Received: 10 October 2019 | Revised: 20 July 2020 | Accepted: 22 July 2020

DOI: 10.1002/asi.24403

RESEARCH ARTICLE

JASIST WILEY

Explicit diversification of search results across multiple dimensions for educational search

Sevgi Yigit-Sert[1] | **Ismail Sengor Altingovde**[1] | **Craig Macdonald**[2] | **Iadh Ounis**[2] | **Özgür Ulusoy**[3]

[1]Computer Engineering Department, Middle East Technical University, Ankara, Turkey

[2]School of Computing Science, University of Glasgow, Glasgow, UK

[3]Computer Engineering Department, Bilkent University, Ankara, Turkey

Correspondence
Ismail Sengor Altingovde, Computer Engineering Department, Middle East Technical University, 06800, Ankara, Turkey.
Email: altingovde@ceng.metu.edu.tr

Funding information
Royal Society, Grant/Award Number: NI140231; The Scientific and Technological Research Council of Turkey (TÜBİTAK), Grant/Award Number: 117E861; Türkiye Bilimler Akademisi, Grant/Award Number: Distinguished Young Scientist Award 2016; Türkiye Bilimsel ve Teknolojik Araştirma Kurumu, Grant/Award Number: BIDEB 2211/A

Abstract

Making use of search systems to foster learning is an emerging research trend known as *search as learning*. Earlier works identified result diversification as a useful technique to support learning-oriented search, since diversification ensures a comprehensive coverage of various aspects of the queried topic in the result list. Inspired by this finding, first we define a new research problem, multidimensional result diversification, in the context of educational search. We argue that in a search engine for the education domain, it is necessary to diversify results across multiple dimensions, that is, not only for the topical aspects covered by the retrieved documents, but also for other dimensions, such as the type of the document (e.g., text, video, etc.) or its intellectual level (say, for beginners/experts). Second, we propose a framework that extends the probabilistic and supervised diversification methods to take into account the coverage of such multiple dimensions. We demonstrate its effectiveness upon a newly developed test collection based on a real-life educational search engine. Thorough experiments based on gathered relevance annotations reveal that the proposed framework outperforms the baseline by up to 2.4%. An alternative evaluation utilizing user clicks also yields improvements of up to 2% w.r.t. various metrics.

1 | INTRODUCTION

Exploiting search as a process for learning is an emerging and exciting research direction (a.k.a. *search as learning*), which has already attracted interest from various fields, such as computer, psychology and learning sciences (Collins-Thompson, Hansen, & Hauff, 2017; Hoppe et al., 2018). A particular direction in recent studies, especially from the perspective of information retrieval research, addressed how general-purpose search engines can be exploited and enriched to satisfy the users' possible learning goals. To this end, earlier works attempted to re-rank the results of a search engine by applying various techniques, most notably, personalization or diversification. For instance, some works (e.g., Collins-Thompson, Bennett, White, de la Chica, & Sontag, 2011) personalized the displayed ranking by incorporating the reading difficulty of documents. In contrast, Raman, Bennett, and Collins-Thompson (2014) proposed a method to diversify the retrieved document result set in terms of the different topical aspects for the so-called exploratory queries.

While the aforementioned previous works paved the way for improving educational search, they essentially focused on leveraging a single-dimension, for example, either the topical aspect or the reading difficulty, during the re-ranking of query results. Instead, in this article, we argue that a search activity for learning can indeed benefit from diversifying the result list—as also suggested in

an earlier study (Syed & Collins-Thompson, 2017)—yet we also argue that the diversification of search results should in contrast be provided for *multiple dimensions*. That is, the result list should not only be diversified for the topical aspects covered by the retrieved documents, but also for other dimensions, such as the type of the document (e.g., text, video, animation, or even a test to assess what has been learnt) or its intellectual level (say, for beginners or experts). Our stand-point is to address a need that has also been recognized by others. For instance, Hoppe et al. (2018) identified the "lack of consideration for multimodal resources" as a major challenge in the search as learning paradigm. Hence, we propose diversification applied to multiple dimensions to obtain a re-ranking of results that can complement learning via search in multiple ways (e.g., presentation of information in alternative *forms* and *levels*), beyond the sole coverage of the topical variety.

Let us assume an illustrative example query, "triangle," which may have several underlying intents (i.e., topical aspects) such as: "types of triangle," "triangle inequalities," "triangle trigonometry." Furthermore, for the type dimension, each document may be related to several possible aspects (such as lecture, exercise, video, etc.), and may target one or more K-12 levels (as aspects) in the educational level dimension. Thus, to compose a result set for this "triangle" query, which would satisfy many users' learning needs, we need to diversify the result set with respect to *each* possible dimension (e.g., topicality, type of documents, education level, etc.).

We envision that, for general-purpose search engines, it may not be necessarily optimal to consider diversification over all the aforementioned dimensions for every query[1] since (a) it will interfere with many other signals for ranking and might cause a quality reduction for the noneducational queries, and (b) the knowledge of all the dimensions and their semantics may not be readily available. Therefore, different from most existing works, rather than a general-purpose search system, our work focuses on an educational search engine, where both the collection (i.e., educational materials) and the users of the system have a richer set of features that could be naturally exploited for search towards learning. Consequently, we examine the multi-dimensional diversification of search results in this context. We employ the data from a real-life educational search engine embedded into a commercial web-based educational framework for K-12 level students in Turkey with around 1.2 M registered users. Note that while the education level dimension typically covers the K-12 aspects range from 1 to 12, our used query log sample covers only the range from 4 to 8. Hence, our results and discussions in the remainder of the manuscript refer only to education levels of 4 to 8. The contributions of this work are four-fold:

1. We define a new result diversification problem that addresses the typical requirements of a search as learning scenario, that is, where there are a wide range of dimensions the search engine needs to consider when returning results that meet the learning goals (i.e., providing comprehensive information on the topic in many forms (e.g., various types of documents) and at many education levels (e.g., from level 4 to 8)).
2. We provide a new framework for diversification, which extends the state-of-the-art diversification methods (namely, xQuAD [Santos, Macdonald, & Ounis, 2010]; a variant of PM2 [Aktolga, 2014] as well as a supervised approach, R-LTR [Zhu, Lan, Guo, Cheng, & Niu, 2014]), to handle multiple dimensions, and provide tailored instantiations for the framework. Specifically, we enrich each diversification method so that while an aspect's coverage in the final ranking is computed, the importance of the dimension which this aspect belongs to is also taken into account. To illustrate our motivation for computing dimension importance values, let an example query be "triangle," and assume that the candidate set has documents from all the types available in the system but they all pertain to the education level 4. In this case, the diversification algorithm should focus on diversifying documents w.r.t. The type dimension, since there are several aspects to cover there, but should not attempt to diversify for the education level dimension. Hence, while setting the importance values for certain dimensions adaptively (i.e., per query), we consider the variety of the aspect values observed in the candidate set.
3. We describe a new rich dataset[2] tailored for the evaluation of diversification algorithms with multiple dimensions, built from user interactions with an existing real-life educational search engine.
4. We carry out an extensive evaluation of our work using a realistic experimental set-up, which is based on query instances and clicks in addition to TREC-style relevance annotations. Our experiments demonstrate the effectiveness of our proposed approach in comparison to strong baselines, showing improvements of 2.6%, 1.4%, and 2.2% for the diversification metrics ERR-IA, α-nDCG and P-IA, respectively; and an improvement of 1.4% for the traditional P@2 metric. Considering the positive impact of diversified result presentation on the learning outcomes (e.g., knowledge gains of users) as shown in Collins-Thompson, Rieh, Haynes, and Syed (2016) and Syed and Collins-Thompson (2017), these improvements in

diversification performance are likely to translate into learning gains in the educational search context, which is the ultimate goal of our present investigation.

The rest of the article is organized as follows. First, we review the related literature. Next, the Diversification across Dimensions section describes our models for diversification across multiple dimensions as adaptations of the xQuAD, PM2 and R-LTR methods, and provides particular instantiations of these models in the context of an educational search engine. The following two sections, Dataset and Experimental Evaluations, present the search evaluation dataset developed for this work and the evaluation results, respectively. The last section provides concluding remarks.

2 | RELATED WORK

We first review related work in the *search as learning* field. Then, we position our work in the search result diversification literature.

2.1 | Search as learning

Enhanced ranking in general-purpose and educational search engines for learning goals. A particular existing direction to enhance the learning experience via search involves the re-ranking of results from general-purpose search engines. Collins-Thompson et al. (2011) proposed to personalize Web search results by re-ranking them with respect to the reading difficulty. More recently, Yilmaz, Ozcan, Altingovde, and Ulusoy (2019) proposed an approach in which they trained classifiers using various educational resources to predict the course category of question-like queries, and then employed these predictions as a signal for re-ranking the initial query results. Both works customized the final result list w.r.t. a single feature of a given user (i.e., reading level) or query (i.e., course category) in a general-purpose search engine. Instead, our work in this manuscript leverages diversification in multiple dimensions as the key methodology to obtain a re-ranking of the results for learning purposes.

