Personalized and Diversified: Ranking Search Results in an Integrated Way

SHUTING WANG, ZHICHENG DOU, and JIONGNAN LIU, Gaoling School of Artificial Intelligence, Renmin University of China, China
QIANNAN ZHU, School of Artificial Intelligence, Beijing Normal University, China
JI-RONG WEN, Gaoling School of Artificial Intelligence, Renmin University of China, China, Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, China, and Beijing Key Laboratory of Big Data Management and Analysis Methods, China

Ambiguity in queries is a common problem in information retrieval. There are currently two solutions: search result personalization and diversification. The former aims to tailor results for different users based on their preferences, but the limitations are redundant results and incomplete capture of user intents. The goal of the latter is to return results that cover as many aspects related to the query as possible. It improves diversity yet loses personality and cannot return the exact results the user wants. Intuitively, such two solutions can complement each other and bring more satisfactory reranking results. In this article, we propose a novel framework, namely, **PnD**, to integrate personalization and diversification reasonably. We employ the degree of refinding to determine the weight of personalization dynamically. Moreover, to improve the diversity and relevance of reranked results simultaneously, we design a reset RNN structure (RRNN) with the "reset gate" to measure the influence of the newly selected document on novelty. Besides, we devise a "subtopic learning layer" to learn the virtual subtopics, which can yield fine-grained representations of queries, documents, and user profiles. Experimental results illustrate that our model can significantly outperform existing search result personalization and diversification methods.

CCS Concepts: • **Information systems** → **Personalization**; **Information retrieval diversity;**

Additional Key Words and Phrases: Personalized search, search result diversification, integration

Authors' addresses: S. Wang, Z. Dou (corresponding author), and J. Liu, Gaoling School of Artificial Intelligence, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing, 100872, China; e-mails: {wangshuting, dou, liujn}@ruc.edu.cn; Q. Zhu, School of Artificial Intelligence, Beijing Normal University, No. 19, Xinjiekouwai St, Haidian District, Beijing, 100875, China; e-mail: zhuqiannan@bnu.edu.cn; J.-R. Wen, Gaoling School of Artificial Intelligence, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing, 100872, China and Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, Beijing, China and Beijing Key Laboratory of Big Data Management and Analysis Methods, No. 59 Zhongguancun Street, Haidian District, Beijing, China; e-mail: jrwen@ruc.edu.cn.

1046-8188/2024/01-ART81 $15.00
https://doi.org/10.1145/3631989

ACM Reference format:
Shuting Wang, Zhicheng Dou, Jiongnan Liu, Qiannan Zhu, and Ji-Rong Wen. 2024. Personalized and Diversified: Ranking Search Results in an Integrated Way. *ACM Trans. Inf. Syst.* 42, 3, Article 81 (January 2024), 25 pages.
https://doi.org/10.1145/3631989

1 INTRODUCTION

As previous studies showed [12, 33], search queries are usually short and ambiguous. For example, when the user issues "Apple," she may want to learn information about fruit apples or the Apple company. Personalized search [2, 15, 17, 24–26, 41, 48, 50–52] and search result diversification [1, 6, 13, 14, 20, 21, 23, 28, 31, 34, 45, 46, 49] are two mainstream methods to eliminate ambiguity and provide more satisfactory search results.

Personalized search tries to capture the profile of users from their search history. In this way, it can mine out the documents that the user is most likely to be interested in and provide personalized results for different users. As shown in Figure 1, the personalized IR system will rerank programming-related documents ahead for a programmer when she issues "JAVA." However, the drawback is that it may return redundant results, e.g., the first two documents are both related to "JAVA program language tutorial," while the user may prefer other documents covering her interests after viewing the first document. It also reveals that personalization cannot capture the intrinsic diversity of user preferences. In contrast, search result diversification aims to diminish ambiguity by improving the probability of satisfying the intents of different users. Specifically, it reduces the similarity between result documents while keeping their relevance to the current query. As Figure 1 illustrates, the diversified system returns documents related to "JAVA language" or "JAVA island" in the top-ranked results. This result has high diversity, yet, if we can infer that the user is a programmer from her search log, then the documents related to "JAVA island" may not be what she desires. Thus, the limitation of search result diversification is that it neglects the user's historical information and cannot provide exact results that meet the user's personalized preferences. Factually, personalization and diversification have their own advantages and disadvantages in different scenarios, they should be integrated to capture the strengths of both while complementing their weaknesses.

Recently, there have been some works [8, 22, 30, 36] on the integration of diversification and personalization. However, most of the efforts simply introduced a user variable to the existing diversification function without an in-depth exploration of the relationship between the two methods. Different from existing works, we think that the proportion of personalization and diversification should not be fixed but dynamic in different situations. Correspondingly, there is a challenge in how to measure the weight properly for controlling personalization and diversification dynamically. As previous works have studied [15, 51], people usually find information that has been searched before, which is called refinding behavior. Intuitively, the refinding degree is a reasonable factor to balance their importance. A high refinding degree means that the user uploads a query similar to the ones she issued before, hence, personalization is critical to providing accurate user interests in such a case. Otherwise, we should focus on diversification, as less favorable information can be extracted from her history for the issued query. Besides, as mentioned before, the user's preferences are varied, so the diversity of user interests should be considered in personalization. By integrating personalization and diversification, the IR system can consider the personalized relevance and diversity of user interests simultaneously. Thus, it provides more satisfactory results. For the same example shown in Figure 1, with a document about "JAVA program language tutorial" being ranked first, the integrated system ranks other programming-related documents that the programmer prefers to click on, e.g., "JAVA program language download," at the top.

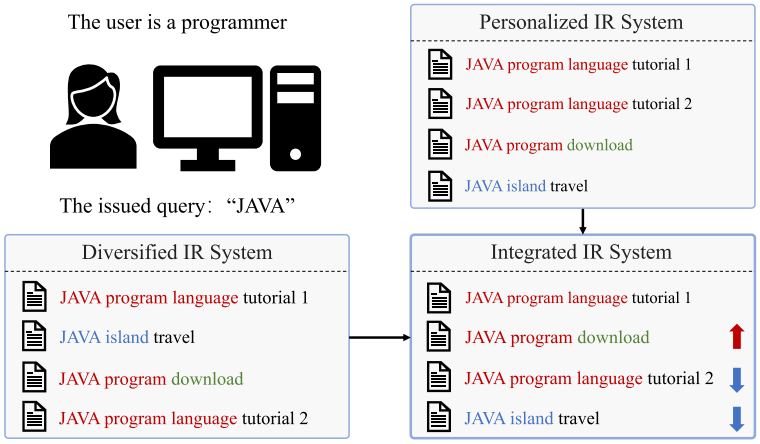

Fig. 1. An example to demonstrate the necessity of integrating personalization and diversification.

