Personalized Search Result Diversification via Structured Learning

Shangsong Liang
s.liang@uva.nl

Zhaochun Ren
z.ren@uva.nl

Maarten de Rijke
derijke@uva.nl

University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

This paper is concerned with the problem of personalized diversification of search results, with the goal of enhancing the performance of both plain diversification and plain personalization algorithms. In previous work, the problem has mainly been tackled by means of unsupervised learning. To further enhance the performance, we propose a supervised learning strategy. Specifically, we set up a structured learning framework for conducting supervised personalized diversification, in which we add features extracted directly from the tokens of documents and those utilized by unsupervised personalized diversification algorithms, and, importantly, those generated from our proposed user-interest latent Dirichlet topic model. Based on our proposed topic model whether a document can cater to a user's interest can be estimated in our learning strategy. We also define two constraints in our structured learning framework to ensure that search results are both diversified and consistent with a user's interest. We conduct experiments on an open personalized diversification dataset and find that our supervised learning strategy outperforms unsupervised personalized diversification methods as well as other plain personalization and plain diversification methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Personalization; diversity; structured SVMs; ad hoc retrieval

1. INTRODUCTION

Search result diversification has recently gained attention as a method to tackle query ambiguity. In search result diversification one typically considers the relevance of a document in light of the other retrieved documents. The goal is to identify the probable "aspects" of the ambiguous query, retrieve documents for each of these aspects and make the search results more diverse [13]. By doing so, in the absence of any knowledge of users' context or preferences, the chance that any user issuing an ambiguous query will find at least one of these results to be relevant to the underlying information need is maximized [10].

In both search result diversification and personalized web search, an issued query is often viewed as an incomplete expression of a user's underlying need [22]. Unlike search result diversification, where the system accepts and adapts its behavior to a situation of uncertainty, personalized web search strives to change this situation by enhancing the system's knowledge about users' information needs. Rather than aiming to satisfy as many users as possible, personalization aims to build a sense of who the user is, and maximize the satisfaction of a specific user [26].

Although different, diversification and personalization are not incompatible and do not have mutually exclusive goals [23]. Search results generated by diversification techniques should be more diverse when a user's preferences are unrelated to the query. Likewise, personalization can improve the effectiveness of aspect weighting in diversification, by favoring query interpretations which are predicted to be more related to each specific user [26].

In this paper we study the problem of *personalized diversification of search results*, with the goal of enhancing both diversification and personalization performances. The problem has previously been investigated by Radlinski and Dumais [19] and Vallet and Castells [26]. They have presented a number of effective *unsupervised learning* approaches that combine both personalization and diversification components to tackle the problem. To further improve the performance we propose a *supervised learning* approach.

There are a couple of advantages to considering a supervised learning approach. Such approaches can leverage useful information underlying labeled training data, apply existing optimization techniques to the problem and are easier to adaptation. Of course, they also have disadvantages, one of which is that it is expensive to create training data for supervised learning methods. This is, however, a shortcoming for any supervised learning strategy and we leave it as future work.

Accordingly, we formulate the task of personalized search result diversification as a problem of predicting a diverse set of documents given a specific user and a query. Specifically, we formulate a discriminant based on maximizing search result diversification, and perform training using the well-known structured support vector machines (SSVMs) framework [25]. The main idea is first to propose a user-interest LDA-style [5, Latent Dirichlet Allocation] topic model, from which we can infer a per-document multinomial distribution over topics and determine whether a document can cater for a specific user. Then, during training we use features extracted directly from the tokens' statistical information in the documents and those utilized by unsupervised personalized diversification algorithms, and, more importantly, those generated from our proposed topic model. Additionally, two types of con-

KDD '14, August 24–27, 2014, New York, NY, USA.

ACM 978-1-4503-2956-9/14/08 ... $15.00.
http://dx.doi.org/10.1145/2623330.2623650 .

straint in SSVMs are explicitly defined to enforce the search results to be both diverse and relevant to a user's personal interest.

We evaluate our proposed approach on a publicly available personalized diversification dataset and compare it to unsupervised approaches, that focus on either personalization or diversification alone, to combined approaches like those in [19] and [26], and to two standard structured learning approaches [32, 33]. The three main contributions of our work are: (1) We tackle the problem of personalized diversification of search results differently, using a supervised learning method. (2) We propose a user-interest latent topic model to capture a user's interest and infer per-document multinomial distributions over topics. (3) We explicitly enforce diversity and personalization through two types of constraints in structured learning for personalized diversification.

2. RELATED WORK

Three major types of research relate to our work: personalized diversification, structured learning, and topic modeling.

2.1 Personalized search result diversification

Two main components, viz., personalized web search and search result diversification, play important roles in tackling the problem of personalized search result diversification. The task of personalized web search aims at identifying the most relevant search results for an individual by leveraging their information. Many personalized web search methods have been proposed, such as the one based on social tagging profiles [27], ranking model adaption for personalized search [29], search personalization by modeling the impact of users' behavior [4], and personalized search using interaction behaviors in search sessions [17]. In contrast, diversification aims to make the search results diversified given an ambiguous query so that users can find at least one of these results to be relevant to their underlying information need [2]. Well-known diversification methods include the maximal marginal relevance model [8], probabilistic model [9], subtopic retrieval model [35], xQuAD [21], RxQuAD [28], IA-select [2], PM-2 [12], and more recently, matroid constraints [1], DSPApprox [13], text-based measures [3], term-level [13], and fusion-based [16]. All of the above methods focus on either personalization or diversification only.

To the best of our knowledge, only Radlinski and Dumais [19] and Vallet and Castells [26] have studied the problem of combining both personalization and diversification. Radlinski and Dumais [19] analyze a large sample of individual users' query logs from a web search engine such that individual users' query reformulations can be obtained. Then they personalize web search by reranking some top results using query reformulations to introduce diversity into those results. Their evaluation suggests that using diversification is a promising method to improve personalized reranking of search results. Vallet and Castells [26] present a number of approaches that combine personalization and diversification components. They investigate the introduction of the user as an explicit variable in state-of-the-art diversification models. Their personalized search result diversification algorithms achieve competitive performance and improve over plain personalization and plain diversification baselines.

All of the previous personalized diversification models are unsupervised. However, we argue that to enhance the performance, it is better to employ a supervised learning approach, and our experiments show that supervised learning can indeed improve the performance of unsupervised approaches. To the best of our knowledge, this is the first attempt to tackle the problem of personalized diversification using supervised learning methods.

2.2 Structured learning

Structured learning has provided principled techniques for learning structured-output models, with the structured support vector machines (SSVMs) being one of the most important ones [25]. In structured learning, a set of training pairs, $\{(\mathbf{x}, \mathbf{y})|\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}\}$, is assumed to be available to the learning algorithm, and the goal is to learning a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ from the input space $\mathcal{X}$ to the output space $\mathcal{Y}$, such that a regularized task-dependent loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ can be minimized, where $\Delta(\mathbf{y}, \bar{\mathbf{y}})$ denotes the cost of predicting output $\bar{\mathbf{y}}$ when the correct prediction is $\mathbf{y}$. In the past few years, Structured SVMs (SSVMs) have been studied and applied in many areas, such as speech recognition [36], optimizing average precision of a ranking [33], and diversification [32]. For us, the most interesting prior application of SSVMs is the one for predicting diverse subsets [32]. However, our personalized search result diversification method differs from that proposed in [32]: we work on personalized diversification where we propose a user-interest LDA-style model to capture a user's interest distribution over topics, whereas they directly apply existing SSVMs algorithm to tackle the problem of search result diversification but not personalized diversification; our model explicitly makes results diverse and consistent to the user's interest by enforcing both diversity and interest constraints, whereas their model only implicitly diversifies the results by adopting standard SSVMs. Prior work on diversification [12, 21, 28], however, has shown that explicit approaches outperform implicit ones in most cases. To the best of our knowledge, this is the first attempt to explicitly enforce diversity and personalization through additional constraints in SSVMs.