Two particular studies employed diversification in a learning-related context. Raman et al. (2014) addressed exploratory Web search queries and the so-called *intrinsically diverse* sessions, where users aim to learn about a topic by seeking information about its multiple aspects. To address such queries, they introduced a greedy diversification algorithm that re-ranks the initially retrieved results. Syed and Collins-Thompson (2017) applied the latter diversification algorithm to enhance the educational benefits in the vocabulary learning task. Contrary to these studies, we employ a more specific educational search setup that enables applying diversification across multiple dimensions, and not only for a single dimension (i.e., topical aspects).

The aforementioned works aimed to improve general-purpose search engines to support search for learning. However, an alternative and complementary research direction is to focus on specialized educational/learning settings that also involve search (referred to as educational search engines here). For instance, Hoppe et al. (2018) mentioned the TIB's web portal,[3] which is dedicated to scientific videos search. Usta, Altingövde, Vidinli, Ozcan, and Ulusoy (2014) presented an analysis of an educational search engine that works on a proprietary education platform for K-12 students. In this work, we also focus on an educational search engine setup, as it is a natural testbed for our proposed diversification approach across multiple dimensions.

Evaluating the impact of search on learning. One particular strategy to evaluate the impact of a search session on gains in the users' knowledge about a given topic is to conduct pre- and post-assessments via tests, summaries, or user studies (e.g., Collins-Thompson et al., 2016; Maxwell, Azzopardi, & Moshfeghi, 2019; Moraes, Putra, & Hauff, 2018; Vanopstal, Stichele, Laureys, & Buysschaert, 2012). In our work, since we exploit real search logs from an educational search engine while not having the possibility to interact with the actual users, we rely on traditional metrics computed over the re-ranked results using the proposed multidimensional diversification framework. A similar evaluation approach has been adopted in the aforementioned works of (Collins-Thompson et al., 2011; Raman et al., 2014). Furthermore, Collins-Thompson et al. (2016) have already shown that an intrinsically diverse presentation of search results yields the highest percentage of users with knowledge gains; and hence, our improvements in terms of the traditional and click-based diversification metrics have a high likelihood of improving users' learning in an educational search context.

2.2 | Diversification of search results

In the literature, diversification approaches are essentially applied to ambiguous queries (such as the query "jaguar," which could be seeking information for either the aspect "animal" or "car") where the user's search intent cannot be clearly determined.

Diversification methods are characterized as either *implicit* or *explicit*, which differ in how the diversification

is conducted. In particular, implicit approaches (e.g., Carpineto, Mizzaro, Romano, & Snidero, 2009; He, Meij, & de Rijke, 2011; Liang, Ren, & de Rijke, 2014; Xia et al., 2017; Zhu et al., 2014) only inspect the attributes of each document itself, usually their contents. In contrast, explicit approaches (such as xQuAD [Santos et al., 2010], PM2 [Dang & Croft, 2012] and aggregation-based methods [Ozdemiray & Altingovde, 2015], DSSA [Jiang et al., 2018]) use an external representation (e.g., common query reformulations) to infer the (topical) *aspects* of queries. In this work, we extend such explicit approaches (namely, xQuAD and PM2) (as well as a more recent supervised implicit approach [R-LTR]) to take into account the multiple *dimensions* that naturally arise in an educational search setting.

There are three approaches in the literature, which are closest to ours in terms of their diversification methodology. First, Hu, Dou, Wang, Sakai, and Wen (2015) introduced the notion of hierarchical intents of topicality. Our work goes further by considering multiple orthogonal dimensions of diversification rather than a strict hierarchy, and goes beyond topicality, to encompass other dimensions that can be estimated (e.g., readability) or derived from document metadata attributes (e.g., document type). Second, Aktolga (2014)[Ch.5] investigated adaptations to PM2 that could achieve a mixed diversification of both topical and nontopical (implicit) dimensions, namely, the sentiments and dates expressed in the documents. Finally, Dou, Hu, Chen, Song, and Wen (2011) proposed a multidimensional topic richness model in a similar fashion to xQuAD for web search diversification. They considered each dimension as a *data source* (such as anchor texts, query logs, web sites, etc.) from which different aspects can be mined. Hu, Dou, Wang, and Wen (2015) extended the latter approach with the aspects derived from an additional data source, namely, the lists appearing in the candidate documents. In contrast to the latter approaches, our experiments focus on dimensions of diversification that are appropriate to an educational search engine. Furthermore, we also extend R-LTR, an implicit diversification method, to exploit explicit aspects for multiple dimensions. To the best of our knowledge, R-LTR has been used with explicit aspects only in (Y. Wang, Luo, & Yu, 2016), but again, not for handling dimensions in the context of educational search. Last but not the least, none of these approaches employ a click-based evaluation setup as we do in this manuscript.

Finally, diversification has been recently studied in recommendation systems. For instance, Noia, Rosati, Tomeo, and Sciascio (2017) applied diversification by taking multiple attributes (i.e., genre, year, actor etc.) of items into account. Instead, our work aims to improve *search experience* in an educational setup.

3 | DIVERSIFICATION ACROSS DIMENSIONS

We now describe the xQuAD (Santos et al., 2010), PM2 (Dang & Croft, 2012) and R-LTR (Zhu et al., 2014) diversification approaches, and show how to adapt them to consider multiple dimensions. Our work builds on xQuAD because: (a) it has been found as the best-performing diversification approach in all TREC campaigns between 2009–2012, and (b) it has only one parameter (namely, λ), which requires tuning. We also choose PM2 for similar reasons, as it has been shown to be as competitive as xQuAD and again has a single parameter, λ. Finally, we employ R-LTR as a representative for the supervised diversification methods.

3.1 | xQuAD

xQuAD iteratively selects documents from an initial ranking of candidate documents for query Q, denoted by $R(Q)$, into the final ranking S that maximizes the following objective:

$$(1-\lambda)\Pr(d|Q)+\lambda\sum_{a\in Q}\left[\Pr(a|q)\Pr(d|a)\prod_{d_j\in S}\left(1-\Pr\left(d_j|a\right)\right)\right], \quad (1)$$

where Q is the user's query, a is an aspect of Q, and S is the set of already selected documents. $\Pr(d|\,Q)$ and $\Pr(d|\,a)$ are identically defined, as being the normalized score of a document with respect to the original query, or an aspect, and can be calculated using any effective document ranking approach, such as BM25 (Santos et al., 2010) or more advanced learned ranking models. The probability $\Pr(a|\,q)$ represents the importance of that aspect for the query, and, by default, is uniform across all aspects (Santos et al., 2010).[4]

We note that the novelty $\prod()$ component of xQuAD may yield small values as more documents are selected into S and the corresponding $(1 - \Pr(d_j|\,a))$ values are multiplied (Ozdemiray & Altingovde, 2015). As a remedy, the product can be replaced by the arithmetic and geometric mean of the probabilities (Ozdemiray & Altingovde, 2015). We refer to these variants as art_xQuAD and geo_xQuAD hereafter.

xQuAD diversifies across any intent space Q, but, typically, common query reformulations are used to identify topics the user may be looking for. However, xQuAD omits other independent factors (i.e., dimensions) affecting the suitability of a document to users.

3.2 | Multidimensional xQuAD

We assume that there are multiple dimensions of diversification $dim \in D$, possibly conditioned on the query Q (denoted by $D(Q)$), which should be covered in a ranking. Each dimension dim has a corresponding set of aspects: a_1, ..., a_n. For the topic dimension, which is generally applied in web search, the aspects are the underlying intents often inferred by mining the query reformulations or knowledge-bases. Although our model is more general, in this article we consider two further dimensions, namely the (educational) level that the document targets and the type of document, which are specific to our target application of search in the education domain. The aspects for such dimensions may also be identified in similar ways to the topic dimension, for example, for a given query, we use related suggestions and their retrieved (and even clicked) results to detect the relevant educational levels or the document types.

