are applied with the standard setting of $\lambda = 0.5$.

The diversification performance of xMMR, xQuAD*, and xQuAD is shown in Fig. 1a for a range of relevance estimation performances. Relevance performance (the x axis) is measured by mean average precision (MAP). Diversification performance (the y axis) is measured by α-nDCG@100 with $\alpha = 1.0$, so as to penalise redundancy the most heavily. Additionally, since all approaches are applied to diversify the top 100 documents, evaluation at rank cutoff 100 ensures that any observed improvements are due to removing redundancy with respect to the aspects already covered, rather than to covering additional query aspects in the ranking. The diversification performance of a standard DPH ranking is also included as a baseline. From the figure, we first observe that the diversification performance of all approaches is highly correlated to their underlying relevance estimation performance. This is somewhat expected, since by improving relevance, the chance of satisfying at least one of the aspects of the query increases, as confirmed by the high correlation observed for the DPH baseline itself (Pearson's $r = 0.8978$). As for the diversification approaches, xMMR is almost indistinguishable from DPH across the query MAP range. Likewise, xQuAD cannot be distinguished from xQuAD*. This further shows that novelty is a generally weak strategy for promoting diversity, both on its own, and when combined with coverage.

Figure 1b provides a complementary view of the results in Fig. 1a. In this second scenario, instead of varying the relevance estimations for the query, we simulate a range of diversity estimations. Once again, besides the diversification performance of xMMR, xQuAD*, and xQuAD over the range of simulated diversity estimations, we include the performance of DPH as an ad hoc retrieval baseline. From Fig. 1b, we observe that the performance of xMMR remains limited by the performance of the baseline ranking, even with increasingly improved aspect relevance estimations. This result further confirms the limitations of novelty as a diversification strategy. In contrast, the performance of xQuAD* substantially increases as the underlying aspect relevance estimations improve. This shows that, besides being more robust as a diversification strategy, coverage can also benefit more from improved evidence of the association of documents to query aspects. More surprisingly, coverage proves to be a more effective strategy for promoting novelty (i.e., for reducing redundancy) than novelty itself, as shown by the striking superiority of xQuAD* compared to xMMR. On the other hand, the performance of xQuAD cannot be distinguished from that of xQuAD*, further confirming the limitations of novelty when combined with coverage.

6.2 Relevance versus non-relevance

The results in Sect. 6.1 emphasise the limitations of novelty as a diversification strategy, based on a range of simulated relevance and diversity performance scenarios. Focusing on the relevance simulation scenario, for a fixed baseline ranking (i.e., a fixed relevance performance), a novelty-based diversification approach re-ranks documents on the basis of their differences from other documents, with no bearing on their likelihood of being relevant to a query aspect. In particular, Zhai et al. (2003) suggested that the performance gains attained by promoting novelty are offset by the corresponding losses due to also promoting non-relevant documents.