2.3 Topic modeling

Topic modeling provides a suite of algorithms to discover hidden thematic structure in a collection of documents. A topic model takes a collection of documents as input, and discovers a set of "latent topics"—recurring themes that are discussed in the collection—and the degree to which each document exhibits those topics [5]. *Latent dirichlet allocation* (LDA) [5] is one of the simplest topic models, and it decomposes a collection of documents into topics—biased probability distributions over terms—and represents each document with a subset of these topics. Many LDA-style models have been proposed, such as the syntactic topic model [6], multilingual topic model [7], topic over time model [30], and more recently, the max-margin model [37], spatio-temporal model [31], fusion-based model [16] and multi-contextual model [24]. We propose a user-interest LDA-style model to capture a multinomial distribution of topics specific to a user. From our model, we infer a per-document multinomial distribution over the topics so that we can easily identify whether a document caters to a user's interest. Our experimental results demonstrate that the model can help to enhance the performance of personalized search result diversification. To the best of our knowledge, this is the first time that a topic model is utilized to enhance the performance of personalized diversification.

3. THE LEARNING PROBLEM

Let $\mathbf{u} = \{d_1, \ldots, d_{|\mathbf{u}|}\} \in \mathcal{U}$ be a set of documents of size $|\mathbf{u}|$ which a user u is interested in. For each query q, we assume that we are given $\mathbf{u}$ and a set of candidate documents $\mathbf{x} = \{x_1, \ldots, x_{|\mathbf{x}|}\} \in \mathcal{X}$, where $\mathcal{X}$ denotes the set of all possible document sets. Our task is to select a subset $\mathbf{y} \in \mathcal{Y}$ of K documents from $\mathbf{x}$ that maximizes the performance of personalized search result diversification given q and $\mathbf{u}$, where we let $\mathcal{Y}$ denote the space

of predicted subsets $\mathbf{y}$. Following the standard machine learning setup, we formulate our task as learning a hypothesis function $h : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ to predict a $\mathbf{y}$ given $\mathbf{x}$ and $\mathbf{u}$. To this end, we assume that a set of labeled training data is available:

$$\{(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)}) \in \mathcal{X} \times \mathcal{U} \times \mathcal{Y} : i = 1, \ldots, N\},$$

where $\mathbf{y}^{(i)}$ is the ground-truth subset of K documents from $\mathbf{x}^{(i)}$, and $\mathbf{u}^{(i)}$ is the set of documents that user u_i is interested in, and N is the size of the training data. We aim to find a function h such that the empirical risk $R_S^{\Delta}(h) = \frac{1}{N}\sum_{i=1}^{N} \Delta(\mathbf{y}^{(i)}, h(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}))$ can be minimized, where we quantify the quality of a prediction by considering a loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ that measures the penalty of choosing $\bar{\mathbf{y}} = h(\mathbf{x}^{(i)}, \mathbf{u}^{(i)})$. Here, given the ground-truth $\mathbf{y}$, viz., the ground truth ranking of relevant documents, and the predicting $\bar{\mathbf{y}}$, viz., the ranking of predicted documents, we define the loss function based on a diversity metric, α-nDCG [10] (other diversity metrics are possible but we obtain the best performance when adopting this metric), as:

$$\Delta(\mathbf{y}, \bar{\mathbf{y}}) = 1 - \alpha\text{-nDCG}(\mathbf{y}, \bar{\mathbf{y}}). \quad (1)$$

We focus on hypothesis functions which are parameterized by a weight vector $\mathbf{w}$, and thus wish to find $\mathbf{w}$ to minimize the risk, $R_S^{\Delta}(\mathbf{w}) \equiv R_S^{\Delta}(h(\cdot; \mathbf{w}))$. We let a discriminant $\mathcal{F} : \mathcal{X} \times \mathcal{U} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ compute how well the predicting $\bar{\mathbf{y}}$ fits for $\mathbf{x}$ and $\mathbf{u}$. Then the hypothesis predicts the $\bar{\mathbf{y}}$ that maximizes $\mathcal{F}$:

$$\bar{\mathbf{y}} = h(\mathbf{x}, \mathbf{u}; \mathbf{w}) = \arg\max_{\mathbf{y} \in \mathcal{Y}} \mathcal{F}(\mathbf{x}, \mathbf{u}, \mathbf{y}). \quad (2)$$

We describe each $(\mathbf{x}, \mathbf{u}, \mathbf{y})$ through a feature vector $\Psi(\mathbf{x}, \mathbf{u}, \mathbf{y})$; the extraction will be discussed later. The discriminant function $\mathcal{F}(\mathbf{x}, \mathbf{u}, \mathbf{y})$ is assumed to be linear in the feature vector $\Psi(\mathbf{x}, \mathbf{u}, \mathbf{y})$ such that:

$$\mathcal{F}(\mathbf{x}, \mathbf{u}, \mathbf{y}) = \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{u}, \mathbf{y}), \quad (3)$$

where $\mathbf{w}$ is a weight vector to be learned from training data.

4. STRUCTURED LEARNING FOR PERSONALIZED DIVERSIFICATION

In this section, we introduce the standard SSVMs learning problem, propose constraints for personalized diversification and describe our optimization problem and the way we make predictions.

4.1 Standard structured SVMs

Our personalized diversification model builds on an existing standard structured learning framework. In our setting, the standard structured learning framework can be described as: given a training set $\{(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)}) \in \mathcal{X} \times \mathcal{U} \times \mathcal{Y} : i = 1, \ldots, N\}$, structured SVMs are employed to learn a weight vector $\mathbf{w}$ for the discriminant function $\mathcal{F}(\mathbf{x}, \mathbf{u}, \mathbf{y})$ through the following quadratic programing problem:

Optimization Problem 1. (Standard structured SVMs)

$$\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{N}\sum_{i=1}^{N} \xi_i \quad (4)$$

subject to $\forall i, \forall \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}^{(i)}, \xi_i \geq 0$,

$$\mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}) + \Delta(\mathbf{y}^{(i)}, \mathbf{y}) - \xi_i.$$

In the objective function (4), the parameter C is a tradeoff between model complexity, $||\mathbf{w}||^2$, and a hinge loss relaxation of the training loss for each training example, $\sum \xi_i$. The constraints enforce the requirement that the ground-truth personalized diversity document set $\mathbf{y}^{(i)}$ should have a greater function value than other alternative $\mathbf{y} \in \mathcal{Y}$, and $\mathbf{y} \neq \mathbf{y}^{(i)}$.

4.2 Additional constraints

As discussed above, we aim at training a personalized diversification model that can enforce both diversity and consistency with the user's interest. This can be achieved by introducing additional constraints to the structured SVM optimization problem defined in (4). To start, diversity requires a set of retrieved documents that should not discuss the same aspects of an ambiguous query. In other words, aspects of documents returned by a diversification model should have little overlap with one another. Formally, we enforce diversity with the following constraint.

Constraint for diversity:

$$\mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)}) \geq \sum_{y \in \mathbf{y}^{(i)}} \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, y) - \xi_i. \quad (5)$$

In (5), the sum of each document's score, $\sum \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, y)$, should not be greater than the overall score when they are considered as an ideal ranking of the document sets. As a result, commonly shared features will be associated with relatively low weights, and a document set with less redundancy will be predicted.

Additionally, personalization requires a set of returned documents to match the user's personal interest. Formally, we enforce personalization with the following constraint.

Constraint for consistency with user's interest:

$$\mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}) + (1 - \text{sim}(\mathbf{y}, \mathbf{u}^{(i)})) - \mu - \xi_i, \quad (6)$$

where $\text{sim}(\mathbf{y}, \mathbf{u}^{(i)}) \in [0, 1]$ is a function (see (14)) that measures subtopic distribution similarity between a set of documents $\mathbf{y}$ and the documents user u_i is interested in, i.e., $\mathbf{u}^{(i)}$, μ is a slack variable that tends to give slightly better performance, which can be defined as $\mu = \frac{1}{N}\sum_{i=1}^{N}(1 - \text{sim}(\mathbf{y}^{(i)}, \mathbf{u}^{(i)}))$.

In (6), $(1 - \text{sim}(\mathbf{y}, \mathbf{u}^{(i)}))$ quantifies how well a set of documents matches a user's interest. If the topics discussed in a set of documents $\mathbf{y}$ are not consistent with a user's personal interest, $\mathbf{w}^T \Psi(\mathbf{x}, \mathbf{u}, \mathbf{y})$ will result in a relatively low score. During prediction, documents consistent with a user's interest will be preferred.