Our proposed model is simple in that it adapts xQuAD by marginalizing over all dimensions:

TABLE 1 Dimensions and aspects used in this work

Dimension	Aspects	Pr(d| a, dim) value
Topicality	Via log mining	Estimated
Education level	{4,5,6,7,8}	Known
Type	{animation, interactive exercise, video, text, game, lecture, conversational exercise, application, summary}	Known

$$(1-\lambda)\Pr(d|Q) + \lambda \sum_{dim \in D(Q)} \sum_{a \in dim} \left[\Pr(dim|Q) \cdot \Pr(a|dim, Q) \cdot \Pr(d|a, dim) \cdot \prod_{d_j \in S} (1 - \Pr(d_j|a, dim)) \right]. \quad (2)$$

In Equation (2): Pr(dim| Q) defines the *dimension importance*, which represents the importance of a dimension for the query; Pr(a| dim, Q) defines the *aspect importance*; and Pr(d| a, dim) is the *document aspect coverage*. Note that we differentiate between dimensions for which the probability Pr(d| a, dim) is *estimated* (such as the relevance of a document to a topical aspect of a query) and dimensions for which this probability can be accurately known based on the available metadata for documents (e.g., given a query related to the *animation* aspect for the type dimension, the diversification algorithm assigns Pr(d| $a_{animation}$, dim_{type}) to either 0 or 1 based on the metadata associated with the document d). Table 1 highlights the dimensions and aspects that we consider in this work. Note that, as mentioned in the introduction, while the aspects for the education level dimension typically cover the range of 1 to 12, for K-12, our query log sample covers only the range of 4 to 8.

To instantiate the proposed multidimensional xQuAD approach, we discuss how to instantiate the dimension and aspect importance probabilities. In particular, the importance of diversification upon each dimension may vary between queries—for example, observing documents with a variety of education levels in the candidate set of documents $R(Q)$ for a particular query Q may suggest that portraying these different levels of content (c.f. Table 1) in the top-ranked documents is likely to benefit a wide range of users. Thus, for the (education) level dimension, we set the dimension importance as follows:

$$\Pr(dim_{level}|Q) = \frac{O(Q) - min_{level}}{max_{level} - min_{level}}, \quad (3)$$

where $O(Q)$ denotes the level aspects observed in $R(Q)$, and max_{level} (min_{level}) denotes the maximum (minimum) number of possible aspects in the level dimension, respectively. For instance, if the documents in $R(Q)$ cover the level aspects {5, 6, 7} and all possible level aspects are {4, 5, 6, 7, 8}, we set Pr(dim_{level}| Q) = (3 − 1)/(5 − 1) = 0.5. Note that, if all possible aspects are observed in $R(Q)$, the importance score is 1, while if only one aspect is observed, it is 0; i.e., no need to diversify for this dimension. The importance of the type dimension is set in the same manner. However, for the topic dimension, we cannot know how many aspects are observed in $R(Q)$, as we can only *estimate* topical relevance. Hence, we intuitively set Pr($dim_{topicality}$| Q) = 1, as we expect relevance to be the first driver of diversification, with diversification

encapsulating other dimensions having relatively lesser importance.

3.3 | PM2

PM2 (Dang & Croft, 2012) adapts the allocation problem of seats to party representatives in some election systems to finding a diversified result list. The diversified result set is constructed with respect to the set of aspects related to the query in proportion to the popularity of these aspects. PM2 starts with a ranked list, $R(Q)$, that represents the candidate documents, with k empty seats, which is the size of the diversified list, S. In each iteration, the winner aspect is determined by the popularity of the aspect (referred as the quotient score). The quotient score is computed for each aspect i via:

$$quotient[i] = \frac{v_i}{(2s_i + 1)} \tag{4}$$

where v_i and s_i indicate the number of votes the party i receives and the number of seats that have been assigned to the party i. A seat (the position in S) is allocated for the winner aspect, that is, i^*, and the document d^* that is relevant to the winner aspect is selected by the following score function:

$$d^* \leftarrow \underset{d_j \in R(q)}{\operatorname{argmax}} \lambda \times qt[i^*] \times \Pr(d_j|q_{i^*}) + (1-\lambda) \sum_{i \neq i^*} qt[i] \times \Pr(d|q_i) \tag{5}$$

where $qt[i]$ is the quotient score and λ is the trade-off parameter between the relevance to the winner aspect and other aspects. Since the selected document is relevant to the other aspects in addition to the winner aspect, PM2 updates the portion of seats in the selected set, S.

3.4 | Multidimensional PM2

We adapt the original PM2 for diversification with multiple dimensions following a similar approach to that of Aktolga (2014)[Ch.5]. Unlike the original PM2 formulation, we have one s_i, which indicates the portion of selected documents in S for aspect i, and v_i, which denotes the number of documents the aspect i should have, for each aspect a of each dimension $dim \in D$. The quotient score is calculated for each aspect a under each dimension dim. Multidimensional PM2 selects the winner aspect i^* for each dimension, and then computes the relevance of the next document in S to the winner aspect versus its relevance to all the other query aspects within that dimension.

Note that our approach is similar to the adaptation of PM2 to multiple dimensions proposed by Aktolga (2014) [Ch.5], in the choice of a winning aspect i^* for each dimension. However, it differs in terms of computing the dimension importance and the λ parameter. We use Equation (3) for the dimension importance instead of using the interpolated weights and we set λ without any smoothing. Furthermore, this previous work employed dimensions (e.g., the document sentiment) that are not applicable to our context.

The scoring equation of multidimensional PM2 is as follows:

$$\begin{aligned} d^* \leftarrow \underset{d_j \in R(q)}{\operatorname{argmax}} & \sum_{dim \in D(Q)} \Pr(dim|Q) \times \lambda \times qt[i^*, dim] \\ & \times \Pr(d_j|q_{i^*,dim}) + (1-\lambda) \sum_{i \neq i^*} qt[i, dim] \times \Pr(d|q_i, dim) \end{aligned} \tag{6}$$

We use the same setting to instantiate the dimension and aspect probabilities for multidimensional PM2 as in multidimensional xQuAD.

3.5 | R-LTR

R-LTR (Zhu et al., 2014) is a supervised implicit diversification method that learns the weights of its scoring function using Stochastic Gradient Descent (SGD). Given a candidate document set $R(Q)$ for a query Q, R-LTR constructs the final ranking S in a greedy manner. In each iteration, R-LTR computes the following scoring function for each document d_i that is not in the ranking S, and the one with the highest score is added to S:

$$\text{R-LTR}_{\text{imp}}(d_i, V_i, S) = \omega_r * x_i + \omega_d * h_S(V_i) \tag{7}$$

The first part of Equation (7) represents the relevance of the scored document, and the second part represents its diversity from the documents already selected in the ranking S. x_i denotes a relevance feature vector that comprises scores expressing query-document matching (e.g., tf-idf, BM25, etc.), while V_i is a matrix capturing the diversity scores of d_i to all other documents in $R(Q)$, in terms of various diversity functions. Hence, V is a 3-way tensor that stores the diversity between each pair of documents in $R(Q)$, each computed using various diversity functions. Finally, ω_r and ω_d are the weight vectors (for the relevance and diversity components, respectively) that are learnt during the training stage.

Since the original R-LTR is an implicit method, it does not employ the knowledge of aspects. Thus, in our setting, for a given pair of documents, the diversity scores are computed as follows. First, based on the content of the documents, we compute two different diversity scores using typical similarity measures from the literature: (a) the tf-idf weighted Cosine similarity, and (b) the Jaccard Coefficient of the document vectors. Second, since the candidate documents' education level and type information are also available (as metadata) during the diversification, we compute their distance using Binary Similarity Coefficient and Jaccard Coefficient, respectively. Thus, for each pair of documents, the tensor V stores a vector of 4 different diversity scores. Note that, in Equation (7), while computing the diversity of d_i to the documents already selected in S, the aggregation function $h_S()$ is invoked, which is the minimum function in our setting (as in Zhu et al. [2014]). We denote this baseline by R-LTR$_{\text{imp}}$.

3.6 | Multidimensional R-LTR

We propose a variant of R-LTR that uses the explicit aspects associated with multiple dimensions, as described in the previous sections. In Equation (8), V^{topic}, V^{level} and V^{type} store the pairwise diversity scores (utilizing the associated aspects) across topic, education level and type dimensions, respectively.

$$\begin{aligned}\text{R-LTR}_{\text{exp}}\left(d_i, V^{topic}, V^{level}, V^{type}, S, Q\right) = \omega_r * x_i + \Pr\left(dim_{topic}|Q\right) * \omega_{topic} * h_S\left(V_i^{topic}\right) + \\ \Pr(dim_{level}|Q) * \omega_{level} * h_S\left(V_i^{level}\right) + \Pr\left(dim_{type}|Q\right) * \omega_{type} * h_S\left(V_i^{type}\right)\end{aligned} \tag{8}$$

In this case, for each dimension, documents are represented w.r.t. their relationship with the related aspects. Specifically, for the topicality dimension, we represent each document as a vector of $\Pr(d|\ a_i)$ scores, which is the score of the document d with respect to each aspect a_i (calculated using an effective ranking approach such as BM25). Then, the tensor V^{topic} stores the Euclidean distance between these document vectors, as a particular type of diversity score. Furthermore, we calculate the pairwise difference of the $\Pr(d|\ a)$ scores between two document vectors and obtain their maximum and minimum as additional diversity scores. For the level and type dimensions, as before, we assign binary values for each aspect according to the metadata associated with the document d. We compute the Euclidean distance between the document vectors to be used as diversity scores in the V^{level} and V^{type} tensors. Finally, for each dimension, the aggregated score is multiplied with $\Pr(dim_i|\ Q)$, which is instantiated as before. We call this method R-LTR$_{\text{exp}}$.