To achieve it, we propose a novel framework, namely, **PnD**, to consider personalization and diversification simultaneously in search result ranking. In summary, PnD greedily selects the local optimal document based on a final score to rerank the results. The final score is produced by aggregating the signals that measure diversity, personality, and ad hoc relevance. Specifically, after generating the user's profiles from her historical search, we first extract virtual subtopics of the current query, candidate documents, and profiles to explicitly model their fine-grained representations. Second, we use the similarity of the query to the user's profiles as the refinding degrees to dynamically determine the importance of personalization. Third, for modeling the diversity of the candidate document relative to previously selected documents, we need to measure the subtopic coverage of selected documents following existing diversification methods [1, 21, 31]. Therefore, we develop a modified RNN structure to update the representations of the current query and the user's profiles based on the selected documents. This process enables the representations of the current query and user profiles to forget the covered information of selected documents. Thus, we can view the similarity of the updated representations with the candidate representation as three novelty scores that measure the candidate's novelty, which includes relevance and diversity, from personalized and general (current query) perspectives. Finally, we assign the refinding degrees to personalized novelty scores and aggregate all novelty scores as the final score for reranking. As we analyzed above, the user's interests are varied, which leads to complex behaviors. Thus, we think that the users' click feedback reflects relevance and diversity simultaneously. Correspondingly, we optimize our model based on the users' click feedback and aim to rerank clicked documents, i.e., satisfactory documents, at the top.

The main contributions of our work are as follows:

(1) We propose a novel framework that dynamically integrates personalization and diversification based on refinding behavior to provide more satisfactory results;
(2) To apply our models atop query logs, we learn virtual subtopics of profiles and queries, which can model users' and queries' various needs explicitly and prompt the performance of our model;
(3) We design a modified RNN structure for modeling the coverage of selected documents on the subtopics of the query and user's profiles, based on which we can measure the scores of remaining documents more reliably.

The rest of this article is organized as follows: Related works are introduced in Section 2. In Section 3, we explain the structure of the proposed model. We demonstrate the experiment settings in Section 4. Then, we present the experiment results and analysis in Section 5. We conclude the whole article and propose further improvements in Section 6.

2 RELATED WORK

2.1 Personalized Search

Personalized search attracts much attention due to its ability to tailor the ranking results based on the user's personality, thereby improving ranking quality and user satisfaction.

2.1.1 Traditional Methods. Traditional personalized search models are almost unsupervised [15, 35] and based on manual features [3, 38, 40]. For example, P-Click[15], proposed by Dou et al., is a typical method based on click data. In the view of the authors, when the user issues a previously searched query, the documents that were frequently clicked are more important than the rarely clicked ones. Considering the importance of topic features for personalizing search results, some works construct user profiles in explicit or implicit topic space. Chirita et al. [9] utilized the **Open Directory Project (ODP)**, which is the directory of artificially annotated topics, to categorize the documents and capture user references. Then, results are re-ranked based on the distance between user topic nodes and document topic nodes. Owing to the coverage limits of the manual topics, Thanh et al. [39] employed **Latent Dirichlet Allocation (LDA)** [4] to extract topic-based features and construct user profiles in the latent topic space. Benefiting from the advanced learning to rank model LambdaMART [5], significant improvement has been achieved by several works [38, 43].

There are some disadvantages to using heuristic rules and features, such as their time-consuming nature and limited coverage. With the development of machine learning, many supervised methods have been proposed and shown superior performance.

2.1.2 Machine Learning-based Methods. SLTB [2] is a feature-based model that aggregates the multiple manual features to rerank the results via a learning-to-rank method. Considering the importance of the user's search history to model the user's current search intent, many researchers have explored such valuable information to solve various tasks, e.g., document retrieval [32], session-based ranking algorithms [18, 29, 42], and so on. In the field of personalized search, Ge et al. proposed HRNN [17] to construct users' long-term and short-term profiles based on a hierarchical RNN structure from their search logs. The authors further introduced the attention technique to learn the importance of historical sessions relative to the current query to gain a reliable representation of the user's long-term profile. To mine out the high-quality negative samples for improving the ranking ability of the trained personalized model, Lu et al. employed **Generative Adversarial Networks (GAN)** for personalized search and proposed PSGAN [25]. Considering the impact of time span on personalization, Ma et al. proposed PSTIM [26] to mine out more accurate personalized intents. Yao et al. devised PEPS [48] to eliminate ambiguity in the stage of word embedding. It learned for each user an individual word embedding matrix that only contains the aspects that the user is interested in. Instead of learning user profiles that may introduce historical noise, Zhou et al. [50] proposed to learn a disambiguation representation of the current query directly. Considering the user histories provide rich contextual information for disambiguating the issued query, the authors applied the transformer encoder [37] to capture more specific user intents and provide satisfactory ranking results. Dou et al. have proven the effectiveness of re-finding behaviors for personalized search. For further utilization of this information, RPMN [51] is designed to identify more complex and potential refinding behaviors. It constructed memory networks from three perspectives to support the refinding of these aspects.

2.2 Search Result Diversification

Instead of capturing the specific intents of the user, search result diversification aims to meet different users' intents by providing search results covering various aspects. According to whether subtopic information is applied, search result diversification approaches can be categorized into explicit and implicit ones. According to whether the algorithms are learnable, they can be divided into unsupervised and supervised approaches.

2.2.1 Unsupervised Methods. Search result diversification has to consider the dissimilarity between the next document and selected documents, thus, most algorithms greedily choose the current optimal document. MMR [6] is a typical implicit diversification method that greedily selects the next optimal document and constructs the reranked results. It measures the relevance of the candidate documents to the query and views the similarity of the candidate to the previously ranked documents as diversity. Then, MMR linearly combines relevance and diversity using a tradeoff factor to generate the ranking score. To consider subtopics explicitly, IA-select [1] and xQuAD [31] are proposed. Rather than calculating the diversity implicitly by document similarity, they regard the subtopic coverage as the diversity indicator, hence explicitly improving the diversity of the ranking results. Dang et al. [13] further introduced term-level subtopics to promote the efficiency of the algorithm. Ozdemiray et al. [27] proposed to rerank the results for each aspect of the query, then conducted a ranking aggregation to produce the final diversified result. Wu et al. [44] designed a fusion-based algorithm to integrate multiple IR search system results. In fact, the query subtopics are usually hierarchical rather than an equal list. Therefore, Hu et al. [20] applied hierarchical intents to calculate diversity and provide ranking results more accurately.