4.3 Our optimization problem

A set of documents produced in response to an ambiguous query should be diverse and consistent to the user's personal interest. To this end we integrate the proposed additional constraints with standard structured SVMs. We propose to train a personalized diversification model by tackling the following optimization problem:

Optimization Problem 2. (Structured SVMs for personalized diversification)

$$\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{N}\sum_{i=1}^{N} \xi_i \quad (7)$$

subject to $\forall i, \forall \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}^{(i)}, \xi_i \geq 0$,

i. $\mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}) + \Delta(\mathbf{y}^{(i)}, \mathbf{y}) - \xi_i,$

ii. $\mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)}) \geq \sum_{y \in \mathbf{y}^{(i)}} \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, y) - \xi_i,$

iii. $\mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}) + ((1 - \text{sim}(\mathbf{y}, \mathbf{u}^{(i)})) - \mu) - \xi_i.$

Algorithm 1: Cutting plane algorithm

Input : $(\mathbf{x}^{(1)}, \mathbf{u}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(N)}, \mathbf{u}^{(N)}, \mathbf{y}^{(N)}), C, \epsilon$

$\mathcal{W}_i \leftarrow \emptyset, \mathcal{W}'_i \leftarrow \emptyset, \mathcal{W}''_i \leftarrow \emptyset$ for all $i = 1, \ldots, N$

$\mu = \frac{1}{N} \sum_{i=1}^{N} (1 - \text{sim}(\mathbf{y}^{(i)}, \mathbf{u}^{(i)}))$

repeat

  for $i = 1, \ldots, N$ **do**

    $H(\mathbf{y}; \mathbf{w}) \equiv \Delta(\mathbf{y}^{(i)}, \mathbf{y}) + \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}) - \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)})$

    $H'(\mathbf{y}; \mathbf{w}) \equiv \sum_{y \in \mathbf{y}^{(i)}} \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, y) - \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)})$

    $H''(\mathbf{y}; \mathbf{w}) \equiv \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}) +$

    $((1 - \text{sim}(\mathbf{y}, \mathbf{u}^{(i)})) - \mu) - \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)})$

    compute $\bar{\mathbf{y}} = \text{argmax}_{\mathbf{y}} H(\mathbf{y}; \mathbf{w})$,

    $\bar{\mathbf{y}}' = \text{argmax}_{\mathbf{y}} H'(\mathbf{y}; \mathbf{w})$ and $\bar{\mathbf{y}}'' = \text{argmax}_{\mathbf{y}} H''(\mathbf{y}; \mathbf{w})$

    compute $\xi_i = \max\{0, \max_{\mathbf{y} \in \mathcal{W}_i} H(\mathbf{y}; \mathbf{w}), \max_{\mathbf{y} \in \mathcal{W}'_i} H'(\mathbf{y}; \mathbf{w}), \max_{\mathbf{y} \in \mathcal{W}''_i} H''(\mathbf{y}; \mathbf{w})\}$

    if $H(\bar{\mathbf{y}}; \mathbf{w}) > \xi_i + \epsilon$ **or** $H'(\bar{\mathbf{y}}'; \mathbf{w}) > \xi_i + \epsilon$ **or** $H''(\bar{\mathbf{y}}''; \mathbf{w}) > \xi_i + \epsilon$ **then**

      Add constraint to working set $\mathcal{W}_i \leftarrow \mathcal{W}_i \cup \{\bar{\mathbf{y}}\}$, $\mathcal{W}'_i \leftarrow \mathcal{W}'_i \cup \{\bar{\mathbf{y}}'\}, \mathcal{W}''_i \leftarrow \mathcal{W}''_i \cup \{\bar{\mathbf{y}}''\}$

      $\mathbf{w} \leftarrow$ optimize (7) over $\bigcup_i \{\mathcal{W}_i, \mathcal{W}'_i, \mathcal{W}''_i\}$

until *no* $\mathcal{W}_i$, $\mathcal{W}'_i$ *and* $\mathcal{W}''_i$ *have changed during iteration*

Algorithm 2: Greedy subset selection for prediction

Input : $\mathbf{w}, \mathbf{x}, \mathbf{u}$

$\bar{\mathbf{y}} \leftarrow \emptyset$

for $k = 1, \ldots, K$ **do**

  $\bar{x} = \arg\max_{x: x \in \mathbf{x}, x \notin \bar{\mathbf{y}}} \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{u}, \bar{\mathbf{y}} \cup \{x\})$

  $\bar{\mathbf{y}} \leftarrow \bar{\mathbf{y}} \cup \{\bar{x}\}$

return $\bar{\mathbf{y}}$

4.4 The learning algorithm

We can solve the optimization problem defined in (7) by employing the cutting plane algorithm [25]. The learning algorithm is shown in Algorithm 1. The algorithm iteratively adds constraints until we have solved the original problem within a desired tolerance ϵ. It starts with empty working sets $\mathcal{W}_i$, $\mathcal{W}'_i$ and $\mathcal{W}''_i$, for $i = 1, \ldots, N$. Then it iteratively finds the most violated constraints $\bar{\mathbf{y}}$, $\bar{\mathbf{y}}'$ and $\mathbf{y}''$ for each $(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}, \mathbf{y}^{(i)})$ in terms of the three constraints (i), (ii) and (iii) in (7), respectively. If they are violated by more than ϵ, we add them into the corresponding working sets. We iteratively update $\mathbf{w}$ by optimizing (7) over the updated working sets. The outer loop in Algorithm 1 can halt within a polynomial number of iterations for any desired precision ϵ; see [25].

4.5 Prediction

After $\mathbf{w}$ has been learned, given an ambiguous query, a set of candidate documents $\mathbf{x}$, and a set of documents $\mathbf{u}$ the user u is interested in, we try to predict a set of documents $\bar{\mathbf{y}}$ by tackling the following prediction problem:

$$\bar{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} \mathcal{F}(\mathbf{x}, \mathbf{u}, \mathbf{y}) = \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{u}, \mathbf{y}). \quad (8)$$

This is a special case of the Budgeted Max Coverage problem [15], and can be efficiently solved by Algorithm 2.

5. USER-INTEREST TOPIC MODEL AND FEATURE SPACE

In this section, we first review the notation and terminology used in our user-interest topic model, and then describe the model and the features used in our structured learning framework.

Table 1: Main notation used in user-interest topic model.

Notation	Gloss	Notation	Gloss
q	query	d	document
u	user	z	topic
T	number of topics	U	number of users
D	number of documents	V	number of tokens
N_d	number of tokens in d		
$\widetilde{\mathbf{u}}$	a set of users	$\widetilde{\mathbf{w}}$	a set of tokens
b_z	Beta distribution parameter for z		
α	the parameter of user Dirichlet prior		
β	the parameter of token Dirichlet prior		
θ_d	multinomial distribution of topics specific to d		
ϕ_z	multinomial distribution of tokens specific to z		
ϑ_u	multinomial distribution of topics specific to u		
z_{di}	topic associated with the i-th token in d		
w_{di}	the i-th token in d		
r_{di}	relevance of the i-th token in d		

5.1 Notation and terminology

We summarize the main notation used in our user-interest topic model (UIT) in Table 1. We distinguish between queries, aspects and topics. A *query* is a user's expression of an information need. An *aspect* (sometimes called *subtopic* at the TREC Web tracks [11]) is an interpretation of an information need. We use *topic* to refer to latent topics as identified by a topic modeling method (LDA).

5.2 User-interest topic model

To capture per-user and per-document multinomial distributions over topics such that we can measure whether a document can cater for the user's interest, we propose a user-interest latent topic model (UIT). Topic discovery in UIT is influenced not only by token co-occurrences, but also by the relevance scores of documents evaluated by users. In our UIT model, we use a Beta distribution over a (normalized) document relevance span covering all the data, and thus various skewed shapes of rising and falling topic prominence can be flexibly represented.

The latent topic model used in UIT is a generative model of relevance and tokens in the documents. The generative process used in Gibbs sampling [18] for its parameter estimation, is as follows:

i. Draw T multinomials ϕ_z from a Dirichlet prior β, one for each topic z;

ii. For each user u, draw a multinomial ϑ_u from a Dirichlet prior α; then for each token w_{di} in document $d \in \mathbf{u}$:

  (a) Draw a topic z_{di} from multinomial ϑ_u;

  (b) Draw a token w_{di} from multinomial $\phi_{z_{di}}$;

  (c) Draw a relevance score r_{di} for w_{di} from Beta $(b_{z_{di}1}, b_{z_{di}2})$.