Finally, in Goynuk and Altingovde (2020), R-LTR was implemented using a neural network framework, which allows a nonlinear formulation and the training of more complex models (i.e., via multiple hidden layers). Similarly, we apply this approach for training a model based on the same input as Equation (8) (denoted by R-LTR$_{\text{expNN}}$).

4 | DATASET

Since there is no suitable TREC collection or existing public benchmark, we describe a new benchmark dataset that we have constructed in the context of an educational search engine for the multidimensional diversification problem. Expanding upon the topic development practices of the TREC Web track diversity tasks (Clarke, Craswell, & Soboroff, 2009), the dataset is created as follows.

4.1 | Identifying the main queries

As our starting point, we use a query log from an educational search engine embedded into a commercial web-based educational framework (called Vitamin[5]) for K-12 level students[6] in Turkey with around 1.2 M registered users. The Vitamin platform hosts a rich set of educational materials (documents, videos, etc.) for various K-12 courses, as well as a search engine to access them. The query log contains a sample of 20 K queries (6,503 of which are unique) from April 2015. To identify the main queries that would benefit from diversification, we follow Dou, Song, Yuan, and Wen (2008) and use click entropy, which indicates the variation of the clicked documents for each query. The selected queries have a total click count of 20 or more, and an entropy greater than 1.5. We also manually eliminated the near-duplicate queries that are extremely short or that are textual variations of each

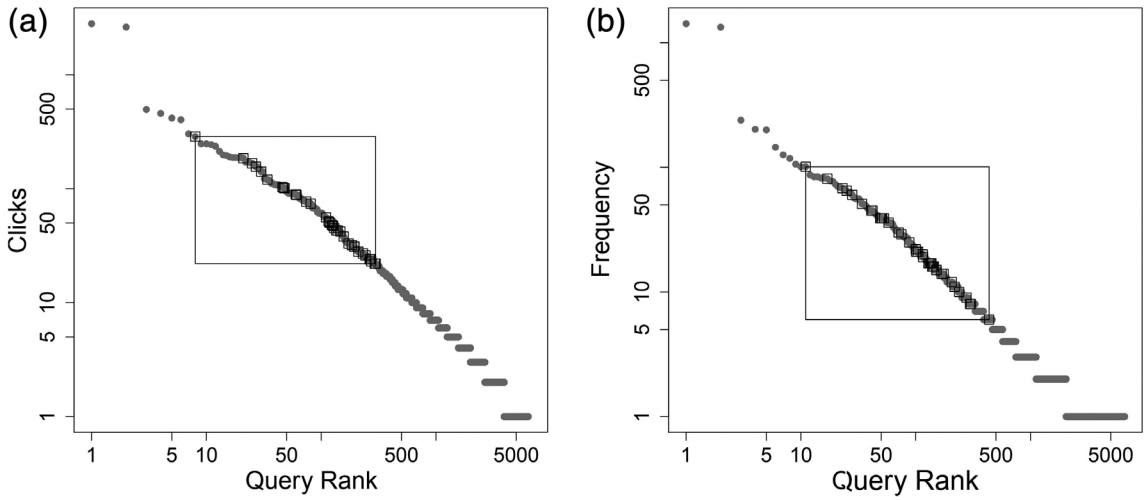

FIGURE 1 Distribution of query click count (left) and frequency (right); our main queries are sampled from the marked regions in each plot

other (i.e., "triangle" vs. "triangles"), keeping the variant with a higher entropy. For the remaining queries, we obtained their related queries (i.e., a related query to q is a query q' either following q in a search session, as in [Clarke et al., 2009], or entirely including the query string q). We then discarded the queries with no or trivial related queries or with completely nonrelevant ones (i.e., no relevance to the original query). This procedure yielded us 40 queries, such as "light," "angles," "electricity" that have a variety of aspects.

With respect to the total click count and occurrence frequency, these 40 queries come from the "torso" of the power law distribution of the query log (as shown in Figure 1), and hence, they are representative of the query volume (as in [Clarke et al., 2009]). Head queries, such as "mathematics," "game," or "science," are too generic and it is unreasonable to determine a set of possible aspects underlying those queries. For tail queries (e.g., "converting poetry to prose"), there is nothing to diversify since they are very specific in nature. Note that the majority of our main queries exhibit similarity to the intrinsic queries of (Raman et al., 2014), that is, they seek information for the various aspects of the same main topic (e.g., for the query "triangle"; the aspects are "triangle inequalities" and "triangle trigonometry"), while just a few of them exhibit extrinsic diversity and involve ambiguity (e.g., the query "voice" [in Turkish] has aspects related to both physics and language).

4.2 | Identifying ground-truth aspects for the topic dimension

First, we clustered the reformulations of a query (using a hierarchical clustering algorithm, as in [Clarke et al., 2009]) to determine the candidate aspects. Next, we applied a manual post-processing process for the noisy clusters, that is, we merged the clusters that are clearly related to the same underlying aspect, and removed those clusters that are redundant, that is, including queries related to aspects already represented by other clusters. Finally, five human annotators (Computer Scientists with teaching experience) labeled these clusters as aspects. The annotators are native Turkish speakers and we verified that they are familiar with the subjects of the assigned queries. During the annotation of the aspects, the annotators took into account the retrieved documents as listed in the query log, as well as the domain knowledge obtained from other educational resources. Thus, even if there has been no cluster representing "trigonometry" for the "triangle" query, the annotator could add it as an aspect. Note that these aspects serve as the "official" aspects for the topic dimension (in the next section, we discuss the aspects for the type and level dimensions).

4.3 | Document-level annotation

For each main query in our set, we obtained all of its occurrences in the query log, and constructed a union of the results (namely, the top-25 documents) for each occurrence. We used the same five judges to annotate the (binary) relevance of each document to the main query and its topical aspects. In general, the documents for each query were annotated by one judge, yielding a total of 12,735 annotations. For a random subset of 4,842 annotations, we also employed a second judge. The obtained Cohen's κ coefficient of inter-rater agreement on these 4,842 annotations is 0.77, which indicates a substantial

agreement (Cohen, 1968). The observed level of agreement suggests that the relevance annotation task is fairly easy for the used query set, and hence, our choice of assigning a single judge per query is adequate.

Finally, for the type and level dimensions, we obtained the official aspects of a query by accessing the metadata of the topically relevant documents in the ground-truth for that query. On average, this yielded 3.55 and 4.53 aspects per query for the type and level dimensions, respectively.

5 | EXPERIMENTAL EVALUATION

We consider two different frameworks for the evaluation: (a) The Annotation-based Evaluation is based on the relevance judgments, that is, annotations obtained via TREC-style topic development procedure, as described in the previous section; (b) The Click-based Evaluation is based on the clicked results for each query instance, separately, and allows setting the aspect importance probabilities more realistically (i.e., by learning from a training set).

5.1 | Annotation-based evaluation

Setup. *Candidate set*. For each query, we re-rank its result documents (obtained from the query log, as described above) using BM25 and identify the top-25 documents as the candidate set to be diversified.

Diversification methods and parameters. As baselines, we use xQuAD and the two variants with the novelty components employing the arithmetic (art_xQuAD) and geometric mean (geo_xQuAD) of the probabilities. We evaluate our multidimensional approach with all three variants. In some experiments, PM2 and R-LTR are also employed.

We have three different dimensions to consider in diversification: education level, type, and topic as specified in Table 1. For the topic dimension, we compute the relevance of the candidate documents (actually, their titles and short descriptions) to the main query and its aspects in the topic dimension, that is, $\Pr(d|\ Q)$ and $\Pr(d|\ a)$, based on the BM25 scores as in earlier works. As the topical aspects, we experiment with the "official" ones (as an ideal scenario). For the (education) level and (document) type dimensions, we assume that the official aspects appropriate for each query are not available at the time of diversification, which may be the case in practice. Hence, we obtain the education level and type of the documents that are in the candidate set of the query, and diversify only based on this knowledge.

For each dimension, we assign the aspect probability $\Pr(a|\ dim, Q)$ assuming a uniform distribution across the aspects[7] in this dimension, as is typical in the literature (e.g., Santos et al., 2010). For all methods, we report the results for the best-performing value of the trade-off parameter λ, which is 1. Note that earlier works (such as Ozdemiray & Altingovde, 2015) also report a similar value of λ and attribute this to the use of the official query aspects.

Finally, R-LTR, being a supervised method, requires training. For all R-LTR variants, we set the learning rate (for stochastic gradient descent) to 0.005 based on the training data. To implement $\text{R-LTR}_{\text{expNN}}$, we train a fully connected two-layer neural network with back-propagation (using the PyTorch framework). The hidden layer has 10 nodes with a sigmoid activation function, and the number of epochs is set to 50. The ground truth ranking is obtained by greedily selecting the document that maximizes the α-nDCG metric as in (Zhu et al., 2014). We apply a five-fold cross-validation to evaluate the performance.