2.2.2 Supervised Methods. Different from unsupervised approaches that use heuristic rules for diversification, supervised approaches introduce learnable parameters to construct a more effective algorithm. R-LTR [53] applies trainable weight to yield relevance and diversity scores. Xia et al. devised a novel loss function to escalate the gap between the ideal ranking and negative rankings, namely, PAMM [45]. As the manual novelty features are limited, they improved R-LTR and PAMM into PAMM-NTN and R-LTR-NTN [46] by introducing a **neural tensor network (NTN)** to model the novelty. As previous supervised methods are implicit without considering subtopics explicitly, Jiang et al. proposed an explicit supervised approach, DSSA [21]. It builds the representation of the selected document sequence using an RNN module. The importance of subtopics is weighted based on the similarity between the subtopics and the document sequence. Liu et al. [23] proposed DVGAN, which conducts diversification based on GAN. Greedily producing ranking lists is the common paradigm of preceding methods, while it inevitably yields sub-optimal results rather than global optimal ones. To tackle this problem, Qin et al. proposed DESA [28]. It applies transformers to measure diversity and produce the ranking score of all candidate documents simultaneously.

2.3 Personalized Search Result Diversification

The goals of search result personalization and diversification seem to be completely opposed. The former tries to find the most satisfying aspect for the current user, and the latter aims to cover as many aspects as possible to meet the diverse intents of all users. However, some studies recently found that they are not opposed but can be combined to provide better results. Radlinski et al. [30] conducted query-query reformulation to generate $R(q)$, the set of queries related to the current query q but different from each other. The personalized diversity result is generated from the original results of queries in $R(q)$. Valle [36] improved traditional diversification methods by introducing a user variable u to score functions. Chen et al. [8] conducted personalized diversity

based on personalized query suggestion diversification. Liang et al. [22] applied a structured SVM model to implement personalized diversification.

Different from previous works that directly inherit the paradigm of diversification, we design our framework based on deep analysis. As users' behaviors are complex and diverse, we believe the clicks on the documents reflect the user's satisfaction, including relevance and diversity. Thus, the main target of our model is improving user satisfaction. As we mentioned above, the manual features are time-consuming and limited, and the deep learning-based model can make up for it by learning high-dimensional features automatically. Thus, we devise our model based on deep learning.

3 OUR PROPOSED METHOD

Most of the existing methods that are applied to disambiguate the search query focus on the two mainstream approaches, personalization, and diversification, while the former will lead to results redundancy and the latter cannot capture accurate user intents. Different from them, we proposed a model that integrates personalization and diversification to make their advantages and disadvantages complementary and return search results that fully meet the user's information needs.

3.1 Problem Definition

In ad hoc search, the engine learns document relevance to the query and returns a candidate document list $\mathcal{D} = \{d_1, d_2, \ldots\}$ when the user u enters a query q. Different from the traditional ad hoc search, our model focuses on integrating personalization and diversification and aims to iteratively find the optimal document from Equation (1):

$$d^{t,*} = \arg\max_{d \in \mathcal{D}_t} P(d|q, \mathcal{D}_t, \mathcal{S}_t, u). \tag{1}$$

Note that our model greedily selects local-optimal documents to build ranking results. For the tth ranking position, there are $t-1$ previously selected documents, which are denoted as $\mathcal{S}_t = \{d^{i,*}|i \in [1, t-1]\}$, and the remained candidate documents construct a tth step candidate document set, $\mathcal{D}_t, \mathcal{D}_t \cup \mathcal{S}_t = \mathcal{D}$. We denote the local optimal document at tth ranking position as $d^{t,*}$, where $d^{t,*} \in \mathcal{D}_t$. After selecting a document in the tth step, the model will consider its impact on measuring the novelty of remaining documents by updating the representation of the current query and the user profile, respectively.

Usually, there are numerous users in the real world, and they have different search intents. To incorporate the user's personality, we apply the long-term and short-term history of the user to learn her profiles. User behavior over a period of time usually has similar search intents, so this period of time is called a session. We first recognize the user's historical behaviors in the current session as her short-term history, i.e., $H_s = \{(q_i, D_i)|i \in [1, m]\}$, where m denotes the number of issued queries in the current session, q_i denotes the ith query, and D_i denotes the sets of corresponding clicked documents. The long-term history $H_l = \{(q_i, D_i)|i \in [1, n]\}$ contains the historical behaviors before the current session, where n denotes the number of issued queries.

Under the initial representation of the current query q^0, our model learns the short- and long-term profiles S^0, L^0 from H_s and H_l, which represent the user's initial profiles. We apply subtopic learning layers to extract the virtual subtopic lists of the query, user profiles, and candidate documents for producing their fine-grained representations, i.e., $\hat{q}^0$, $\hat{S}^0$, $\hat{L}^0$, and $\hat{d}^0$.

After selecting a document in the tth step, the model will measure its coverage across the diverse personalized interests and query aspects by updating the representations of the user profiles and the current query, respectively. Specifically, suppose that we have selected the tth documents, the model will update $\hat{q}^{t-1}$, $\hat{S}^{t-1}$, $\hat{L}^{t-1}$ to $\hat{q}^t$, $\hat{S}^t$, $\hat{L}^t$. In such a way, $\hat{q}^t$, $\hat{L}^t$, and $\hat{S}^t$ denote the uncovered

Table 1. Notations Used in Our Framework

Notation	Explanation
u	Current user.
q	The entered query.
$\mathcal{D}$	Set of all candidate documents.
$\mathcal{S}_t$	Set of selected candidate documents.
$\mathcal{D}_t$	Set of remaining candidate documents.
d, $\hat{d}$	Representation of document d, $\hat{d}$ considers virtual subtopics.
$\hat{d}^{t,*}$	Representation of the selected document at tth step considering virtual subtopics.
q^0	Initial representations of query.
S^0	Initial representations of the short-term profile.
L^0	Initial representations of the long-term profile.
$\hat{q}^0$	Initial representations of the query considering virtual intents.
$\hat{S}^0$	Initial representations of the short-term profile considering virtual intents.
$\hat{L}^0$	Initial representations of long-term profile considering virtual intents.
$\hat{q}$, $\hat{S}$, $\hat{L}$	Representations at the tth step, based on virtual subtopics.