Fig. 1 shows a graphical representation of our model. In the generative process, the relevance scores of tokens observed in the same document are the same and evaluated by a user, although a relevance score is generated for each token from the Beta distribution. In our experiments, there is a fixed number of latent topics, T, although a non-parametric Bayes version of UIT that automatically integrates over the number of topics would certainly be possible. The posterior distribution of topics depends on the information from two modalities: tokens and the documents' relevance scores.

Inference is intractable in this model. Following [6, 7, 14, 18, 30], we employ Gibbs sampling to perform approximate inference. We adopt a conjugate prior (Dirichlet) for the multinomial distributions, and thus we can easily integrate out ϑ and ϕ, analytically capturing the uncertainty associated with them. In this way we

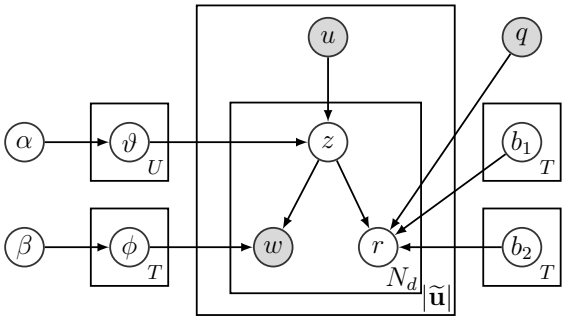

Figure 1: Graphical representation of user-interest topic model.

facilitate the sampling, i.e., we need not sample ϑ and ϕ at all. Because we use the continuous Beta distribution rather than discretizing document relevance scores, sparsity is not a big concern in fitting the model. For simplicity and speed we estimate these Beta distributions (b_{z1}, b_{z2}) by the method of moments, once per iteration of Gibbs sampling. We find that the sensitivity of the hyperparameters α and β is not very strong. Thus, for simplicity, we use fixed symmetric Dirichlet distributions ($\alpha = 50/T$ and $\beta = 0.1$) in all our experiments.

In the Gibbs sampling procedure above, we need to calculate the conditional distribution $P(z_{di}|\widetilde{\mathbf{w}}, \mathbf{r}, \mathbf{z}_{-di}, \widetilde{\mathbf{u}}, \alpha, \beta, \mathbf{b}, q)$, where $\mathbf{z}_{-di}$ represents the topic assignments for all tokens except w_{di}. We begin with the joint probability of a dataset, and using the chain rule, we can obtain the conditional probability conveniently as:

$$P(z_{di}|\widetilde{\mathbf{w}}, \mathbf{r}, \mathbf{z}_{-di}, \widetilde{\mathbf{u}}, \alpha, \beta, \mathbf{b}, q) \propto \frac{n_{z_{di}w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V}(n_{z_{di}v} + \beta_v) - 1} \frac{n_{u_{di}z_{di}} + \alpha_{z_{di}} - 1}{\sum_{z=1}^{T}(n_{u_{di}z} + \alpha_z) - 1} \times \frac{(1 - r_{di})^{b_{z_{di}1}-1} r_{di}^{b_{z_{di}2}-1}}{B(b_{z_{di}1}, b_{z_{di}2})},$$

where n_{zv} is the total number of tokens v that are assigned to topic z, n_{uz} represents the number of topics z that are assigned to user u.

After the Gibbs sampling procedure, we can easily infer a user's interest, i.e., multinomial distributions over topics for user u as:

$$\vartheta_{uz} = p(z|u) = \frac{n_{uz} + \alpha_z}{\sum_{z=1}^{T}(n_{uz} + \alpha_z)}, \tag{9}$$

and easily infer multinomial distributions over tokens for topic z:

$$\phi_{zv} = p(v|z) = \frac{n_{zv} + \beta_v}{\sum_{v=1}^{V}(n_{zv} + \beta_v)}, \tag{10}$$

where n_{zv} is the number of tokens of word v that are assigned to topic z. To obtain the multinomial distribution over topics for document d, i.e., θ_{dz}, we first apply the Bayes' rule:

$$\theta_{dz} = p(z|d) = \frac{p(d|z)p(z)}{p(d)}, \tag{11}$$

where $p(d|z)$ is the probability of d belonging to topic z, and $p(z)$ is the probability of topic z. According to (10), $p(d|z)$ can be obtained as $p(d|z) = \prod_{v \in d} p(v|z) = \prod_{v \in d} \phi_{zv}$. According to (9), $p(z)$ can be obtained as $p(z) = \sum_{u=1}^{U} p(z|u)p(u)$, where U is the total number of users. Therefore, 11 can be represented as:

$$\theta_{dz} = \frac{\prod_{v \in d} \phi_{zv} \sum_{u=1}^{U} p(z|u)p(u)}{p(d)}. \tag{12}$$

As any d has the same chance to be considered to be returned in response to q, we can assume that $p(d)$ is a constant, and likewise we also assume that $p(u)$ is a constant, such that (12) becomes:

$$\theta_{dz} = \frac{1}{E} \prod_{v \in d} \phi_{zv} \sum_{u=1}^{U} \vartheta_{uz}, \tag{13}$$

where $E = \sum_{z=1}^{T} \prod_{v \in d} \phi_{zv} \sum_{u=1}^{U} \vartheta_{uz}$ is a normalization constant. Then, the topic distribution similarity $\mathrm{sim}(\mathbf{y}, \mathbf{u})$ between a set of documents $\mathbf{y}$ and the documents $\mathbf{u}$ a user u is interested in can be measured as:

$$\mathrm{sim}(\mathbf{y}, \mathbf{u}) = \frac{1}{|\mathbf{y}|} \sum_{d \in \mathbf{y}} \cos(\theta_d, \vartheta_u), \tag{14}$$

where vectors $\theta_d = (\theta_{d1}, \cdots, \theta_{dT})$ and $\vartheta_u = (\vartheta_{d1}, \cdots, \vartheta_{dT})$ are the multinomial distribution of topics specific to document d and user u, respectively. We use the cos function in (14); other distance functions such as one based on Euclidean distance can be employed but we found that the results were not significantly different.

5.3 Feature space

The feature representation Ψ must enable meaningful discrimination between high quality and low quality predictions [32]. To predict a set of documents in the personalized diversification task, we propose to consider three main types of feature space.

Tokens. Following [32], we define L token sets $V_1(\mathbf{y}), \ldots, V_L(\mathbf{y})$. Each token set $V_l(\mathbf{y})$ contains the set of tokens that appear at least l times in some document in $\mathbf{y}$. Then we use thresholds on the ratio $|D_l(v)|/|\mathbf{u}|$ (or $|D_l(v)|/|\mathbf{x}|$) to define feature values of $\psi_l(v, \mathbf{u})$ (or $\psi_l(v, \mathbf{x})$) that describe word v at l-th importance level. Here, $D_l(v)$ is the set of documents that have at least l copies of v in the whole set of documents $\mathbf{u}$ (or $\mathbf{x}$). We let $L = 20$ in our experiments, as quite a few tokens can appear more than 20 times in a document. Besides, we propose to directly utilize the tokens' statistics to capture similarity between a document $x \in \mathbf{y}$ and a set of documents $\mathbf{u}$ that a user u is interested in as features. We consider cosine, Euclidean and Kullback-Leibler (KL) divergence similarity metrics. For each of these three metrics, we compute the minimal, maximal, and average similarity scores of the document $x \in \mathbf{y}$ and the standard deviations to a set of documents $\mathbf{u}$ based on the content of the documents and the standard LDA model [5]. In total, we have 49 features that fall in this feature category.

Interest. In addition, based on our UIT topic model, we also compute the cosine, Euclidean and KL similarity between a document $x \in \mathbf{y}$ and a set of documents $\mathbf{u}$ based on a multinomial distribution over topics and the user's multinomial distribution over topics generated by UIT. Again, for each of these three similarity metrics, we compute the minimal, maximal, and average similarity scores and the standard deviation scores. In total, we have $S = 36$ features $\omega_s(x, \mathbf{u})$ that fall in this feature category.