Evaluation metrics. The ground truth includes the relevance labels of documents for the union of the aspects (of all dimensions), given a query. Based on this ground truth, we compute typical and well-known diversification metrics in the literature (ERR-IA, α-nDCG, P-IA, Subtopic(ST)-Recall, and D#-nDCG), all at rank cut-off 10. For computing the metric scores, we employ two approaches. The *Flat* evaluation is the traditional setup that does not take the dimensions into account, while the *DimAware* evaluation computes a metric score for each of the three dimensions and then obtains their average as the overall performance (i.e., as the layer-aware metrics in X. Wang, Dou, Sakai, and Wen [2016]).

We use the Student's paired t test (at 95% confidence level) for analyzing statistical significance.

Results. Our experiments answer the following research questions:

- Does using three dimensions altogether yield a better diversification performance than using each of these dimensions on its own?
- Do the multidimensional xQuAD, PM2 and R-LTR variants yield better diversification than their so-called *flat* counterparts, that is, the original algorithms that use all the aspects belonging to all dimensions as a flat set of aspects?

To answer the first question, Table 2 compares the diversification effectiveness of three cases: (a) the non-diversified BM25 baseline, (b) the original xQuAD algorithm that uses the aspects of each dimension, namely, topic, education level and type, separately; and (c) the original xQuAD algorithm using the union of aspects

TABLE 2 Diversification performances of Single Dimension and Flat xQuAD (using the Flat and DimAware evaluation)

		Flat evaluation					DimAware evaluation				
Div	Method	ERR-IA	α-nDCG	P-IA	ST-Recall	D#-nDCG	ERR-IA	α-nDCG	P-IA	ST-Recall	D#-nDCG
None	BM25	0.425	0.766	0.304	0.796	0.817	0.447	0.766	0.321	0.811	0.848
Single Dim	xQuAD topic	**0.468**	0.831	**0.309**	0.857	0.864	0.482	0.813	**0.325**	0.858	0.883
	xQuAD level	0.437	0.792	0.299	0.852	0.839	0.457	0.791	0.316	0.867	0.869
	xQuAD type	0.433	0.783	0.300	0.827	0.828	0.455	0.786	0.317	0.841	0.858
Flat	xQuAD	**0.468***,‡	**0.845***,‡	0.299†	**0.923**†,*,‡	**0.880***,‡	**0.483***,‡	**0.833***,‡	0.315†	**0.929**†,*,‡	**0.901***,‡

Note: The best values for each case are shown in bold.
†Statistically significant difference from xQuAD topic at 0.05 level.
*Statistically significant difference from xQuAD level at 0.05 level.
‡Statistically significant difference from xQuAD type at 0.05 level.

TABLE 3 Diversification performances of the flat and multidimensional methods (with the Uniform and Adaptive Instantiations of the dimensions' importance) using the Flat evaluation

		Flat evaluation				
Div.	Method	ERR-IA	α-nDCG	P-IA	ST-Recall	D#-nDCG
None	BM25	0.425	0.766	0.304	0.796	0.817
Flat	xQuAD	0.468	0.845	0.299	0.923	0.880
	art_xQuAD	**0.477**	**0.865**	0.313	0.910	0.894
	geo_xQuAD	0.472	0.849	0.300	**0.925**	0.883
	PM2	0.475	0.862	**0.315**	0.914	**0.896**
M-Dim	xQuAD	0.467 (−0.3%)	0.843 (−0.2%)	0.299	**0.923**	0.880 (−0.1%)
(Uniform)	art_xQuAD	0.476 (−0.1%)	**0.865**	0.313	0.916	0.893 (−0.1%)
	geo_xQuAD	0.469 (−0.6%)	0.846 (−0.4%)	0.299 (−0.3%)	**0.923** (−0.2%)	0.880 (−0.3%)
	PM2	**0.479** (0.9%)	0.861 (−0.1%)	**0.316** (0.4%)	0.912 (−0.2%)	**0.897** (0.1%)
M-Dim	xQuAD	0.481 (2.7%)	0.859 (1.7%)	0.302 (1%)	0.931 (0.8%)	0.890 (1.1%)
(Adaptive)	art_xQuAD	**0.489** (2.6%)	**0.877** (1.4%)	**0.320*** (2.2%)	0.913 (−0.4%)	0.903* (1%)
	geo_xQuAD	0.482 (2.1%)	0.860 (1.2%)	0.301 (0.5%)	0.929 (0.4%)	0.889 (0.6%)
	PM2	0.477 (0.5%)	0.865 (0.4%)	0.317 (0.7%)	**0.936*** (2.4%)	**0.908*** (0.5%)

Notes: In parentheses, we report the percentage change w.r.t. the corresponding flat method. The best values for each case are shown in bold.
*Statistically significant difference using the Student's paired *t* test (at 95% confidence level) w.r.t. the corresponding flat method.

from all three dimensions as a *flat* input. Our findings reveal that diversification using aspects from even one dimension is superior to a nondiversified baseline for the majority of metrics, while among the three dimensions, diversification via the topic dimension yields the best performance for all metrics. Furthermore, using aspects from all three dimensions (as a flat diversification) yields considerably better results than using a single dimension for most of the metrics. In other words, diversification considering just one dimension (say, topic) is not likely to yield results that are also sufficiently diverse for the other dimensions. Although there may be some correlations between the aspects of different dimensions (e.g., topic and education level), the algorithms should better use all the dimensions explicitly for the best performance, as we aim to do in this article.

Table 3 addresses our second research question, that is, can the multidimensional algorithms that explicitly model the query dimensions along with their aspects outperform their flat versions? To begin with, Table 3 (using the Flat Evaluation) diversification methods (based on xQuAD and PM2) usually provide a notable improvement over the nondiversified BM25 baseline for all metrics. For instance, while BM25 yields an α-nDCG score of 0.766, the best performing flat and multidimensional method, namely, art_xQuAD, yields 0.865 and 0.877, respectively.

Next, we compare the performance of the multidimensional diversification approaches under two

TABLE 4 Diversification performances of the flat and multidimensional methods (with the Uniform and Adaptive Instantiations of the dimensions' importance) using the DimAware evaluation

		DimAware evaluation				
Div.	**Method**	**ERR-IA**	**α-nDCG**	**P-IA**	**ST-Recall**	**D#-nDCG**
None	BM25	0.447	0.766	0.321	0.811	0.848
Flat	xQuAD	0.483	0.833	0.315	0.929	0.901
	art_xQuAD	**0.493**	**0**.854	0.329	0.927	0.915
	geo_xQuAD	0.487	0.838	0.316	**0.931**	0.904
	PM2	**0.493**	**0.855**	**0.331**	0.925	**0.919**
M-Dim	xQuAD	0.483	0.832 (−0.1%)	0.315	**0.929**	0.900 (−0.1%)
(Uniform)	art_xQuAD	0.494 (0.3%)	**0.856** (0.2%)	0.329 (0.1%)	0.928 (0.2%)	0.916
	geo_xQuAD	0.485 (−0.4%)	0.834 (−0.4%)	0.315 (−0.3%)	**0.929** (−0.2%)	0.901 (−0.3%)
	PM2	**0.497** (0.9%)	0.853 (−0.2%)	**0.333** (0.5%)	0.923 (−0.1%)	**0.921** (0.1%)
M-Dim	xQuAD	0.497 (2.9%)	0.848 (1.8%)	0.318 (1%)	0.936 (0.7%)	0.911* (1%)
(Adaptive)	art_xQuAD	**0.507** (2.8%)	**0.866** (1.4%)	**0.337*** (2.2%)	0.923 (−0.4%)	0.925* (1.1%)
	geo_xQuAD	0.498 (2.1%)	0.848 (1.3%)	0.317 (0.4%)	0.934 (0.3%)	0.909 (1.3%)
	PM2	0.495 (0.4%)	0.856 (0.1%)	0.333 (0.6%)	**0.943*** (2%)	**0.930*** (1.1%)

Notes: In parentheses, we report the percentage change w.r.t. the corresponding flat method. The best values for each case are shown in bold.
*Statistically significant difference using the Student's paired *t* test (at 95% confidence level) w.r.t. the corresponding flat method.