(a) Simulated relevance

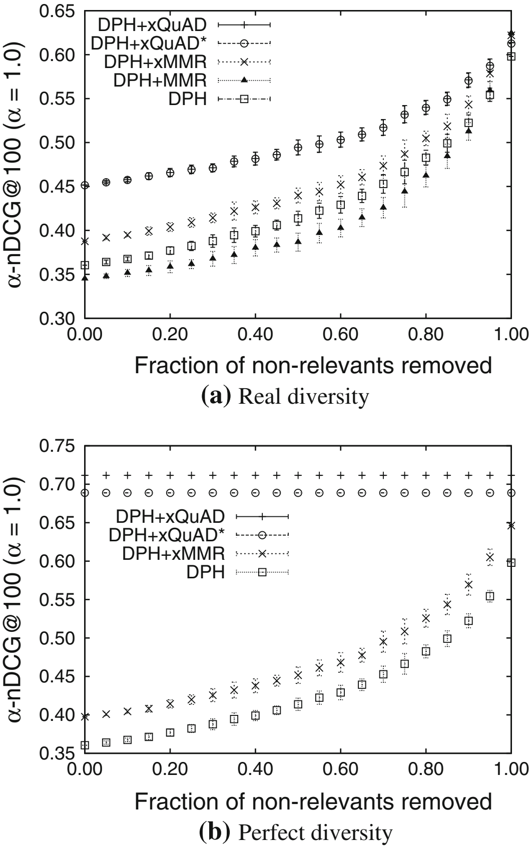
(b) Simulated diversity

Fig. 1 Diversification performance of explicit novelty-based (xMMR), coverage-based (xQuAD*), and hybrid (xQuAD) approaches for a range of **a** relevance and **b** diversity performances

To fully investigate this intuition in a Web search setting, we perform a complementary simulation to the one shown in Fig. 1a. In particular, while the previous simulation produced baseline rankings with various performances, these rankings still contained both relevant and non-relevant documents. Instead, we simulate a different scenario, where the baseline ranking is gradually improved by randomly removing non-relevant documents. This allows us to assess the impact of non-relevant documents on the performance of novelty-based diversification. In particular, Fig. 2a shows the diversification performance of MMR, xMMR, xQuAD*, and xQuAD, as we increase the fraction of non-relevant documents removed from a baseline ranking produced by DPH. MMR (Carbonell and Goldstein 1998) is included so as to allow the analysis of the impact of non-relevant documents under an implicit novelty-based approach. The performance of DPH itself is also shown as a baseline. We test removal fractions from 0 to 1, in steps of 0.05. For instance, a removal fraction of 0 represents the original DPH ranking, while a fraction of 1 means that all non-relevant results have been removed from this ranking. For a given fraction, each random removal of non-relevant documents is repeated 20 times, and we report diversification performances averaged across these 20 repetitions, with error bars denoting standard deviations.

From Fig. 2a, we first note, as expected, that the performance of DPH improves as non-relevant documents are removed from its ranking. What we are interested to know, however, is whether a novelty strategy can take advantage of these gradually improving

Fig. 2 Diversification performance of explicit novelty-based (xMMR), coverage-based (xQuAD*), and hybrid (xQuAD) approaches as non-relevant documents are removed



(a) Real diversity

(b) Perfect diversity

baseline performances. Looking at MMR, we observe that the performance of this implicit novelty-based approach is lower than that of DPH. Moreover, the gap between MMR and DPH remains almost unaltered as non-relevant documents are removed. A similar observation can be made for xMMR. Although it performs above DPH, the gap between the two approaches does not increase with the removal of non-relevant documents. Another important observation is that the hybrid combination of coverage and novelty implemented by xQuAD does not benefit from an improved baseline ranking when compared to xQuAD*—indeed, the performance of these two approaches is indistinguishable from one another in the figure. These results are surprising, as they show that, contrarily to the established intuition, a baseline ranking with only relevant documents is not sufficient to improve novelty-based diversification.

To investigate what could help improve novelty as a diversification strategy, we perform a similar simulation to the one presented in Fig. 2a, however under an extreme scenario. In particular, while the diversification approaches in Fig. 2a leverage 'real' aspect-document relevance estimates (i.e., those provided by DPH), we propose a scenario where these approaches are deployed under ideal conditions, so as to stress their maximum potential. In this idealised scenario, all approaches are deployed with 'perfect' aspect-document relevance estimates, based on the relevance assessments of the diversity task of the TREC 2010 Web track (Clarke et al. 2010). Moreover, all approaches are deployed to make full use of these perfect estimates. To achieve this, xMMR is deployed with $\lambda = 0$ (see Eq. 3), while xQuAD and xQuAD* are deployed with $\lambda = 1.0$ (see Eqs. 7 and 9).[5]

Figure 2b shows the results of this 'perfect' simulation scenario. From the figure, we first observe that xMMR can consistently outperform DPH. However, as in Fig. 2a, the gap between xMMR and DPH remains roughly constant as non-relevant documents are removed. This surprising result shows that removing non-relevant documents from the baseline ranking does not necessarily improve novelty, even when novelty is deployed under idealised conditions.