intents of the query and profiles after selecting t documents and will be used in the next steps for ranking the remaining documents. In the rest of this article, we will abbreviate $\hat{q}^t$, $\hat{S}^t$, $\hat{L}^t$ to $\hat{q}$, $\hat{S}$, $\hat{L}$ to save space, and $\hat{q}^0$, $\hat{S}^0$, $\hat{L}^0$ refer to the initial representations of the query and profiles. Since our model considers the diversity of user preferences, **we will call the personalization-related part "diversified personalization" to make the content easier to understand. Accordingly, we refer to the query-related part as "general diversification."** The score function in Equation (1) can be transformed as:

$$P(d|q, \mathcal{D}_t, \mathcal{S}_t, u) = \zeta\left(f_g(d, q), r^S f_S(d, S), r^L f_L(d, L), \text{rel}(q, d)\right), \tag{2}$$

where ζ denotes an MLP, which is used to aggregate different signals and produce the final score of d. $f_g(d, q)$ denotes the general novelty score, which is produced by the general diversification part without considering personality. Our model uses it to measure general diversity while maintaining ad hoc relevance. The diversified personalization parts yield two signals, $f_S(d, S)$ and $f_L(d, L)$, which consider the candidate document's relevance to the user's short- and long-term interests that have not been covered by selected documents. Thus, we call them personalized novelty scores. r^L and r^S represent the similarity between the query and the profiles, i.e., the degree of refinding. They are used to adjust the weights of personalization-related components. $\text{rel}(q, d)$ denotes the additional relevance of document d to query q, which is produced by aggregating some relevance features. We show our notations in Table 1. The content of our model will be introduced in detail in the following subsections.

3.2 Overview of Our Model

Our model is proposed to reasonably integrate personalization and diversification. When the user u issues the query q, it reranks the documents she wants at the top. As described in the problem definition, we achieve it by following these steps:

(1) **Virtual subtopics learning**. We design a "subtopic learning layer" to extract virtual subtopics from the current query, user profiles, and documents. The lists of subtopics are viewed as their fine-grained representations and used in subsequent steps.
(2) **Calculation of refinding degree**, which measures the refinding degrees by the similarity between the current query and the user's profiles. The refinding degrees are used to weigh the personalization part of our model. Steps (1) and (2) construct the "weighting process" of Figure 2.

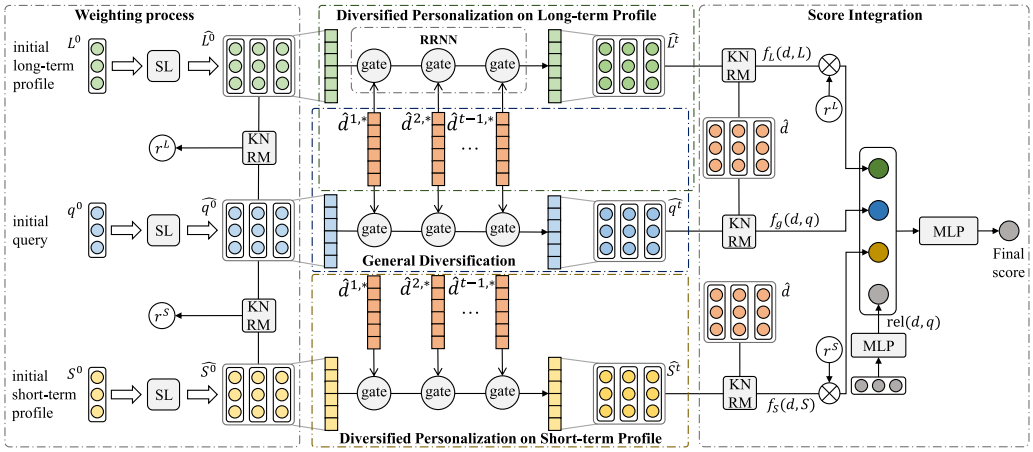

Fig. 2. The architecture of PnD. The gray circle with "gate" represents the "reset" gate introduced in Section 3.5. The gray square with "SL" denotes the subtopic learning layer introduced in Section 3.3. Note that the representations of selected document $\hat{d}^{i,*}$ and candidate document $\hat{d}$ are generated in the same process by the SL structure, but we have no space to draw them explicitly.

(3) **Diversity modeling**, which updates the representation of the query, profiles in the greedy document selection to forget the aspects covered by the selected documents. These updated representations contain aspects that have not been covered by selected documents, and their similarity with remaining candidate documents can measure the novelty of candidates to produce reliable ranking scores.
(4) **Aggregated result scoring**, which calculates and aggregates a general novelty score, two personalized novelty scores, and an ad hoc relevance score to select the local optimal document. We repeat Steps (3) and (4) until all candidate documents are reranked.

The overall model structure of PnD is shown in Figure 2. We will explain each component in detail as follows.

3.3 Virtual Subtopics Learning

As the user interests and query aspects are various, explicitly applying the subtopics is conducive to generating their fine-grained representations, which will introduce richer information. However, the dataset containing users' click feedback always lacks subtopic information. Furthermore, manual subtopics are also time-consuming and narrow in coverage. To solve this problem, we devise a **"subtopic learning layer" (SL)** to capture virtual subtopics of queries, documents, and user profiles. We call them virtual subtopics, as we do not know their exact meaning.

3.3.1 Modeling the Virtual Subtopics. Since the initial representations of the short-term profile and long-term profile, S^0 and L^0, are produced from the user's search log and the current query via a hierarchical transformer [50], we deem that they contain all the interests of the user about the query. To capture the fine-grained user interests in specific aspects, we apply a learnable matrix with nonlinear activation to map the original representation to a specific space, and the obtained vector represents a subtopic that contains the user interest in this aspect. We construct c learnable matrices $M_i, i \in \{1, \ldots, c\}$ to decode virtual subtopics of the user's profiles from c aspects. In this way, the representation of the profile is represented as a $c \times h$ matrix, where h denotes the

dimension of the subtopic vector:

$$\hat{S}^0 = \left[\hat{S_1}^0, \hat{S_2}^0, \ldots, \hat{S_c}^0\right], \quad \hat{S_i} = \phi(M_{S,i}^T S^0), \quad \hat{S}^0 \in \mathbb{R}^{c \times h}, \quad \hat{S_i} \in \mathbb{R}^{1 \times h}, \tag{3}$$

$$\hat{L}^0 = \left[\hat{L_1}^0, \hat{L_2}^0, \ldots, \hat{L_c}^0\right], \quad \hat{L_i} = \phi(M_{L,i}^T L^0), \quad \hat{S}^0 \in \mathbb{R}^{c \times h}, \quad \hat{S_i} \in \mathbb{R}^{1 \times h}, \tag{4}$$

where ϕ denotes the activation layer; we apply LeakyReLU in this article. Note that we build subtopic layers with different parameters for short- and long-term profiles, as they capture the user's preferences from different perspectives. We also apply a different subtopic learning layer to yield $\hat{q}^0$ and $\hat{d}$, as the information they contain is word-level.