TABLE 5 Diversification performances of R-LTR using the Flat evaluation

		Flat evaluation				
Div.	**Method**	**ERR-IA**	**α-nDCG**	**P-IA**	**ST-Recall**	**D#-nDCG**
Imp.	R-LTR$_{imp}$	0.430	0.785	0.282	0.883	0.836
M-Dim Exp.	R-LTR$_{exp}$	0.435 (1.2%)	0.806 (2.7%)	0.291 (3.2%)	0.933* (5.7%)	0.863* (3.2%)
M-Dim Exp.	R-LTR$_{expNN}$	**0.461*** (7.2%)	**0.849*** (8.2%)	**0.305*** (8.2%)	**0.958*** (8.5%)	**0.897*** (7.3%)
M-Dim Exp.	art_xQuAD	0.489	0.877	0.320	0.913	0.903

Notes: In parentheses, we report the percentage change w.r.t. R-LTR$_{imp}$. The best values for each case are shown in bold.
*Statistically significant difference using the Student's paired *t* test (at 95% confidence level) w.r.t. R-LTR$_{imp}$.

TABLE 6 Diversification performances of R-LTR using the DimAware evaluation

		DimAware evaluation				
Div.	**Method**	**ERR-IA**	**α-nDCG**	**P-IA**	**ST-Recall**	**D#-nDCG**
Imp.	R-LTR$_{imp}$	0.450	0.783	0.299	0.891	0.864
M-Dim Exp.	R-LTR$_{exp}$	0.454 (0.9%)	0.798 (1.9%)	0.307 (2.7%)	0.935* (4.9%)	0.882* (2.1%)
M-Dim Exp.	R-LTR$_{expNN}$	**0.481*** (6.9%)	**0.842*** (7.5%)	**0.321*** (7.4%)	**0.960*** (7.7%)	**0.918*** (6.3%)
M-Dim Exp.	art_xQuAD	0.507	0.866	0.337	0.923	0.925

Notes: In parentheses, we report the percentage change w.r.t. R-LTR$_{imp}$. The best values for each case are shown in bold.
*Statistically significant difference using the Student's paired *t* test (at 95% confidence level) w.r.t. R-LTR$_{imp}$.

instantiations: setting the importance of each dimension, Pr($dim|Q$), as proposed in the section entitled Multi-dimensional xQuAD (referred to as Adaptive) versus under a uniform distribution assumption (i.e., to 1/3 in this case). We find that the multidimensional approaches with the Uniform instantiation does not yield any better

performance than their flat versions (except in a few cases). In contrast, the multidimensional approaches with the dimensions' importance set using the Adaptive method achieve the best performance and consistently outperform their flat counterparts on all metrics. For most of the metrics, the best performing multidimensional diversification method is art_xQuAD, which yields the scores of 0.489, 0.877, and 0.320 for ERR-IA, α-nDCG and P-IA, while its flat counterpart can only achieve 0.477, 0.865, and 0.313, suggesting a relative improvement of 2.6%, 1.4%, and 2.2%, respectively. Note that similar trends are also observed for the DimAware Evaluation, reported in Table 4.

Tables 5 and 6 provide the findings for the approaches based on the supervised R-LTR method (to facilitate comparisons, the results for the art_xQuAD is repeated). Our results reveal that (a) our multidimensional R-LTR$_{exp}$ approach (using Adaptive instantiation) with explicit aspects outperforms the baseline R-LTR$_{imp}$ (which is an implicit diversification method), (b) our implementation of the multidimensional R-LTR approach using a two-layer neural network (as in Goynuk and Altingovde [2020]) further improves the performance (since R-LTR$_{expNN}$ outperforms R-LTR$_{exp}$), and (c) multidimensional R-LTR$_{expNN}$ yields the best performance only for the ST-Recall metric, while multidimensional art_xQuAD performs better for the remaining metrics.

Overall, our experiments confirm a positive answer to our second research question: multidimensional approaches with our instantiations are superior to the original diversification algorithms, i.e., the flat versions of xQuAD and PM2, and the baseline R-LTR$_{imp}$.

5.2 | Click-based evaluation

In this section, we provide an alternative evaluation based on user clicks. We focus on the second research question of the previous section, that is, whether multidimensional diversification approaches can outperform their flat counterparts in educational search, which lies at the very core of this article.

Setup. *Query instances and candidate results*. We use the same query set as before. However, instead of constructing a candidate ranking per query and then diversifying it (as in the section entitled Annotation-based Evaluation), here for each query instance (i.e., an occurrence in the query log), we obtain the result list that has been actually presented to the user, again from the query log, and then diversify the latter, which serves as the candidate ranking in this setup. Note that for a given query, say, "triangle," the result lists generated by the underlying retrieval system for different instances are usually quite similar, but there might be occasional variations due to updates in the document collection and other system-dependent factors. However, the clicked results in each instance may vary widely, as different users may differ in their learning interests for one or more aspects of a given query. The latter type of information, clicks observed for each instance (together with our relevance judgments) are exploited for evaluation in this section. Our goal is to re-rank the candidate result list (via diversification) of a given query instance so that the clicked results appear as early as possible in the list (more details are provided later).

Different from the previous section where we had a candidate result set of 25 documents, here we restrict our candidate set to the top-10 documents per query instance, since our evaluation is based on the users' clicks and due to the well-known rank (or position) bias, it is less likely to observe clicks for the documents ranked too low, that is, after a cutoff value of 10. Overall, for our 40 main queries, we extracted 926 instances together with their top-10 results, which form our dataset for the experiments in this section.

Diversification methods and parameters. In this section, we employ only the flat and multidimensional versions of xQuAD since the results from the previous section show them to be representative. As before, we compute the relevance of the candidate documents to the main query and its aspects in the topic dimension, (i.e., $\Pr(d|q)$ and $\Pr(d|a)$), using BM25. For the (educational) level and (document) type dimensions, we use binary values, as defined in the Multidimensional xQuAD section. We estimate the dimension importance probability through Equation (3) for the level and type dimensions while we set the importance of topic dimension to 1. For all diversification experiments, we report the results for the best-performing value of the trade-off parameter λ, which is found to be 0.9.

A crucial issue is determining the aspect importance, $\Pr(a|dim, Q)$, for each dimension and its aspects, which was previously assigned a uniform distribution. Since our evaluation in this section is based on user clicks, the diversification algorithm should accurately model the preferences of the user population towards different aspects of a query, as they may markedly vary. For instance, Figure 2 displays the user clicks' distribution over the aspects of each dimension for the query "light." For the education level dimension, the aspect *level 7* is the most popular aspect with a considerably large click-rate, that is, 85.3% of clicks observed over all instances of this query are for the documents covering this aspect. The documents with education levels 4 and 5 are very rarely clicked, while the other levels (6 and 8) are not

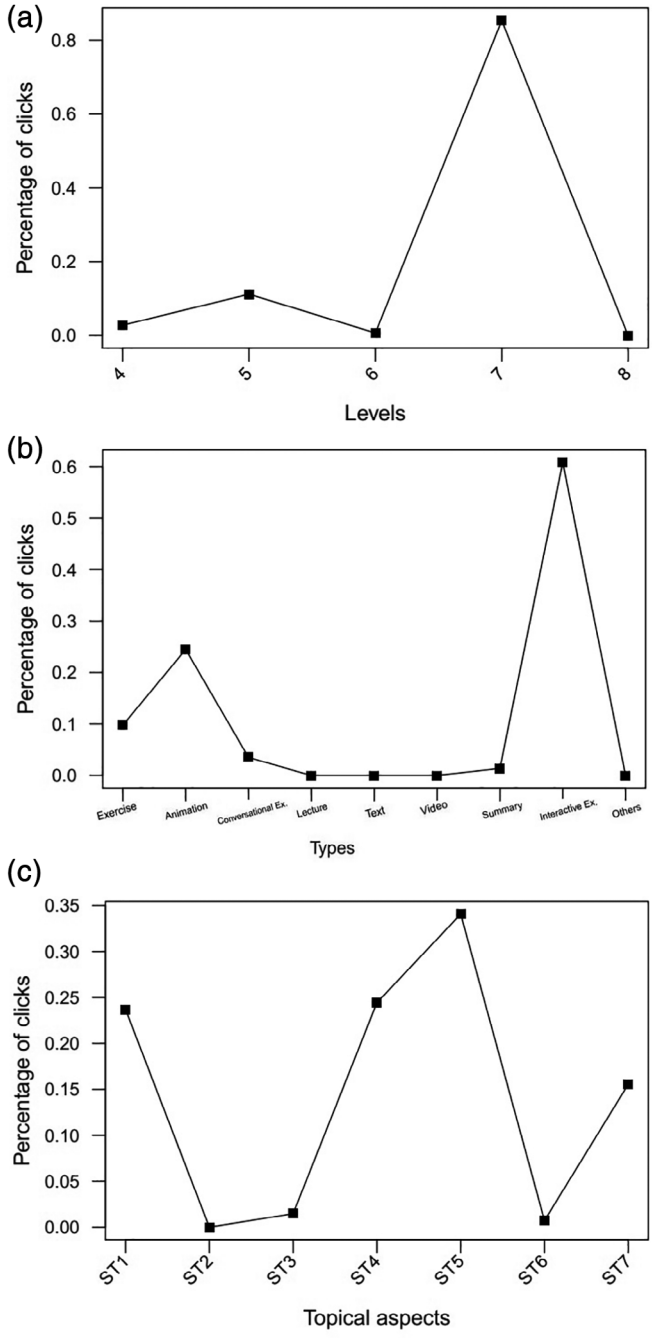

FIGURE 2 Distribution of click counts for the query "light" across each dimension: education level (top-left), type (top-right), topic (bottom). For the latter, the topical aspects shown as ST1 to ST7 on the *x*-axis correspond to "light and color," "light filter," "white light," "absorption," "refraction," "light year," and "light sources," respectively

clicked at all. Similarly, for the topic dimension, documents covering four of the aspects that are identified in the ground truth (namely, *ST1*, *ST4*, *ST5*, *ST7* in Figure 2) are often clicked, while the others are neglected.