In terms of absolute performance, although xMMR performs slightly better in contrast to its performance in the 'real' scenario in Fig. 2a, the benefits of deploying novelty as a standalone strategy seem quite low. Indeed, while xMMR struggles to improve over DPH, xQuAD* largely outperforms both DPH and xMMR. To understand why this is the case, we can look at the right end of Fig. 2b. In particular, when there are only relevant documents to be diversified (i.e., when the fraction of non-relevants removed is 1), xQuAD* still outperforms xMMR. This is because, different from coverage, novelty does not take into account how well each *individual* document covers *multiple* query aspects. In contrast, coverage provides a much stronger diversification performance, by placing more emphasis on 'highly diverse' documents (i.e., documents relevant to multiple aspects). Lastly, compared to xQuAD*—a purely coverage-based approach—the hybrid strategy deployed by xQuAD is finally shown to bring significant improvements. This shows that, although rather limited as a standalone strategy, novelty can still play a role in combination with coverage, as a tie-breaking criterion—i.e., whenever two documents have similar coverage, the one that covers the least seen aspects (i.e., the most novel) should be ranked higher.

7 Conclusions

We have thoroughly investigated the role of novelty as a diversification strategy. In particular, we placed existing diversification approaches in a common framework based on two complementary dimensions: diversification strategy and aspect representation. Moreover, we have introduced four new diversification approaches to enable the assessment of novelty as a diversification strategy, independently of the query aspect representation dimension. Based on a thorough investigation, we have provided empirical evidence of the limitations of novelty-based diversification in a standard Web search scenario. Finally, through a comprehensive analysis based on simulations, we have shed light on the limitations of novelty, and its role as a diversification strategy.

In particular, we found that novelty is generally an ineffective diversification strategy when deployed on its own. As it ignores how diverse individual documents are, its performance is inherently limited by the relevance of the underlying baseline ranking. However, when deployed in combination with a coverage-based strategy, it can still provide improvements, provided that an effective aspect-document relevance estimation mechanism is available. To this end, future research should focus on constructing aspect representations that better reflect the multiple possible information needs underlying an

[5] Note that MMR is discarded from this simulation, as it cannot leverage aspect-document relevance estimates.

ambiguous query (Santos and Ounis 2011), e.g., based on the needs of previous users who issued similar queries, as identified from a query log. Another promising direction for investigation is on developing improved retrieval approaches for estimating how different documents cover the identified query aspects, e.g., by leveraging machine learned models (Santos et al. 2011).

Appendix A: Breakdown analysis

Table 6 provides a summary breakdown performance analysis of all approaches investigated in this paper. In particular, performance is measured by ERR-IA@20 across the 98 TREC 2009 and 2010 Web track queries (WT09+WT10), and corresponds to the approaches' test performance in Tables 3, 4, and 5. Each row in Table 6 contrasts a given pair of approaches according to their (mean, maximum, and minimum) difference in performance, as well as the number of queries improved (+), hurt (−), or unchanged (=) by applying the leftmost approach in contrast to the rightmost approach. Likewise, the most improved (best) and most hurt (worst) queries are also shown for each pair of approaches compared. For uniformity, all explicit diversification approaches use the WT aspect representation, as described in Sect. 4.3.