3.3.2 Guaranteeing Diversity of Virtual Subtopics. To prevent the learning of homogeneous virtual subtopics, thus affecting the coverage of subtopics, we further construct an auxiliary task to ensure the diversity of the produced virtual subtopics. Inspired by Reference [54], we calculate the **intra-list similarity (ILS)** of the given subtopic set to represent its diversity. Take the query subtopics $[\hat{q}_1^0, \hat{q}_2^0, \ldots, \hat{q}_c^0]$ as an example. We calculate the similarity of every pair of virtual subtopics and normalized it as below:

$$\text{ILS}(\hat{q}^0) = \frac{2\sum_{i=1}^{c}\sum_{j=i+1}^{c}\text{sim}(\hat{q}_i^0, \hat{q}_j^0)}{c(c-1)}. \tag{5}$$

sim() denotes the similarity function; we initialize it by cosine similarity. For the virtual subtopics of user profiles and candidate documents, we employ the same function to calculate the corresponding intra-list similarity. Thus, we devise the loss function to promote the diversity of virtual subtopics as follows:

$$\mathcal{L}_{div} = \text{ILS}(\hat{q}^0) + \text{ILS}(\hat{L}^0) + \text{ILS}(\hat{S}^0) + \text{ILS}(\hat{d}^0). \tag{6}$$

This loss function is utilized to construct the final loss function for optimizing our model, which will be introduced in Section 3.7.

3.4 Calculation of Refinding Degree

As we mentioned above, the proportion of personalization and diversification is not fixed, but dynamically based on the refinding degree. It denotes the degree to which the current query is related to the user's search history, thus, we can view the similarity between the user's profiles and the query as the refinding degree. The high similarity means that the user has searched for relevant queries before, so we should assign a high weight to diversified personalization. Otherwise, we should focus on diversification, since the search history can not provide favorable information. Then, we calculate the similarity between the current query and the profiles. r^S denotes the similarity between $\hat{q}^0$ and $\hat{S}^0$, and r^L denotes the similarity between $\hat{q}^0$ and $\hat{L}^0$. As the new representations of the query and profiles are lists of subtopics, we should consider the matching information between each subtopic pair. Therefore, we can capture the fine-grained similarity score between the query and profiles. Inspired by KNRM [47], we construct a similarity function to calculate the matching score between two lists of subtopics. We first build a cross-matching matrix M^S of $\hat{q}^0$ and $\hat{S}^0$, i.e.,

$$M_{ij}^S = \cos(\hat{q}_i^0, \hat{S}_j^0); \tag{7}$$

$$\cos(x, y) = \frac{< x, y >}{\|x\|_2 \|y\|_2}. \tag{8}$$

It can capture the rich matching information between the query and short-term profile at the subtopic level. Then, we apply k RBF kernels to integrate the matching score from k aspects:

$$f_o(M^S) = \sum_i \log\left(\sum_j \exp\left(-\frac{(M_{ij}^S - u_o)^2}{2\sigma_o^2}\right)\right), \quad o \in \{1, \ldots, k\}, \tag{9}$$

where k, u_o, and σ_o are hyper-parameters. Following existing work, the value of k is set to 11 in our model, u_o is evenly distributed in $(-1, 1)$ based on k and σ_o is 0.1 in our framework.

Finally, we aggregate the k matching scores to yield r^S by an MLP layer ϕ with tanh as activation:

$$r^S = \phi\left(f_1(M^S), \ldots, f_o(M^S), \ldots, f_k(M^S)\right). \tag{10}$$

We use F_k to represent the steps above and r^L is calculated by the same function:

$$r^S = F_k(\hat{q}^0, \hat{S}^0), \quad r^L = F_k(\hat{q}^0, \hat{L}^0). \tag{11}$$

Note that the refinding degrees, r^S and r^L, are generated based on the original representations of the query and user's profiles, $\hat{q}^0$, $\hat{S}^0$, and $\hat{L}^0$. It means that the importance of the personalization parts will not change in the document selection process.

3.5 Diversity Modelling

Suppose that $\hat{S}^0$ and $\hat{L}^0$ represent the initial short- and long-term profiles, which contain multiple aspects that user u prefers, and $\hat{q}^0$ represents the original subtopics that query q contains. To enhance the novelty of the results, we hope our model can choose the local optimal document based on the subtopics not covered by previously selected documents. Thus, the query should forget the information related to the selected documents. Inspired by the gate mechanism from LSTM [19]. It uses the gate mechanism to forget and learn information based on the input state. In our study, we construct a "reset gate" to update $\hat{q}^{t-1}$, which can model the subtopic coverage of the newly selected document and enable $\hat{q}^{t-1}$ to forget this information by a reset vector, thus the updated representation $\hat{q}^t$ only contains the aspects that have not been covered by selected documents. Note that in the process of the reset gate, all the lists of subtopics are flattened as a $ch-$dimension vector.

$$\hat{q}^t = \text{gate}(\hat{q}^{t-1}, \hat{d}^{t,*}) = r \otimes \hat{q}^{t-1}; \tag{12}$$

$$r = f_2(W_2^T f_1(W_1^T[\hat{q}^{t-1}; \hat{d}^{t,*}] + b_1) + b_2), \tag{13}$$

where $\hat{d}^{t,*} \in \mathbb{R}^{ch}$ represents the tth selected document. $\otimes$ denotes the Hadamard product. $r \in (-1, 1)^{ch}$ represents the reset vector to update the last hidden state. We apply a learnable linear layer with W_1, b_1 as parameters to learn the coverage information of $\hat{d}^{t,*}$ on $\hat{q}^{t-1}$, and the second layer with W_2, b_2 is used to map the information to the reset vector. f_1 and f_2 denote activation functions. This process can capture the nonlinear interaction between the document and the query. Thus, it can learn the coverage on subtopics of the query and update the hidden state.

As we mentioned above, we should consider the user's diverse preferences to avoid redundant personalized ranking results and satisfy the comprehensive intents of the user. To achieve it, we build the same structures with different parameters to capture the impact of selected documents on the user's personalized interests and update the representation of $\hat{S}$ and $\hat{L}$ iteratively.

We call the RNN structure based on the reset gate "RRNN." Through it, we will obtain the latest representations of the query, long- and short-term profiles. They contain the remaining aspects after selecting t documents. The update processes of $\hat{L}$, $\hat{q}$, and $\hat{S}$ correspond to the parts of "Diversified Personalization on Long-term Profile," "General Diversification," and "Diversified Personalization on Short-term Profile" in Figure 2.