Probability. The main probabilities used in state-of-the-art unsupervised personalized diversification methods are utilized in our learning model as features, i.e., $\gamma_m(x, \mathbf{x}, \mathbf{u})$. These probabilities include $p(d|q)$, the probability of d relevant to q, $p(c|d)$, the probability of d belonging to a category c, $p(c|q, u)$, the personalized query aspect distribution, $p(c|d, u)$, the personalized aspect distribution over d, and $p(d|c, u)$, the personalized aspect-dependent document distribution, where c is a category that d belongs to in the Textwise Open Directory Project category service.[1] For $p(d|q)$, we obtain 3 versions of this feature value produced by BM25 [20], Jelinek-Mercer and Dirichlet language models [34]. To get the feature value of $p(c|d)$, we make use of the Textwise service which returns up to 3 possible categories for d, ranked by a score in $[0, 1]$, and we use the normalized scores as features. We adopt 5 ways of

[1] http://textwise.com

computing $p(c|q,u)$ as feature values [26]; for details on how to compute $p(c|q,u)$, $p(c|d,u)$ and $p(d|c,u)$ we refer to [26].

Then, we define $\Psi(\mathbf{x},\mathbf{u},\mathbf{y})$ as follows:

$$\Psi(\mathbf{x},\mathbf{u},\mathbf{y}) = \begin{bmatrix} \frac{1}{|\mathbf{y}|}\sum_{v\in V_1(\mathbf{y})}\psi_1(v,\mathbf{u}) \\ \frac{1}{|\mathbf{y}|}\sum_{v\in V_1(\mathbf{y})}\psi_1(v,\mathbf{x}) \\ \vdots \\ \frac{1}{|\mathbf{y}|}\sum_{v\in V_L(\mathbf{y})}\psi_L(v,\mathbf{u}) \\ \frac{1}{|\mathbf{y}|}\sum_{v\in V_L(\mathbf{y})}\psi_L(v,\mathbf{x}) \\ \frac{1}{|\mathbf{y}|}\sum_{x\in\mathbf{y}}\omega_1(x,\mathbf{u}) \\ \vdots \\ \frac{1}{|\mathbf{y}|}\sum_{x\in\mathbf{y}}\omega_S(x,\mathbf{u}) \\ \frac{1}{|\mathbf{y}|}\sum_{x\in\mathbf{y}}\gamma_1(x,\mathbf{x},\mathbf{u}) \\ \vdots \\ \frac{1}{|\mathbf{y}|}\sum_{x\in\mathbf{y}}\gamma_M(x,\mathbf{x},\mathbf{u}) \end{bmatrix}.$$

6. EXPERIMENTAL SETUP

In this section, we describe our experimental setup; §6.1 lists our research questions; §6.2 describes our dataset; §6.3 and §6.4 lists the baselines and metrics for evaluation, respectively; §6.5 details the settings of the experiments.

6.1 Research questions

The research questions guiding the remainder of the paper are: (RQ1) Can supervised personalized diversification methods outperform state-of-the-art unsupervised methods? Can our method beat state-of-the-art supervised methods? See §7.1. (RQ2) What is the contribution of the user-interest topic model in our proposed method? See §7.2. (RQ3) What is the effect of the constraints for diversity and consistence with user's interest in our method? See §7.3. (RQ4) Does our method outperform the best supervised baseline method on each query? See §7.4. (RQ5) Can our method retrieve a competitive number of subtopics per query? See §7.5. (RQ6) What is the performance of our supervised methods when the C parameter is varied? See §7.6.

6.2 Dataset

In order to answer our research questions we work with a publicly available personalized diversification dataset.[2] This dataset contains private evaluation information from 35 users on 180 search queries. The queries are quite ambiguous, as the length of each query is no more that two keywords. In total, there are 751 subtopics for the queries, with most of the queries having more than 2 subtopics. Over 3800 relevance judgements are available, for at least the top 5 results for each query. Each relevance judgement includes 3 main assessments: a 4-grade scale assessment on how relevant the result is to the user's interests (resulting in the *user relevance* ground truth and the set of users' interesting documents being created); a 4-grade scale assessment on how relevant the result is to the evaluated query (resulting in the *topic relevance* ground truth being created); and a 2-grade assessment whether a specific subtopic is related to the evaluated query (resulting in the subjective subtopics related to the search query being created). Details of this dataset can be found in [26]. For pre-processing, we apply Porter stemming, tokenization, and stopword removal (using the INQUERY list) to the documents using the Lemur toolkit.[3]

[2] http://ir.ii.uam.es/~david/persdivers/

[3] http://www.lemurproject.org

Two well-known corpora, ClueWeb09 and ClueWeb12,[4] have been proposed for search result diversification tasks in the TREC 2009–2013 Web tracks [11]. However, they do not contain any user information or relevance judgments provided by specific users, and thus do not fit our experiments.

6.3 Baselines

Let PSVM$_{div}$ denote our personalized diversification via structured learning method. We compare PSVM$_{div}$ to 11 baselines: a traditional web search algorithm, BM25 [20]; 2 well-known plain (in the sense of "not personalized") search result diversification approaches, IA-Select [2] and xQuAD [21]; a plain (in the sense of "not diversified") personalized search approach based on BM25 [27], Pers$_{BM25}$; a two-stage diversification and personalization approach, xQuAD$_{BM25}$, as suggested by [19], that first applies the xQuAD algorithm and then Pers$_{BM25}$; 4 state-of-the-art unsupervised personalized diversification methods [26], PIA-Select, PIA-Select$_{BM25}$, PxQuAD, and PxQuAD$_{BM25}$. As PSVM$_{div}$ builds on standard structured learning framework, we also consider 2 structured learning algorithms: SVM$_{div}$ [32] that directly tries to retrieve relevant documents covering as many subtopics as possible, and a standard structured learning method, denoted as SVM$_{rank}$ [33] that directly ranks documents by optimizing a relevance-biased evaluation metric (we use α-nDCG and nDCG to define the loss functions for SVM$_{div}$ and SVM$_{rank}$, respectively).[5]

For the supervised methods, PSVM$_{div}$, SVM$_{div}$ and SVM$_{rank}$, we use a 130/40/10 split for our training, validation and test sets, respectively. We train PSVM$_{div}$, SVM$_{div}$ and SVM$_{rank}$ using values of C (see (7)) that vary from 1e-4 to 1.0. The best C value is then chosen on the validation set, and evaluated on the test queries. The train/validation/test splits are permuted until all 180 queries were chosen once for the test set. We repeat the experiments 10 times and report the average evaluation results.

6.4 Evaluation

We use the following diversity metrics for evaluation, most of which are official evaluation metrics in the TREC Web tracks [11] and are widely used in the literature on result diversification:

α-nDCG@k. A version of normalized discounted cumulative gain at k in which the role of the parameter α is emphasized in computing the novelty of the top k documents. α-nDCG@k scores a ranking by rewarding newly-found subtopics and penalizing redundant subtopics geometrically, discounting all rewards with a log-harmonic discount function of rank. See [10] for details on how α-nDCG@k is computed.

S-Recall@k. Subtopic recall at k [35] is computed at retrieval depth k using the following procedure. Assume there are Q ambiguous queries. Let z be an aspect of query q and N_q the number of aspects (subtopics) associated with q. Then, the subtopic recall at rank k [35] is defined as the percentage of subtopics covered by one of the top k documents:

$$\text{S-Recall@}k = \frac{1}{Q}\sum_{q=1}^{Q}\frac{|\bigcup_{i=1}^{k}\text{subtopics}(d_i|q)|}{N_q},$$

where subtopics$(d_i|q)$ is the number of aspects covered by d_i in response to q.

ERR-IA@k. Intent-aware expected reciprocal rank at retrieval depth k, similarly, is computed as

$$\text{ERR-IA@}k = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{N_q}\sum_{z=1}^{N_q}\text{ERR}(k|z,q),$$

[4] http://boston.lti.cs.cmu.edu/clueweb12/

[5] The source code for SVM$_{rank}$ [33] and SVM$_{div}$ [32] is available at http://www.cs.cornell.edu/People/tj/.

where $\text{ERR}(k|z,q)$ is the expected reciprocal rank score at k in terms of aspect z of query q.

Prec-IA@k. Intent-aware precision at k [2] is defined as

$$\text{Prec-IA@}k = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{N_q}\sum_{z=1}^{N_q}\text{Prec}(k|z,q),$$

where $\text{Prec}(k|z,q)$ is the precision at k in terms of the aspects z of q, and can be computed as $\frac{1}{k}\sum_{j=1}^{k} j_q(z,j)$. Here, $j_q(z,j)=1$ if the document returned for q at depth j is judged relevant to aspect z of q; otherwise, $j_q(z,j)=0$.