The first group of comparisons in Table 6 contrasts the performance of each diversification approach to the non-diversified DPH baseline. In this group, we first observe that implicit novelty-based approaches (MMR and MVA) differ marginally from DPH on average (−0.0027 for MMR, −0.0087 for MVA). Moreover, the range of differences (max − min) is also small, indicating that these approaches have little impact in the ranking. This observation is corroborated by the fact that most of the queries are unchanged by deploying an implicit novelty-based approach (47 for MMR, 39 for MVA). This appears to be an artifact of the high dimensionality of the term space where these approaches operate, as discussed in Sect. 3.1. Indeed, their explicit counterparts (xMMR and xMVA) show a much stronger impact, dramatically reducing the number of unchanged queries (15 for xMMR, 17 for xMVA) compared to that observed when deploying implicit approaches. While this means that more queries are improved, we can also observe that more queries are hurt, showing that explicit novelty-based approaches are also unstable. In particular, despite affecting more queries than their implicit counterparts, their average performance difference with respect to DPH remains negligible (0.0015 for xMMR, −0.0071 for xMVA). Coverage-based approaches (IA-Select* and xQuAD*) improve this scenario, by consistently improving more queries (48 for IA-Select*, 51 for xQuAD*), while hurting fewer queries (38 for IA-Select*, 35 for xQuAD*). Additionally, their average performance difference with respect to DPH is significantly higher (0.0527 for IA-Select*, 0.0489 for xQuAD*) compared to that observed for novelty-based approaches, both implicit and explicit. Finally, combining coverage and novelty in a hybrid approach (IA-Select and xQuAD) further improves this scenario, yet not significantly, as denoted by the slightly higher number of improved queries (51 for IA-Select, 55 for xQuAD) and smaller number of hurt queries (36 for IA-Select, 32 for xQuAD).

The above observations are further corroborated by the comparisons in Groups 2 through 4 in Table 6. In particular, the second group contrasts the performance of explicit novelty-based approaches (xMMR and xMVA) and their implicit counterparts (MMR and MVA), providing a breakdown of the results in Table 3. From these comparisons, we observe that, while an explicit aspect representation helps more than it hurts for MMR (42

Table 6 Breakdown ERR-IA@20 comparison of all approaches investigated in this paper

	Performance difference			Affected queries		
	Mean	Max	Min	+	=	−
Group 1						
MMR–DPH	−0.0027	0.0896	−0.1826	29	47	22
	Best: *26—lower heart rate*			Worst: *12—djs*		
MVA–DPH	−0.0087	0.0396	−0.2718	23	39	36
	Best: *77—bobcat*			Worst: *12—djs*		
xMMR–DPH	0.0015	0.2088	−0.2137	35	15	48
	best: *50—dog heat*			Worst: *12—djs*		
xMVA–DPH	−0.0071	0.3720	−0.6252	34	17	47
	Best: *55—iron*			Worst: *84—continental plates*		
IA-Select*–DPH	0.0527	0.6627	−0.8789	48	12	38
	Best: *88—forearm pain*			Worst: *86—bart sf*		
xQuAD*–DPH	0.0489	0.6627	−0.9029	51	12	35
	Best: *88—forearm pain*			Worst: *86—bart sf*		
IA-Select–DPH	0.0516	0.5959	−0.5302	51	11	36
	Best: *43—the secret garden*			Worst: *86—bart sf*		
xQuAD–DPH	0.0513	0.6468	−0.8688	55	11	32
	Best: *88—forearm pain*			Worst: *86—bart sf*		
Group 2						
xMMR–MMR	0.0105	0.2526	−0.2135	42	27	29
	Best: *50—dog heat*			Worst: *12—djs*		
xMVA–MVA	0.0016	0.3925	−0.6249	36	19	43
	Best: *55—iron*			Worst: *84—continental plates*		
Group 3						
IA-Select*–xMMR	0.0512	0.5506	−0.8796	56	12	30
	Best: *43—the secret garden*			Worst: *86—bart sf*		
xQuAD*–xMMR	0.0474	0.5506	−0.9036	55	12	31
	Best: *43—the secret garden*			Worst: *86—bart sf*		
IA-Select*–xMVA	0.0598	0.7304	−0.4087	50	11	37
	Best: *88—forearm pain*			Worst: *86—bart sf*		
xQuAD*–xMVA	0.0560	0.7304	−0.4327	52	11	35
	Best: *88—forearm pain*			Worst: *86—bart sf*		
Group 4						
IA-Select–IA-Select*	−0.0012	0.3801	−0.2699	34	35	29
	Best: *80—keyboard reviews*			Worst: *1—obama family tree*		
xQuAD–xQuAD*	0.0024	0.3206	−0.1350	20	68	10
	Best: *80—keyboard reviews*			Worst: *76—raised gardens*		

queries helped, 29 hurt), the same is not true for MVA (36 helped, 43 hurt). Nonetheless, as previously observed in Sect. 5.1, the average difference xMMR–MMR and xMVA–MVA is marginal and not significant. The third group contrasts the performance of explicit coverage-based approaches (IA-Select* and xQuAD*) to explicit novelty-based