3.6 Aggregated Result Scoring

As we explained in the previous chapter, the latest representations of $\hat{q}$, $\hat{S}$, and $\hat{L}$ only contain the aspects that have not been covered by selected documents. Thus, the matching scores of the candidate to them can be viewed as the degree to which the user's or query's remaining intents are satisfied. We call them novelty scores, as they measure the relevance and diversity of the candidate simultaneously. We apply similar functions with Equation (11) to produce the general novelty score $f_f(d, q)$, the long-term personalized novelty score $f_L(d, L)$, and the short-term personalized novelty score $f_S(d, S)$ as below:

$$f_g(d, q) = F_k(\hat{d}, \hat{q});\ f_L(d, L) = F_k(\hat{d}, \hat{L});\ f_S(d, S) = F_k(\hat{d}, \hat{S}). \tag{14}$$

Note that we omit $t-1$ in $\hat{q}$, $\hat{S}$, and $\hat{L}$ for simplification. We further introduce an additional component to measure the ad hoc relevance of d, $\text{rel}(d, q)$. More specifically, $\text{rel}(d, q) = \phi(f(d, q))$, where $\phi(\cdot)$ denotes an MLP layer, and $f(d, q)$ denotes the relevance features between document d and query q. We use the same feature set as Reference [51].

Finally, the novelty scores and the ad hoc relevance score can be integrated to produce the final score of d. To control the importance of the personalization part, we multiply the personalized novelty scores by refinding degrees. The document with the highest final score will be selected as the next optimal document. **MMR (Maximal Marginal Relevance)** is a typical integration pattern that combines relevance and diversity scores linearly based on a tradeoff factor λ. As a hyper-parameter, λ is designed manually and may not be optimal. To find proper tradeoff factors, we use a learnable matrix W_3 to aggregate the scores. Consistent with Equation (2), we have:

$$P(d|q, \mathcal{D}_t, \mathcal{S}_t, u) = \zeta(s) = \tanh(W_3^T s), \tag{15}$$

where ζ is an MLP layer with $\tanh(\cdot)$ as activation function and s is a vector consists of scores we obtained, i.e.,

$$s = \left(f_g(d, q), r^S f_S(d, S), r^L f_L(d, L), \text{rel}(q, d)\right)^T.$$

3.7 Optimization of Our Model

Section 3.3.2 has demonstrated an auxiliary task of our optimization that ensures the heterogeneity of the virtual subtopics. In fact, the main task of our model is reranking the candidate documents and providing more satisfactory ranking results for the user, so employing a learning-to-rank loss function to train our model is indispensable. In previous works on personalized search [17, 25, 48, 50], researchers always thought that the clicked documents only revealed the relevance between documents and queries. However, human behaviors and their motivations are complex and diverse. When a user is searching for desired documents, she may neglect the redundant documents, because she has browsed another similar document before, despite the fact that these redundant documents can satisfy her original information need. **It means that users' click behaviors are not just decided by the relevance of documents, but also depend on the diversity of results. Thus, we can view clicked documents as satisfactory documents that reflect the user's satisfaction with relevance and diversity, and non-clicked documents as unsatisfactory ones.** The training samples of our model can be viewed as pairwise samples. We denote them as $x = (q, H, C = \{d_1 | i \in [1, t-1]\}, (d_a, d_b)), y = \{0, 1\}$, where C denotes the previous $t-1$ documents, which are sampled from the original ranking. H represents the user's search history, and d_a and d_b denote a pair of candidate documents for the tth position. y is the ground truth label. Under selecting the C, if d_a is a better choice than d_b at the tth position, then $y = 1$, otherwise, $y = 0$. Our model generates the final scores of d_a and d_b based on the input data, and calculate $P(d_a, d_b | q, H, C)$, which denotes the probability that d_a is better than d_b. For brevity,

we will refer to $P(d_a, d_b|q, H, C)$ as $P(d_a, d_b)$. We have:

$$P(d_a, d_b) = \frac{1}{1 + \exp(-(\text{score}(d_a) - \text{score}(d_b)))}, \tag{16}$$

where $\text{score}(d_a)$ denotes $P(d_a|q, \mathcal{D}_t, \mathcal{S}_t, u)$. Thus, we can use binary classification loss to construct our loss function:

$$\mathcal{L}_{rank} = -\sum_{q \in Q} \sum_{o=1}^{|\Delta_q|} y^o \log(P(d_a^o, d_b^o)) + (1 - y^o)(\log(P(d_b^o, d_a^o))), \tag{17}$$

where Q represents all the queries and $\Delta_q = \{(q, H, S, (d_a^o, d_b^o))\}$ denotes the set consists of all pairs w.r.t. query q, y^o is the ground truth that d_a^o is better than d_b^o.

Therefore, the final loss function of our proposed model combines the diversity loss function, $\mathcal{L}_{div}$, which is introduced in Section 3.3.2, and the ranking loss function, $\mathcal{L}_{rank}$:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rank} + \lambda_2 \mathcal{L}_{div}, \tag{18}$$

where λ_1 and λ_2 are hyper-parameters that control the importance of two loss functions.

4 EXPERIMENT SETTINGS

4.1 Dataset

We evaluate our model and baselines based on a large-scale search log collected from an English-language commercial search engine between 1 January, 2013, and 28 February, 2013. Each piece of data contains an anonymous user ID, a session ID, a query string, query issued time, top retrieved URLs and corresponding document contents, click labels, documents' dwelling time, and so on. During the collection of this dataset, the search engine was not equipped with personalization, so the click-through was ensured not to be biased toward other personalization signals. The dataset contains 5,317 users and 2,665,625 queries. Considering the noise in click-through, e.g., the user may fault-click a document or find it is not the right one and close it quickly, we view a document that has a dwell time longer than 30 seconds or is clicked on the last as a satisfactory document following Reference [48]. The sessions are constructed by treating 30 minutes of inactivity as the boundary. Moreover, to construct the historical search of users, we segment the users' data in the first six weeks as the basic search history, which is used to filter users with inadequate search logs, and the last two weeks as experimental data. Note that the long-term history is not only built based on the first six weeks but also on all the search behaviors before the current session to ensure the timeliness and adequacy of user histories. Since we consider the personalization of search results, the division strategy of our dataset is applied to the search logs of each user rather than the entire search logs. Specifically, we divide the experimental data of every user into the training set, valid set, and test set according to the ratio of 4:1:1. The statistics information of the experimental data is shown in Table 2. Note that there exist some publicly available datasets for personalized search, i.e., AOL, ORCAS, and WEBIS logs. However, AOL only contains clicked documents for each query without actual original ranking results. Though there are methods to crawl candidate documents for queries to build pseudo original ranking lists, However, considering that users usually browse documents from top to bottom, a real original ranking list is critical to inferring the user's diverse interests, since some aspects of her interests may be satisfied by previously reviewed documents. Therefore, such fake original ranking lists make it hard to model the user's diverse interests, which may not be suitable for evaluating our method. Similarly, the ORCAS and WEBIS logs both lack user identifiers. Moreover, WEBIS only contains 13,651 queries and 16,739 clicks, which may prevent sufficiently optimizing our model. Thus, even if it may introduce some