MAP-IA@k. Intent-aware MAP at k [2] is computed as

$$\text{MAP-IA@}k = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{N_q}\sum_{z=1}^{N_q}\text{MAP}(k|z,q),$$

where $\text{MAP}(k|z,q)$ is the MAP score for top k returned documents in terms of aspect z of q.

For evaluating accuracy, we use nDCG [10], ERR, Prec@k and MAP. Since users mainly evaluated the top 5 returned results [26], we compute the scores at depth 5 for all metrics. Statistical significance of observed differences between the performance of two runs is tested using a two-tailed paired t-test and is denoted using $^{\blacktriangle}$ (or $^{\blacktriangledown}$) for significant differences for $\alpha = .01$, or $^{\triangle}$ (and $^{\triangledown}$) for $\alpha = .05$.

6.5 Experiments

We report on 6 main experiments aimed at answering the research questions listed in §6.1. Our first experiment aims at understanding whether supervised personalized diversification methods outperform unsupervised ones and whether PSVM$_{div}$ beats the supervised algorithms that apply structured learning technique directly. We compare PSVM$_{div}$ to 2 supervised baselines, SVM$_{div}$ and SVM$_{rank}$, and the 9 unsupervised baselines with both topic relevance and user relevance ground truths, respectively.

To understand the contribution of the user-interest topic model, we conduct our second experiment where we perform comparisons between PSVM$_{div}$ using all features ("token", "interest" and "probability," see §5.3) including those extracted from the topic model and PSVM$_{div}$ using basic features ("token" and "probability" only, see §5.3). In our third experiment, aimed at understanding the effect of our new constraints in PSVM$_{div}$, a series of experiments is conducted by employing different sets of constraints while training.

In order to understand how PSVM$_{div}$ compares to the best baseline, our fourth and fifth experiment provide a query- and subtopic-level analysis, respectively. Finally, to understand the influence of the key parameter in our structured learning framework, C, we train PSVM$_{div}$, SVM$_{div}$ and SVM$_{rank}$ by varying C from 1e-4 to 1.0 and report the performance.

7. RESULTS AND ANALYSIS

The following subsections report, analyze and discuss our experimental results.

7.1 Supervised vs. unsupervised

Table 2 lists the diversity scores of the unsupervised baseline methods. For all metrics in terms of either user relevance or topic relevance, none of the plain methods, viz., BM25, IA-Select, Pers$_{BM25}$, xQuAD and xQuAD$_{BM25}$, beats the best unsupervised personalized diversification methods, viz., PIA-Select, PIA-Select$_{BM25}$, PxQuAD or PxQuAD$_{BM25}$. Moreover, in some cases the performance differences between the best plain method and the best unsupervised personalized diversification method are significant. This indicates that diversity and personalization are complementary and can enhance each other. The same observation can be found in Table 5 where performance is evaluated by relevance-oriented metrics.

Table 2: Performance of unsupervised methods on diversification metrics. The best performance per metric is in boldface. The best plain retrieval method (BM25, IA-Select, Pers$_{BM25}$, xQuAD and xQuAD$_{BM25}$) is underlined. Statistically significant differences between the best performance per metric and the best plain retrieval method are marked in the upper left hand corner of the best performance score.

	User relevance				
	α-nDCG	S-Recall	ERR-IA	Prec-IA	MAP-IA
BM25	.6443	.4557	.2267	.1659	.1245
IA-Select	.6099	.4282	.2241	.1624	.1177
Pers$_{BM25}$	.6427	.4541	.2318	.1639	.1206
xQuAD	.6421	.4635	.2299	.1675	.1267
xQuAD$_{BM25}$	.6270	.4558	.2249	.1646	.1123
PIA-Select	.5766	.4407	.2006	.1480	.1085
PIA-Select$_{BM25}$	.6457	$^{\blacktriangle}$**.4752**	.2364	.1581	.1180
PxQuAD	.6409	.4588	.2313	.1629	.1296
PxQuAD$_{BM25}$	**.6497**	.4713	$^{\triangle}$**.2367**	**.1676**	**.1296**
	Topic relevance				
BM25	.7599	.4456	.2315	.1717	.1241
IA-Select	.7685	.4425	.2365	.1767	.1212
Pers$_{BM25}$	.7746	.4555	.2330	.1794	.1219
xQuAD	.7711	.4600	.2348	.1747	.1245
xQuAD$_{BM25}$	.7763	.4741	.2336	.1773	.1225
PIA-Select	.7410	.4641	.2227	.1650	.1206
PIA-Select$_{BM25}$	$^{\triangle}$**.7854**	**.4798**	$^{\triangle}$**.2415**	.1740	$^{\blacktriangle}$**.1300**
PxQuAD	.7744	.4543	.2350	.1747	.1278
PxQuAD$_{BM25}$	.7827	.4718	.2396	**.1797**	.1245

Table 3 shows the diversity-oriented evaluation results of 3 supervised methods using basic features ("token", and "probability" features, see §5.3) in terms of both ground truths. In terms of diversity-oriented evaluation metrics all of the supervised methods significantly outperform the best unsupervised methods when making comparisons between the scores and the scores of unsupervised methods in Table 2 in most cases. We make further comparisons in Tables 5 and 6 in terms of relevance-oriented metrics, and find that supervised methods can statistically significantly outperform unsupervised ones. These two findings attest to the merits of taking supervised personalized diversification methods for the task of personalized search result diversification.

Next, we compare supervised strategies to each other. Tables 3 and 4 show the diversity-oriented evaluation resutls in terms of both grounth truths. It is clear from both tables that our supervised method PSVM$_{div}$ statistically significantly beats plain supervised methods, SVM$_{rank}$ and SVM$_{div}$. This is because PSVM$_{div}$ considers both personalization and diversity factors, whereas the other two do not take both two factors into account. SVM$_{rank}$ only tries to return more relevant documents, and SVM$_{div}$ directly utilizes standard structured learning for diversification.

As shown in Table 6, in terms of the relevance-oriented metrics, PSVM$_{div}$ does not significantly outperform SVM$_{rank}$ and SVM$_{div}$. This is because PSVM$_{div}$ returns the same number of relevant documents that do, however, cover more subtopics than the other supervised methods. Hence, PSVM$_{div}$ mainly outperforms the other two in terms of diversity-oriented metrics. We provide further analyses in §7.4 (query-level) and §7.5 (subtopic-level).

7.2 Effect of the proposed UIT model

Next, to understand the contribution of our UIT topic model, we compare the performance of the supervised methods using basic

Table 3: Performance of supervised methods utilizing basic features on diversification metrics. The best performance per metric is in boldface. Statistically significant differences between supervised and the best unsupervised method (in Table 2) per metric, between PSVM$_{div}$ and SVM$_{div}$, are marked in the upper left hand corner of the supervised method' score, in the right hand corner of the PSVM$_{div}$ score, respectively.

	User relevance				
	α-nDCG	S-Recall	ERR-IA	Prec-IA	MAP-IA
SVM$_{rank}$	▲.6667	△.4837	.2396	.1683	▲.1856
SVM$_{div}$	▲.6750	△.4887	.2412	△.1698	▲.1974
PSVM$_{div}$	**▲.7234▲**	**▲.5756▲**	**▲.2514▲**	**▲.1702△**	**▲.2037△**
	Topic relevance				
SVM$_{rank}$	.7889	.4805	.2437	△.1812	▲.1848
SVM$_{div}$	▲.8003	△.4893	△.2479	△.1833	▲.2045
PSVM$_{div}$	**▲.8533▲**	**▲.5834▲**	**▲.2649▲**	**▲.1846△**	**▲.2113▲**

Table 4: Performance of supervised methods utilizing all features on diversification metrics. The best performance per metric is in boldface. All the scores here are statistically significant compared to those in Table 2. Statistically significant differences between the method here and the method in Table 3, between PSVM$_{div}$ and SVM$_{div}$, are marked in the upper left hand corner of the corresponding score, in the right hand corner of the PSVM$_{div}$ score, respectively.

	User relevance				
	α-nDCG	S-Recall	ERR-IA	Prec-IA	MAP-IA
SVM$_{rank}$	△.6782	△.4973	.2416	△.1710	▲.2887
SVM$_{div}$	△.6867	△.4973	.2456	△.1729	▲.2911
PSVM$_{div}$	**▲.7513▲**	**▲.6140▲**	**△.2628▲**	**△.1742△**	**▲.2979△**
	Topic relevance				
SVM$_{rank}$	▲.8422	▲.5068	▲.2554	△.1903	▲.3001
SVM$_{div}$	△.8569	▲.5068	▲.2628	▲.1907	▲.3036
PSVM$_{div}$	**▲.9549▲**	**▲.6730▲**	**▲.2849▲**	**△.1917△**	**▲.3096△**

Table 5: Performance of unsupervised methods on relevance metrics. Notational conventions are the same as in Table 2.