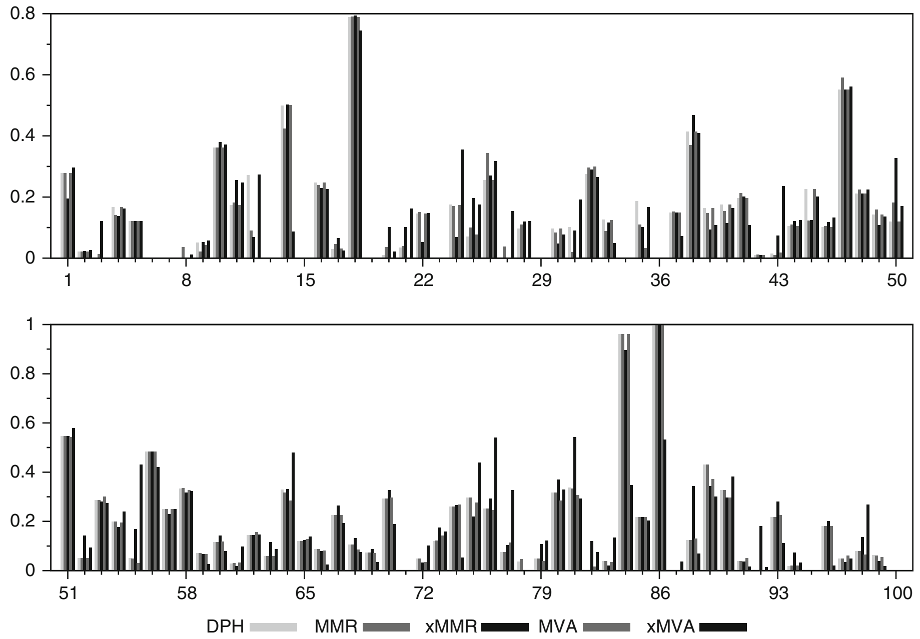


Fig. 3 Breakdown diversification performance of the results in Table 3 (ERR-IA@20, test) across the WT09 (top: 1–50) and WT10 (bottom: 51–100) queries. For each query, five columns denote the performances of DPH, MMR, xMMR, MVA, and xMVA. xMMR and xMVA use the WT aspect representation from Sect. 4.3

approaches (xMMR and xMVA), breaking down the results in Table 4. These comparisons confirm that coverage is a consistently superior strategy than novelty, with IA-Select* and xQuAD* improving upon xMMR and xMVA for the majority of the queries. Finally, the fourth group contrasts the performance of hybrid approaches (IA-Select and xQuAD) to their coverage-only counterpart (IA-Select* and xQuAD*), breaking down the results in Table 5. This last group of comparisons highlights two observations: (1) the novelty component of xQuAD is less sensitive than the one implemented by IA-Select, having no impact whatsoever for 68 queries (against 35 unchanged queries for IA-Select); and (2) when in operation, the novelty component of xQuAD is also safer, improving 20 queries while hurting only 10 (against 34 improved and 29 hurt queries for IA-Select), although marginally.

Overall, while Table 6 lists the most improved and most hurt query for each pair of approaches compared, there is no apparent correlation between these approaches and the characteristics of these queries (e.g., whether these queries are ambiguous or underspecified (Clarke et al. 2008), or their number of relevant aspects). Therefore, it is difficult to draw any conclusions regarding the particular queries that would benefit from deploying novelty as a diversification strategy (either by itself or in combination with coverage). Instead, for completeness, in Figs. 3, 4, and 5, we provide a full query-by-query breakdown of the test ERR-IA@20 results in Tables 3, 4, and 5, respectively.