To illustrate why it is crucial to accurately model the aspect probabilities during diversification, consider the following toy scenario. Assume a query q with three different aspects A, B, and C (say, in the topic dimension) and three candidate documents d1, d2, and d3 covering these aspects, respectively. If all aspects are equally likely in the query log, then the top-2 rankings (obtained after diversification) as (d1, d2) or (d2, d3) would be equally good, as each ranking covers two different aspects. However, if we further assume that the users' click rates on the documents covering these aspects A, B, and C are 90%, 5%, and 5%, respectively, then it is obvious that a click-based evaluation will favor the ranking (d1, d2) over the ranking (d2, d3), since in the majority of its instances, documents covering aspect A will be clicked, yielding a higher evaluation score. This suggests that the top-ranked results should not only cover the diverse aspects, but those diverse aspects that are popular, so that we can improve the click-based metrics.

We learn the aspect probabilities for each query and dimension by splitting our dataset into training and test sets. In particular, for each query, we use the first 75% of its instances (in timestamp order) as the training set (adding up to 699 instances) and the rest as the test set (including 227 instances in total). The aspect priors are then obtained from the training set using Equation (9):

$$P(a|dim,Q) = \frac{\text{relevant clicks for aspect } a}{\sum_{a\in(dim,Q)} \text{relevant clicks}} \tag{9}$$

As mentioned before, most users click the top-ranked result(s) regardless of its relevance, a phenomena known as the rank bias. For instance, in our dataset, for about 25% of the instances, only the top-1 or top-2 results are clicked. Naturally, for such rankings, it is almost impossible to improve the click-based metrics via re-ranking (i.e., after the diversification). This is a common issue that arises in the case of conducting a click-based evaluation by re-ranking previously obtained results, usually from a query log. The ideal solution—of conducting an A/B test with the previous and treated rankings—is rarely attainable as the researchers are usually not in control of the underlying retrieval system, which also holds for our case. Hence, following the practice in some previous works (e.g., Bai, Cambazoglu, Gullo, Mantrach, & Silvestri, 2017), we combine the initial ranking with the diversified one (using the well-known Borda Voting method (e.g., see Aslam & Montague, 2001) so that the final ranking is not extremely different from the initial ranking. Furthermore, we always preserve the first document in the initial ranking and apply diversification for the rest of the documents in the list.

Ground truth and evaluation. In this setup, the ground truth is based on the clicked results per query

instance, following Bai et al. (2017). Furthermore, we filter them so that only those clicks on the documents that are labeled as relevant in the annotation-based evaluation (see the section entitled Document-level Annotation) are kept in each instance's ground truth. Among the total of 1,895 clicks for all our query instances, only 12% of the clicks are for documents labeled as nonrelevant. In other words, 88% of the users' clicks were for documents judged "relevant." For the remaining 12% of clicks, a manual analysis of some randomly chosen results revealed that these clicks are noisy (i.e., the user—most likely to be a young student, as our dataset covers the educational levels between 4 and 8—may have clicked unintentionally) as they seem definitely nonrelevant, and hence we discard them.

However, note that not every relevant result may be clicked in every instance; as discussed before, certain users may be interested in certain aspects only, and thus may skip documents that are labeled as relevant yet covering the other aspects that are not interesting for such users. Therefore, the evaluation framework presented in this section differs from that of the previous section.

To summarize, for each instance, the ground truth involves those results that are both clicked by the user in this instance's result list and also labeled as relevant by our judges. Based on this ground truth, we compute the traditional relevance metrics as well as the diversification metrics (this is possible, since the ground truth aspects are available for the documents labeled as relevant). We report the Precision and nDCG metrics for relevance, and P-IA and α-nDCG for diversity, at early rank cutoff values of 2 and 5.

Results. Table 7 presents the diversification performance of the multidimensional xQuAD algorithm with different aspect importance weights. The "Uniform" tag in the table denotes that the aspects' importance under each dimension are assumed to be uniform, whereas the "with Priors" tag denotes that the aspects' importance of a query across each dimension are learned from the training data and using Equation (9). For the multidimensional approaches, the dimension importance is set using the Adaptive strategy described in the section entitled Multidimensional xQuAD.

Note: The best values for each case are shown in bold.

Table 7 reveals that both the flat and multidimensional diversification methods (with Priors) outperform the baseline especially for the top-2 results. We also find that the multidimensional approach with Uniform aspect priors yield inferior results both to the flat and multidimensional approaches (with Priors), and sometimes, even to the nondiversified baseline. This confirms our intuition that the diversification methods in this setup should incorporate realistic aspect priors learned from the user interactions, that is, clicks. The multidimensional approach with the Priors achieves the best results overall, with relative improvements over its flat counterpart reaching up to 2.0% (i.e., 0.350 vs. 0.343) and 1.4% (i.e., 0.442 vs. 0.436) for the diversification and relevance metrics P-IA@2 and P@2, respectively.

In our query log, since the number of instances for each query varies (i.e., the minimum and maximum number of instances is 4 and 97, respectively), it is worthwhile to investigate what happens if the diversification scores are first averaged over the instances of each query, and then over the queries (i.e., a macro averaging perspective); so that a query with too many instances does not dominate the overall performance and conclusions drawn.

Table 8 presents the diversification performance of the flat and multidimensional xQuAD approaches (both with Priors) by macro averaging the scores over queries. The trends are similar to those in Table 7, as multidimensional xQuAD outperforms both of its competitors, with even larger margins for the diversification metrics. In particular, the latter method achieves improvements of 1.6% and 2.6% over its flat counterpart in terms of α-nDCG@2 and P-IA@2, respectively. In other words, our gains presented in Table 7 still occur when the query frequency effect is eliminated from our evaluation.

Note: The best values for each case are shown in bold.

Overall, our evaluations based on the user clicks and relevance annotations (as presented in the current and

TABLE 7 Performances of flat and multidimensional xQuAD using the click-based evaluation

		Relevance				Diversity			
Div	**Method**	**P@2**	**P@5**	**nDCG@2**	**nDCG@5**	**P-IA@2**	**P-IA@5**	**α-nDCG@2**	**α-nDCG@5**
NonDiv		0.411	**0.306**	0.462	0.542	0.345	**0.228**	0.524	0.625
Flat	withPriors	0.436	0.303	0.477	0.544	0.343	0.226	0.532	0.630
M-Dim	Uniform	0.422	0.305	0.461	0.537	0.336	0.224	0.517	0.619
	withPriors	**0.442**	0.305	**0.484**	**0.545**	**0.350**	**0.228**	**0.539**	**0.632**

Note: The best values for each case are shown in bold.

TABLE 8 Performances of flat and multidimensional xQuAD with Priors (macro-averaging over queries)

	Relevance				Diversity			
	P@2	**P@5**	**nDCG@2**	**nDCG@5**	**P-IA@2**	**P-IA@5**	**α-nDCG@2**	**α-nDCG@5**
NonDiv	0.403	**0.285**	0.449	0.525	0.295	0.210	0.467	0.566
Flat	0.405	0.278	0.452	0.527	0.310	0.213	0.488	0.580
M-Dim	**0.407**	0.282	**0.453**	**0.528**	**0.318**	**0.214**	**0.496**	**0.583**

Note: The best values for each case are shown in bold.

previous sections, respectively) reveal that the proposed multidimensional diversification approach yields improvements of up to 2.6% for various relevance and diversification metrics (c.f. Tables 3–8), a finding that indicates the robustness of our approach for the educational search scenario addressed in this article.

6 | CONCLUSIONS

We introduced the multidimensional diversification of results in the context of educational search to help the users' learning-oriented search activities. Our proposed enhancement of the xQuAD diversification model (also applied to PM2 and R-LTR) allows the multiple dimensions that are available in this context to be taken into account when ranking documents, such as the type and target educational level of each document. Our extensive experiments upon a newly-created test collection using the logs of a real-life educational search engine show that our proposed approach can surface a variety of document types, education levels and topics within the top-ranked documents, and exhibits 2.6% improvement over traditionally strong "flat" diversification approaches and a marked 15.1% improvement over a BM25-based initial ranking obtained within a TREC-style evaluation framework, that is based on relevance annotations, for the ERR-IA metric.