	User relevance				Topic relevance			
	nDCG	ERR	Prec	MAP	nDCG	ERR	Prec	MAP
BM25	.5697	.9364	.7113	.2038	.7775	.9440	.9146	.2239
IA-Select	.5126	.9389	.6796	.1813	.7340	.9452	.9250	.2299
Pers$_{BM25}$	.5713	.9276	.7183	.2076	.7741	.9374	.9298	.2316
xQuAD	.5526	.9352	.6858	.1915	.7518	.9367	.9125	.2231
xQuAD$_{BM25}$	.5540	.9133	.6921	.1841	.7605	.9278	.9312	.2281
PIA-Select	.4783	.9034	.6417	.1774	.6709	.9062	.8667	.2043
PIA-Select$_{BM25}$	.5482	.9271	.6687	.1803	.7264	.9418	.9042	.2223
PxQuAD	.5631	.9246	.7050	.2073	.7679	.9435	.9229	.2306
PxQuAD$_{BM25}$	**.5764**	**.9374△**	**.7258△**	**.2145**	**.7793**	**.9466**	**.9396**	**.2355**

features, i.e., all other features but not the features generated from the UIT model, with those using all the features.

We turn to Tables 3 and 4, that list the results of the supervised methods in terms of diversity-oriented metrics when using the basic features and all features, respectively. For all supervised methods, the performance of using all features is better than that of only using the basic features. That is, our proposed UIT model can capture users' interest distributions and this kind of information can be applied to improve performance. Due to space limitations, we do not report the results in terms of relevance-oriented metrics; the findings there are qualitatively similar.

Table 6: Performance of supervised methods utilizing basic features on relevance metrics. The best performance per metric is in boldface. Statistically significant differences between supervised and the best unsupervised method (in Table 5) per metric, between PSVM$_{div}$ and SVM$_{div}$, are marked in the upper left hand corner of the supervised method' score, in the right hand corner of the PSVM$_{div}$ score, respectively.

	User relevance				Topic relevance			
	nDCG	ERR	Prec	MAP	nDCG	ERR	Prec	MAP
SVM$_{rank}$	△.5805	△.9456	△.7345	△.2238	△.7864	.9478	▲.9763	△.2446
SVM$_{div}$	△.5813	△.9467	△.7396	△.2240	△.7858	.9493	▲.9806	△.2482
PSVM$_{div}$	**△.5833**	**△.9485**	**△.7412**	**△.2281**	**△.7922△**	**△.9521**	**▲.9834**	**△.2496**

Table 7: Performance of PSVM$_{div}$ involving different constraints using basic features on diversification metrics with user relevance ground truth. The best performance per metric is in boldface. Statistically significant differences against PSVM$_{div}$-C$_{i}$ are marked in the upper right hand corner of the corresponding scores.

	User relevance				
	α-nDCG	S-Recall	ERR-IA	Prec-IA	MAP-IA
PSVM$_{div}$-C$_{i}$	.6713	.4842	.2403	.1673	.1969
PSVM$_{div}$-C$_{i,ii}$	.6973▲	.5262▲	.2437	.1681	.1977
PSVM$_{div}$-C$_{i,iii}$	.6994▲	.5275▲	.2478△	.1687△	.1983
PSVM$_{div}$-All	**.7234▲**	**.5756▲**	**.2514▲**	**.1702▲**	**.2037△**

Table 8: Performance of PSVM$_{div}$ involving different constraints using all features on diversification metrics with user relevance ground truth. Statistically significant differeneces between the score here and that in Table 7 are marked in the upper left hand corner of the scores. Other notational conventions are the same as in Table 7.

	User relevance				
	α-nDCG	S-Recall	ERR-IA	Prec-IA	MAP-IA
PSVM$_{div}$-C$_{i}$	▲.6843	▲.4965	△.2434	△.1714	▲.2906
PSVM$_{div}$-C$_{i,ii}$	▲.7156▲	▲.5334▲	△.2494△	△.1720△	▲.2932
PSVM$_{div}$-C$_{i,iii}$	▲.7194▲	▲.5388▲	△.2501△	△.1723△	▲.2937△
PSVM$_{div}$-All	**▲.7513▲**	**▲.6140▲**	**△.2628▲**	**△.1742▲**	**▲.2979△**

7.3 Effect of the proposed constraints

Next, to understand the effect of the newly proposed constraints, we conduct experiments by employing different sets of constraints while training. The comparisons are again divided into those using all features and those using basic features. We write PSVM$_{div}$-C$_{i}$, PSVM$_{div}$-C$_{i,ii}$, PSVM$_{div}$-C$_{i,iii}$, and PSVM$_{div}$-All to denote the methods trained with the standard constraint (constraint i in (7)), standard and diversity-biased constraints (constraints i and ii in (7)), standard and interest-biased (constraints i and iii in (7)), and all constraints involved (constraints i, ii and iii in (7)), respectively. Again, we only report results on diversity-oriented metrics.

According to Tables 7 and 8, when employing one more constraint, either diversity-biased or interest-biased, the performance is statistically significantly better than that of only employing the standard constraint. In terms of all metrics, the performance of PSVM$_{div}$ employing all constraints statistically significantly outperforms the performance of using at most two constraints. The positive effect of the proposed constraints again demonstrates that combining diversification (the diversity-biased constraint) and personalization (the interest-biased constraint) boosts the performance.

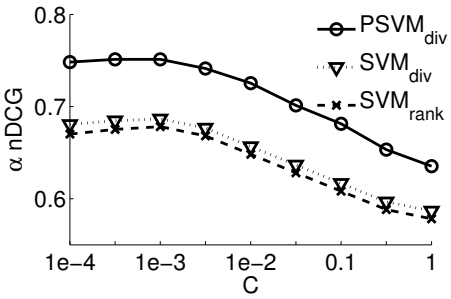

Figure 4: Performance of the supervised methods using all features when varying the value of parameter C.

7.4 Query-level analysis

In order to figure out why PSVM$_{div}$ enhances other supervised baselines, we take a closer look at per test query improvements of PSVM$_{div}$ over the best supervised baseline method, viz., SVM$_{div}$, which outperforms SVM$_{rank}$ in most cases. Fig. 2 shows the per query performance differences in terms of the diversify-oriented metrics of PSVM$_{div}$ against SVM$_{div}$ when they use all the features. PSVM$_{div}$ achieves performance improvements for many queries, especially in terms of α-nDCG, S-Recall, ERR-IA.

In a very small number of cases, PSVM$_{div}$ performs poorer than SVM$_{div}$. This appears to be due to the fact that PSVM$_{div}$ promotes some non-relevant documents when it tries to cover as many subtopics as possible for a given query.

7.5 Subtopic-level analysis

Next, we zoom in on the number of different subtopics that are returned by PSVM$_{div}$ and SVM$_{div}$, respectively, to further analyze why PSVM$_{div}$ beats SVM$_{div}$. Here, again, we use SVM$_{div}$ as a representative. Specifically, we report changes in the number of subtopics for PSVM$_{div}$ against SVM$_{div}$ in Fig. 3 when they use all features. Red bars indicate the number of subtopics that appear in the run of PSVM$_{div}$ but not in the run of SVM$_{div}$, white bars indicate the number of subtopics in both runs, whereas blue bars indicate the number of subtopics that are not in PSVM$_{div}$ but in SVM$_{div}$; queries are ordered first by the size of the red bar, then the size of the white bar, and finally the size of the blue bar.

Clearly, the differences between PSVM$_{div}$ and SVM$_{div}$ in the top 2 and 3 are more limited than the differences in the top 4 and 5, but in all cases PSVM$_{div}$ outperforms SVM$_{div}$. E.g., in total there are 68 more subtopics in the top 5 of the run produced by PSVM$_{div}$ than those in the SVM$_{div}$ run (in terms of all the 180 test queries, 68 subtopics in PSVM$_{div}$ but not in SVM$_{div}$, 7 subtopics in SVM$_{div}$ but not in PSVM$_{div}$).