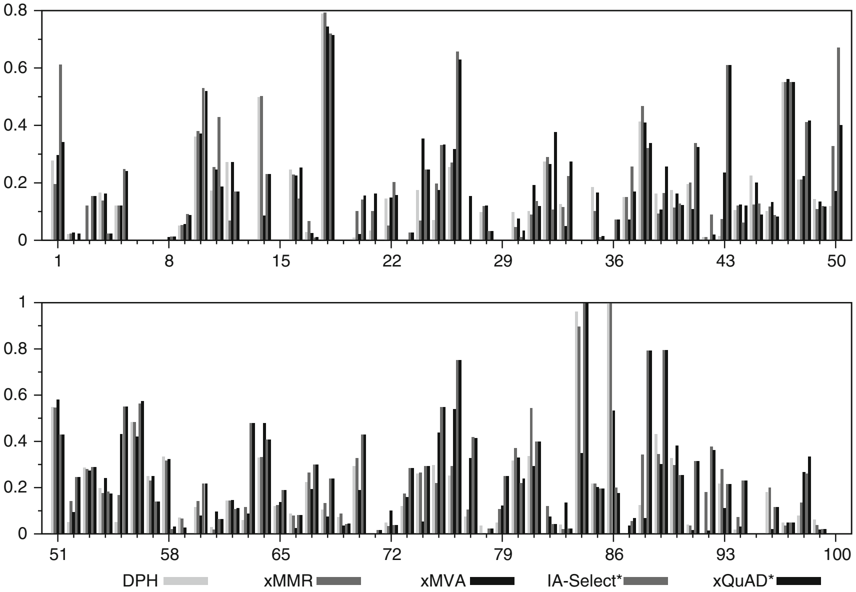


Fig. 4 Breakdown diversification performance of the results in Table 4 (ERR-IA@20, test) across the WT09 (top: 1–50) and WT10 (bottom: 51–100) queries. For each query, five columns denote the performances of DPH, xMMR, xMVA, IA-Select*, and xQuAD*. xMMR, xMVA, IA-Select*, and xQuAD* use the WT aspect representation from Sect. 4.3

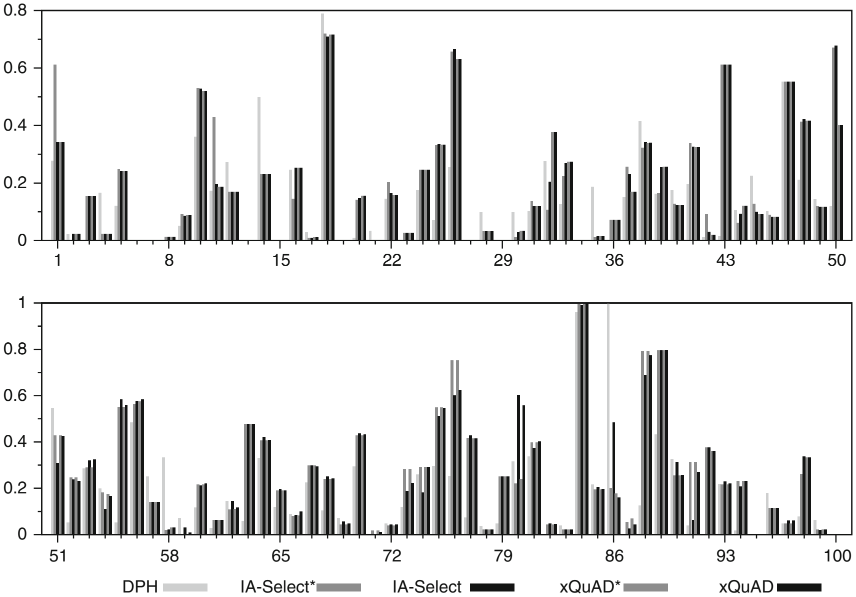


Fig. 5 Breakdown diversification performance of the results in Table 5 (ERR-IA@20, test) across the WT09 (top: 1–50) and WT10 (bottom: 51–100) queries. For each query, five columns denote the performances of DPH, IA-Select*, IA-Select, xQuAD*, and xQuAD. IA-Select*, IA-Select, xQuAD*, and xQuAD use the WT aspect representation from Sect. 4.3

References

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the 2nd ACM international conference on web search and data mining* (pp. 5–14).