We also employed another evaluation framework based on the user clicks. Contrary to the annotation-based evaluation, the click-based setup is sensitive to the users' learning preferences for query aspects, which vary wildly in practice, and hence, the diversification methods use aspect importance priors that are also obtained from the query logs. In this realistic evaluation framework, multidimensional diversification again proves to be useful, for instance, providing good gains of 1.4% and 7.5% for the P@2 metric over the "flat" diversification and nondiversified initial ranking, respectively.

In light of previous works (Collins-Thompson et al., 2016; Syed & Collins-Thompson, 2017), which showed the positive impact of diversified result presentation on the learning outcomes (e.g., knowledge gains of users), we would like to end this article by highlighting that our observed improvements in diversification performance using the traditional metrics are likely to materialize into human learning gains in the educational search context, which is the ultimate goal of our present investigation.

As future work, we plan to extend our multidimensional diversification framework by taking personalization into account, again for the purposes of enhancing learning in the educational search context.

ACKNOWLEDGMENTS

This work is partially funded by the Royal Society under the Newton International Exchanges Scheme with grant no. NI140231 and The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant no. 117E861. I.S. Altingovde is partially supported by Turkish Academy of Sciences Distinguished Young Scientist Award (TUBA-GEBIP 2016). Sevgi Yigit-Sert is supported by TÜBİTAK-BIDEB 2211/A program. The authors are also grateful for the collaboration with Sebit Inc. for providing the query logs used in this work.

ENDNOTES

[1] As discussed in the Related Work section, for general-purpose search engines, there are several earlier works, which aimed to diversify the results for ambiguous queries when a single (topical) dimension is used, as well as a few approaches that addressed a hierarchy of (or, dimensions for) the query aspects.

[2] https://github.com/syigitsert/multi-dim-diversification

[3] https://av.tib.eu

[4] In the section entitled Click-based Evaluation, we go beyond this assumption and learn the aspect importances from the users' clicks.

[5] http://www.vitaminegitim.com/

[6] i.e., targeting students aged 5–17 and covering primary, middle and high school educations (as in U.S.A.).

[7] For the type and education level dimensions, we only consider the aspects observed in the candidate set.

REFERENCES

Aktolga, E. (2014). *Integrating non-topical aspects into information retrieval* (Unpublished doctoral dissertation). Amherst: University of Massachusetts.

Aslam, J. A., & Montague, M. (2001). Models for metasearch. In *SIGIR* (pp. 276–284). New York, NY: ACM.

Bai, X., Cambazoglu, B. B., Gullo, F., Mantrach, A., & Silvestri, F. (2017). Exploiting search history of users for news personalization. *Information Sciences*, *385*(C), 125–137.

Carpineto, C., Mizzaro, S., Romano, G., & Snidero, M. (2009). Mobile information retrieval with search results clustering: Prototypes and evaluations. *Journal of the American Society for Information Science and Technology*, *60*(5), 877–895.

Clarke, C. L., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 Web track. In TREC. Gaithersburg, Maryland: NIST.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Collins-Thompson, K., Bennett, P. N., White, R. W., de la Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In *CIKM* (pp. 403–412). New York, NY: ACM.

Collins-Thompson, K., Hansen, P., & Hauff, C. (2017). Search as learning (Dagstuhl Seminar 17092). *Dagstuhl Reports*, *7*(2), 135–162.

Collins-Thompson, K., Rieh, S. Y., Haynes, C. C., & Syed, R. (2016). Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *CHIIR* (pp. 163–172). New York, NY: ACM.

Dang, V., & Croft, W. B. (2012). Diversity by proportionality: An election-based approach to search result diversification. In *SIGIR* (pp. 65–74). New York, NY: ACM.

Dou, Z., Hu, S., Chen, K., Song, R., & Wen, J.-R. (2011). Multi-dimensional search result diversification. In *WSDM* (pp. 475–484). New York, NY: ACM.

Dou, Z., Song, R., Yuan, X., & Wen, J. (2008). Are click-through data adequate for learning web search rankings? In *CIKM* (pp. 73–82). New York, NY: ACM.

Goynuk, B., & Altingovde, I. S. (2020). Supervised learning methods for diversification of image search results. In *ECIR* (pp. 158–165). Cham, Switzerland: Springer.

He, J., Meij, E., & de Rijke, M. (2011). Result diversification based on query-specific cluster ranking. *Journal of the American Society for Information Science and Technology*, *62*(3), 550–571.

Hoppe, A., Holtz, P., Kammerer, Y., Yu, R., Dietze, S., & Ewerth, R. (2018). Current challenges for studying search as learning processes. In *7th Workshop on Learning & Education with Web Data (LILE)*. https://www.tib.eu/fileadmin/Daten/dokumente/forschung-entwicklung/LILE_Workshop_SALIENT_position_paper.pdf.

Hu, S., Dou, Z., Wang, X., Sakai, T., & Wen, J.-R. (2015). Search result diversification based on hierarchical intents. In *CIKM* (pp. 63–72). New York, NY: ACM.

Hu, S., Dou, Z., Wang, X., & Wen, J. (2015). Search result diversification based on query facets. *Journal of Computer Science and Technology*, *30*(4), 888–901.

Jiang, Z., Dou, Z., Zhao, W. X., Nie, J., Yue, M., & Wen, J. (2018). Supervised search result diversification via subtopic attention. *IEEE Transactions on Knowledge and Data Engineering*, *30*(10), 1971–1984.

Liang, S., Ren, Z., & de Rijke, M. (2014). Personalized search result diversification via structured learning. In *KDD* (pp. 751–760). New York, NY: ACM.

Maxwell, D., Azzopardi, L., & Moshfeghi, Y. (2019). The impact of result diversification on search behaviour and performance. *Information Retrieval Journal*, *22*(5), 422–446.

Moraes, F., Putra, S. R., & Hauff, C. (2018). Contrasting search as a learning activity with instructor-designed learning. In *CIKM* (pp. 167–176). New York, NY: ACM.

Noia, T. D., Rosati, J., Tomeo, P., & Sciascio, E. D. (2017). Adaptive multi-attribute diversity for recommender systems. *Information Sciences*, *382*, 234–253.

Ozdemiray, A. M., & Altingovde, I. S. (2015). Explicit search result diversification using score and rank aggregation methods. *Journal of the American Society for Information Science and Technology*, *66*(6), 1212–1228.

Raman, K., Bennett, P. N., & Collins-Thompson, K. (2014). Understanding intrinsic diversity in web search: Improving whole-session relevance. *ACM Transactions on Information Systems*, *32*(4), 20:1–20:45.

Santos, R., Macdonald, C., & Ounis, I. (2010). Exploiting query reformulations for web search result diversification. In *WWW* (pp. 881–890). New York, NY: ACM.

Syed, R., & Collins-Thompson, K. (2017). Optimizing search results for human learning goals. *Information Retrieval Journal*, *20*(5), 506–523.

Usta, A., Altingövde, I. S., Vidinli, I. B., Ozcan, R., & Ulusoy, Ö. (2014). How k-12 students search for learning? Analysis of an educational search engine log. In *SIGIR* (pp. 1151–1154). New York, NY: ACM.

Vanopstal, K., Stichele, R. V., Laureys, G., & Buysschaert, J. (2012). Pubmed searches by dutch-speaking nursing students: The impact of language and system experience. *Journal of the American Society for Information Science and Technology*, *63*(8), 1538–1552.

Wang, X., Dou, Z., Sakai, T., & Wen, J.-R. (2016). Evaluating search result diversity using intent hierarchies. In *SIGIR* (pp. 415–424). New York, NY: ACM.

Wang, Y., Luo, Z., & Yu, Y. (2016). Learning for search results diversification in twitter. In *WAIM* (pp. 251–264). Cham, Switzerland: Springer.

Xia, L., Xu, J., Lan, Y., Guo, J., Zeng, W., & Cheng, X. (2017). Adapting Markov decision process for search result diversification. In *SIGIR* (pp. 535–544). New York, NY: ACM.

Yilmaz, T., Ozcan, R., Altingovde, I. S., & Ulusoy, Ö. (2019). Improving educational web search for question-like queries through subject classification. *Information Processing and Management*, *56*(1), 228–246.

Zhu, Y., Lan, Y., Guo, J., Cheng, X., & Niu, S. (2014). Learning for search result diversification. In *SIGIR* (pp. 293–302). New York, NY: ACM.

How to cite this article: Yigit-Sert S, Altingovde IS, Macdonald C, Ounis I, Ulusoy Ö. Explicit diversification of search results across multiple dimensions for educational search. *J Assoc Inf Sci Technol*. 2021;72:315–330. https://doi.org/10.1002/asi.24403