7.6 Performance of parameter tuning

To understand the performance of the tradeoff parameter C used in (4) and (7), which balances between weights and slacks, we show the performance of PSVM$_{div}$ as well as the 2 supervised baselines using all features. To save space, we only report the performance on α-nDCG. Fig. 4 plots the results and it illustrates that PSVM$_{div}$ performs best when C is small. This indicates the merit of our new constraints (as well as the standard constraint used in the baselines) focusing on weight modification rather than on low training loss.

8. CONCLUSION

Most previous work on personalized diversification of search results produce a ranking using unsupervised methods, either implicitly or explicitly. In this paper, we have adopted a different perspective on the problem, based on structured learning. We propose to boost the diversity and match to users' personal interests of search results by introducing two additional constraints into the standard structured learning framework. We also propose a user-interest topic model to capture users' multinomial distribution of interest over topics and infer per-document multinomial distributions over topics. Based on this a number of user interest features are extracted and the similarity between a user and a document can be effectively measured for our learning method.

Our evaluation shows that supervised personalized diversification approaches outperforms state-of-the-art unsupervised personalization diversification, plain personalization and plain diversification algorithms. The two proposed constraints are shown to play a significant role in the supervised method. We also find that the user-interest topic model helps to improve performance. Our proposed learning method is able to return more subtopics.

As to future work, we aim to study other types of learning strategies for personalized diversification of search results. Our method employed the α-nDCG metric in the loss function; we plan to use other alternative metrics. Finally, our experimental results were only evaluated on a single dataset. In future work we plan to invite users to label the existing datasets, e.g., ClueWeb09, such that they can also be used for personalized diversification algorithms.

Acknowledgments. This research was partially supported by the China Scholarship Council, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 288024 (LiMoSINe) and nr 312827 (VOX-Pol), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, the Center for Creation, Content and Technology (CCCT), the QuaMerdes project funded by the CLARIN-nl program, the TROVe project funded by the CLARIAH program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

9. REFERENCES

[1] Z. Abbassi, V. S. Mirrokni, and M. Thakur. Diversity maximization under matroid constraints. In *KDD*, pages 32–40, 2013.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

[3] K. Bache, D. Newman, and P. Smyth. Text-based measures of document diversity. In *KDD*, pages 23–31, 2013.

[4] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *SIGIR*, pages 185–194, 2012.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] J. Boyd-Graber and D. M. Blei. Syntactic topic models. In *NIPS*, 2008.

[7] J. Boyd-Graber and D. M. Blei. Multilingual topic models for unaligned text. In *UAI*, pages 75–82, 2009.

[8] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, 1998.

[9] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, pages 429–436, 2006.

[10] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and

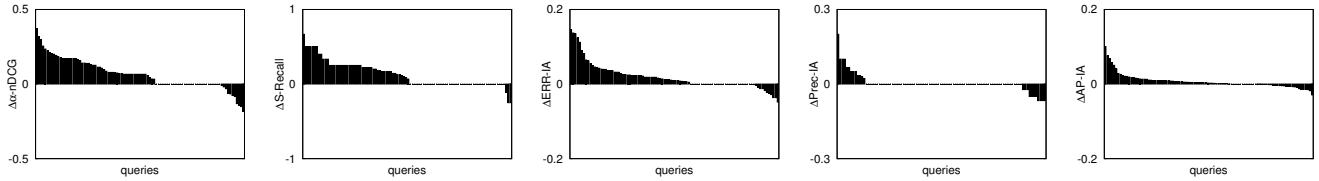

Figure 2: Per query performance differences of PSVM$_{div}$ against SVM$_{div}$. The figures shown are for α-nDCG, S-Recall, ERR-IA, Prec-IA and MAP-IA, respectively. A bar extending above the center of a plot indicates that PSVM$_{div}$ outperforms SVM$_{div}$, and vice versa for bars extending below the center. Note that figures are not in the same scale.

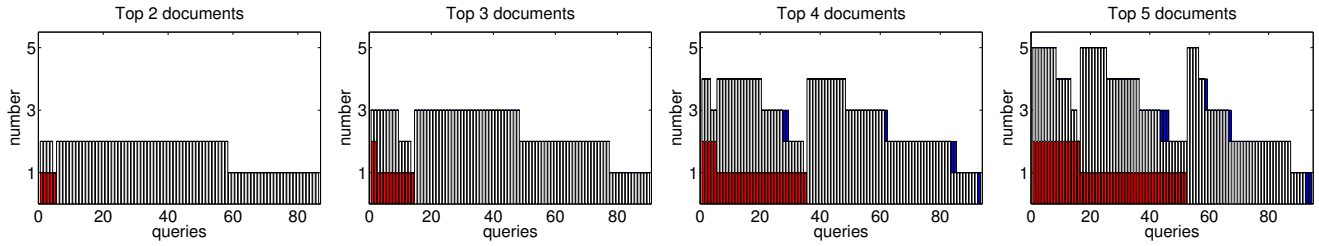

Figure 3: How runs produced by PSVM$_{div}$ and SVM$_{div}$ differ. Red, white, blue bars indicate the number of different subtopics that appear in PSVM$_{div}$ but not in SVM$_{div}$, in both runs and not in PSVM$_{div}$ but in SVM$_{div}$, respectively, at corresponding depth k (for k=2, 3, 4, 5). Figures should be viewed in color.

diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.

[11] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 web track. In *TREC*, pages 1–8, 2012.

[12] V. Dang and W. B. Croft. Diversity by proportionality: An election-based approach to search result diversification. In *SIGIR*, pages 65–74, 2012.

[13] V. Dang and W. B. Croft. Term level search result diversification. In *SIGIR*, pages 603–612, 2013.

[14] S. Jameel and W. Lam. An unsupervised topic segmentation model incorporating word order. In *SIGIR*, pages 203–212, 2013.

[15] S. Khuller, A. Moss, and J. S. Naor. The budgeted maximum coverage problem. *Inf. Proc. Lett.*, 70(1):39–45, 1999.

[16] S. Liang, Z. Ren, and M. de Rijke. Fusion helps diversification. In *SIGIR*, 2014.

[17] C. Liu, N. J. Belkin, and M. J. Cole. Personalization of search results using interaction behaviors in search sessions. In *SIGIR*, 2012.

[18] J. S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, 89(427):958–966, 1994.

[19] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *SIGIR*, pages 691–692. ACM, 2006.

[20] S. E. Robertson and D. A. Hull. The TREC-9 filtering track final report. In *TREC*, pages 25–40, 2000.

[21] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *WWW*, pages 881–890, 2010.

[22] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *CIKM*, pages 824–831. ACM, 2005.

[23] Y. Shi, X. Zhao, J. Wang, M. Larson, and A. Hanjalic. Adaptive diversification of recommendation results via latent factor portfolio. In *SIGIR*, pages 175–184. ACM, 2012.

[24] J. Tang, M. Zhang, and Q. Mei. One theme in all views: Modeling consensus topics in multiple contexts. In *KDD*, pages 5–13, 2013.

[25] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, 2005.

[26] D. Vallet and P. Castells. Personalized diversification of search results. In *SIGIR'12*, 2012.

[27] D. Vallet, I. Cantador, and J. M. Jose. Personalizing web search with folksonomy-based user and document profiles. In *ECIR*, pages 420–431. Springer, 2010.

[28] S. Vargas, P. Castells, and D. Vallet. Explicit relevance models in intent-oriented information retrieval diversification. In *SIGIR*, pages 75–84, 2012.

[29] H. Wang, X. He, M.-W. Chang, Y. Song, R. W. White, and W. Chu. Personalized ranking model adaptation for web search. In *SIGIR*, pages 323–332, 2013.

[30] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *KDD*, pages 424–433, 2006.

[31] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: Discover spatio-temporal topics for twitter users. In *KDD*, pages 605–613, 2013.

[32] Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *ICML*, pages 1224–1231. ACM, 2008.

[33] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *SIGIR*, pages 271–278. ACM, 2007.

[34] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, pages 334–342, 2001.

[35] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pages 10–17, 2003.

[36] S.-X. Zhang and M. Gales. Structured SVMs for automatic speech recognition. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 21(3):544–555, 2013.

[37] J. Zhu, X. Zheng, L. Zhou, and B. Zhang. Scalable inference in max-margin topic models. In *KDD*, pages 964–972, 2013.