Table 2. Basic Statistics of the Experiment Dataset

Item	Train	Valid	Test
#data pair	579,862	124,525	149,854
#queries	82,067	17,356	21,150
#sessions	66,157	14,176	21,150
Avg query length	13.0	12.9	15.8
Avg session length	2.74	2.73	2.83
Avg #click per query	1.19	1.18	1.19

reproducibility limitations, we conduct our experiment on a large-scale commercial query log with actual original rankings to guarantee the reliable analysis of real user behaviors.

4.2 Baselines

For evaluating the performance of our models, we select some personalization, diversification, and personalized diversification methods to compare with our models.

(1) Ad hoc baselines:

Original We use the original ranking result of the commercial dataset as the ad hoc baseline result.

(2) Personalized Search Methods:

P-Click [15]. Dou et al. proposed P-Click to conduct the personalized search. It calculates the relevance of documents to the query and the user based on the click feature.

SLTB [2]. SLTB applies LambdaMART [5] to optimize a personalized ranking model based on 102 features extracted from the user search log, including query-doc-user Features features, query-history features, and so on.

HRNN [17]. HRNN dynamically models the user's short-term and long-term profiles from her search history using hierarchical RNN structures. To further consider the discrepant importance of each historical behavior, the authors conduct the attention mechanism to capture a more reliable and stable long-term user profile.

HTPS [50]. HTPS learns personalized information from user history by constructing a transformer-based structure, thereby disambiguating the issued query sentences and capturing accurate user intents.

PEPS [48]. PEPS proposes an alternative method for personalizing search results. It builds a personalized word embedding matrix for each user, which only contains the aspects that the user is interested in. Then, the K-NRM is utilized to calculate the similarity between the word embeddings of the query and the document.

(3) Diversification Method:

MMR [6]. MMR is a typically unsupervised implicit method for search result diversification. It calculates the relevance score and diversity score of documents and linearly combines them by trade factor λ.

IA-Select [1]. IA-Select was proposed to consider the subtopic to satisfy the diverse intents of different users.

xQuAD [31]. xQuAD is another explicit diversification method that follows the MMR paradigm to combine relevance and diversity linearly.

R-LTR-NTN [46]. R-LTR-NTN is a supervised method. It introduces the **neural tensor network (NTN)** structure, which is learnable, to model the dissimilarity between a document and selected documents as the novelty of the document and applies the ranking algorithm of R-LTR [53] to train the model.

DESA-IM [28]. DESA is a search result diversification model that discards the paradigm of greedy ranking. It applies the transformer structure to diversify the search result. We selected the implicit part of DESA, DESA-IM, as the representative of diversification methods to conduct the experiment.

(4) Personalized Diversification Method.

P-IA-Select & P-xQuAD [36]. Vallet introduced a user variant u to existing heuristic diversity methods, IA-Select and xQuAD, hence improved them to personalized diversity methods, P-IA-Select and P-xQuAD.

Pnd. In this article, we propose a framework that integrates search result personalization and diversification in a deep-learning manner.

4.3 Implement Details

For baselines, we implement the best parameters based on their papers. In our experiment, the hyper-parameter configuration is as follows: The dimension of the word embedding is 100; for the profile generator, the hidden size of the transformer is 256; and the number of heads in multi-head self-attention is 8. For the subtopic learning layer, the number of project matrices, c, is 8, the activation is LeakyReLU, and the dimension of the virtual subtopic, h, is 32. For the reset gate, its hidden size is 128, and the activation functions of the first and second layers are LeakyReLU and Tanh. For our similarity function F_k, the number of RBF kernels is 11, with the mean value u_o evenly distributed in $(-1, 1)$ based on k and the variance σ_o is set to 0.1. The learning rate is 5e-4, and the weight decay coefficient is set to 1e-5. We apply the Adam algorithm to optimize our model.

4.4 Evaluation Metrics

4.4.1 Metrics of Personalization. **In this article, we believe that the click data reflects the users' satisfaction with results, which includes relevance and diversity. Therefore, we can select the personalization metrics, which also focus on click data, to evaluate the overall performance of models.** Following previous works on personalization, we choose three widely used metrics, MAP, MRR, and P@1. Their calculations are shown as follows:

MAP. Mean average precision (MAP) is a widely used metric in information retrieval. It considers the rank and relevance of the result documents, since people not only want to search for satisfactory documents but also hope that they can be ranked at the top. Considering that there are M queries and the number of clicked (relevant) documents of jth query is n_j, the formulation of MAP is:

$$\text{MAP} = \frac{1}{M}\sum_{j=1}^{m}\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{i}{\text{Pos}(i)}, \tag{19}$$

where $\text{Pos}(i)$ represents the position of the ith relevant document. While considering the relevance, MAP is also concerned about the position: A relevant document with a lower position contributes less to MAP. This property is in line with users' expectations.

MRR. Indeed, many people finish the browser after finding the first relevant document, thus, the first relevant document is rather important. **Mean reciprocal rank (MRR)** is the metric that focuses on the position of the first relevant document, rp:

$$\text{MRR} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{rp}, \tag{20}$$

where M denotes the number of all queries.

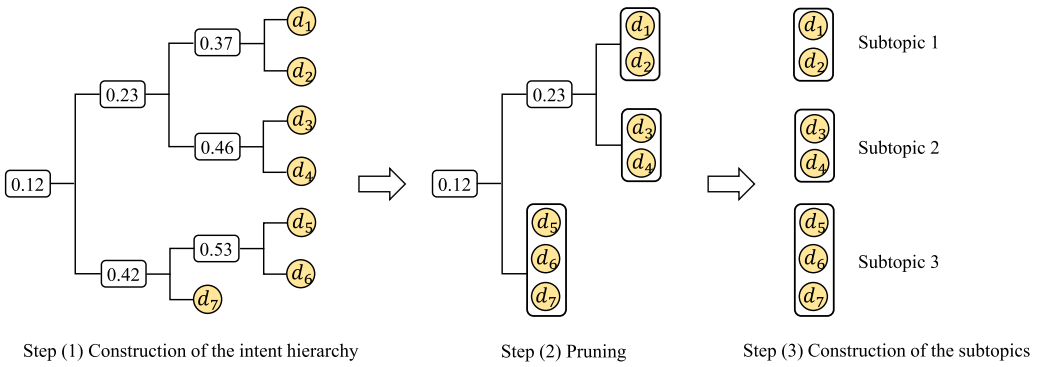

Fig. 3. The processing flow of the first pseudo-subtopic generation method.