Amati, G., Ambrosi, E., Bianchi, M., Gaibisso, C., & Gambosi, G. (2007). FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog track. In *Proceedings of the 16th text REtrieval conference*.

Carbonell, J., & Goldstein, J. (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 335–336).

Carterette, B., & Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 1287–1296).

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In*Proceedings of the 18th ACM conference on information and knowledge management* (pp. 621–630).

Chen, H., & Karger, D. R. (2006). Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 429–436).

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., et al. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 659–666).

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 web track. In *Proceedings of the 18th text retrieval conference*.

Clarke, C. L. A., Craswell, N., Soboroff, I., & Cormack, G. V. (2010). Preliminary overview of the TREC 2010 Web track. In *Proceedings of the 19th text retrieval conference*.

Clarke, C. L. A., Craswell, N., Soboroff, I., & Ashkan, A. (2011). A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the 4th ACM international conference on web search and data mining* (pp. 75–84).

Cooper, W. S. (1971). *The inadequacy of probability of usefulness as a ranking criterion for retrieval system output*. Technical report, University of California, Berkeley.

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008) An experimental comparison of click position-bias models. In *Proceedings of the 1st ACM international conference on web search and data mining* (pp. 87–94).

Goffman, W. (1964). On relevance as a measure. *Information Storage and Retrieval, 2*(3), 201–203.

Gordon, M. D., & Lenk, P. (1991). A utility theoretic examination of the probability ranking principle in information retrieval. *Journal of the American Society for Information Science and Technology, 42*(10), 703–714.

Hochbaum, D. S. (Ed.). (1997). Approximation algorithms for NP-hard problems. Boston, MA, USA: PWS Publishing Co.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Radlinski, F., & Dumais, S. (2006). Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 691–692).

Rafiei, D., Bharat, K., & Shukla, A. (2010) Diversifying Web search results. In *Proceedings of the 19th international conference on world wide web* (pp. 781–790).

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation, 33*(4), 294–304.

Sakai, T., & Song, R. (2011). Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 1043–1052).

Sanderson, M. (2008). Ambiguous queries: Test collections need more sense. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 499–506).

Santos, R. L. T., & Ounis, I. (2011). Diversifying for multiple information needs. In *Proceedings of the 1st International Workshop on Diversity in Document Retrieval* (pp. 37–41).

Santos, R. L. T., Macdonald, C., & Ounis, I. (2010a). Exploiting query reformulations for Web search result diversification. In *Proceedings of the 19th international conference on world wide web* (pp. 881–890).

Santos, R. L. T., Macdonald, C., & Ounis, I. (2010b). Selectively diversifying Web search results. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 1179–1188).

Santos, R. L. T., Peng, J., Macdonald, C., & Ounis, I. (2010c). Explicit search result diversification through sub-queries. In *Proceedings of the 31st European conference on information retrieval* (pp. 87–99).

Santos, R. L. T., Macdonald, C., & Ounis, I. (2011). Intent-aware search result diversification. In *Proceedings of the 34th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 595–604).

Song, R., Luo, Z., Nie, J. Y., Yu, Y., Hon, H. W. (2009). Identification of ambiguous queries in web search. *Information Processing and Management, 45*(2), 216–229.

Spärck-Jones, K., Robertson, S. E., & Sanderson, M. (2007). Ambiguous requests: implications for retrieval tests, systems and theories. *SIGIR Forum, 41*(2), 8–17.

Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 11–18).

Wang, J., & Zhu, J. (2009). Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 115–122).

Zhai, C., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 10–17).