P@k. More generally, we should pay more attention to the specific position, as users will neglect documents with low positions even if they are relevant. Precise@k (P@k) allows us to consider the top-k documents only. Assuming that the number of relevant documents on the top K of results is N_i for the ith query, the formula of P@k is shown as:

$$\text{P@}k = \frac{1}{M}\sum_{i=1}^{M}\frac{N_i}{k}. \tag{21}$$

In this article, we select P@1 to focus on the performance of the top-1 results.

4.4.2 Metrics of Diversification. Since the dataset we use is common for personalized search and lacks subtopic information, we have to construct pseudo subtopics to further evaluate the diversity of results explicitly. To ensure the reliability of the evaluation, we adopt two methods:

(1) Following Reference [16], given the candidate documents of the query, we apply a hierarchical clustering algorithm to them to produce an intent hierarchy. Then, we utilize a pruning technique to aggregate similar nodes whose similarity exceeds the threshold $\alpha = 0.35$. The yielded clusters denote the pseudo subtopics of the query. An example is shown in Figure 3.
(2) Following Reference [11], we view each query q as a seed and mine out the queries that occurred in the same session, i.e., "co-session" relationship, and undertake this expansion twice. Then, we construct the graph on these queries, each node denotes a query in the set, and two queries that have the same clicked documents, i.e., "co-click" relationship, will be linked by an edge. Finally, we apply a clustering algorithm to yield query clusters, which represent subtopics of q. The processing flow is shown in Figure 4.

After that, we choose two common metrics, ERR-IA@10 [7] and α-NDCG@10 [10], for evaluation of diversity. Their computations are presented below.

ERR-IA. The idea of ERR is similar to that of MRR; the difference is that ERR takes into account the diversity between documents.

$$\text{PP}(k) = \prod_{i=1}^{k-1}(1 - r(i))r(k), \tag{22}$$

$$\text{ERR} = \sum_{k=1}^{n}\phi(k)\text{PP}(k). \tag{23}$$

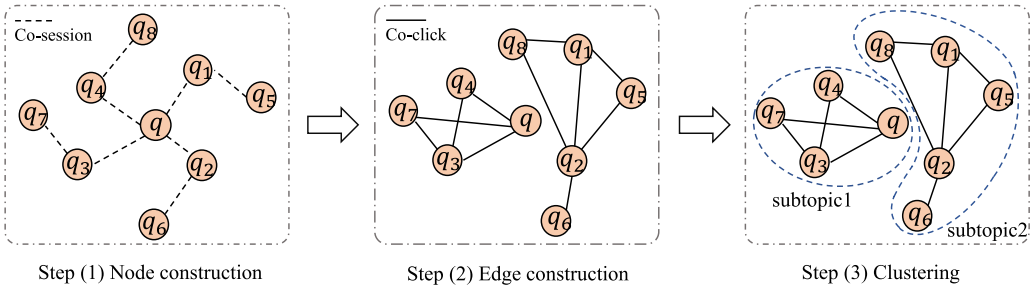

Fig. 4. The processing flow of the second pseudo-subtopic generation method.

$r(i)$ stands for the probability that the ith document can satisfy the user. $PP(k)$ denotes the probability that the top-k documents can satisfy the user. $\phi(k)$ denotes the position function, and n is the number of top results that are used to calculate the metric value. ERR-IA is the improved version of ERR that considers the **intent-aware (IA)** component.

$$\text{PP-IA}(k,s) = \prod_{i=1}^{k-1}(1 - r(i,s))r(k,s), \tag{24}$$

$$\text{ERR-IA} = \frac{1}{m}\sum_{s=1}^{m}\sum_{k=1}^{n}\phi(k)\text{PP-IA}(k,s), \tag{25}$$

where $r(i,s)$ denotes the relevance of the ith document to the sth subtopic. In this article, we produce the $r(i,s)$ based on the cosine similarity between the document and the subtopic.

α-NDCG. α-NDCG is an updated version of NDCG, which is a classical metric in information retrieval. The main idea of NDCG is that (1) the higher the relevance of documents, the higher the influence the evaluation result will have. (2) The higher the position of relevant documents, the higher the influence, too. Given a ranking list R of a set of candidate documents D, and its optimal ranking R^*, the DCG of top-k results on R is that:

$$\text{DCG}(R,k) = \sum_{i=1}^{k}\frac{2^{r(i)} - 1}{\log(i+1)}, \tag{26}$$

where $r(i)$ denotes the relevance score of the ith document in R. NDCG is the normalization of DCG:

$$\text{NDCG}(R,k) = \frac{\text{DCG}(R,k)}{\text{DCG}(R^*,k)}. \tag{27}$$

To evaluate the diversity and subtopic coverage of the ranking list, α-NDCG introduces the intent-aware component and models the redundancy by introducing the penalty factor of redundancy α. Its DCG value is yielded as follows:

$$\text{D}\mathfrak{C}\text{G}(R,k) = \sum_{s=1}^{m}\sum_{i=1}^{k} r(k,s)(1-\alpha)^{C(k-1,s)}, \tag{28}$$

$$C(k-1,s) = \sum_{j=1}^{k-1} r(j,s), \tag{29}$$

where $r(k,s)$ evaluates the relevance of the kth result to the sth subtopic and $C(k-1,s)$ measures the relevance of the top k-1 documents to the sth subtopic. If the sth subtopics have been covered

Table 3. Overall Performances of Baselines and Our Model

Task	Model	MAP		MRR		P@1	
Ad hoc Search	Original	.7399	−10.0%	.7506	−9.8%	.6162	−15.9%
Personalized Search	P-Click	.7509	−8.66%	.7634	−8.3%	.6260	−13.7%
	SLTB	.7921	−3.65%	.7998	−3.9%	.6901	−4.8%
	HRNN	.8065	−1.90%	.8191	−1.6%	.7127	−1.7%
	HTPS	.8220	−0.01%	.8318	−0.04%	.7291	0.55%
	PEPS	.8221	–	.8321	–	.7251	–
Search Result Diversification	MMR	.4212	−48.77%	.4304	−48.28%	.2044	−71.81%
	IA-Select	.7137	−13.19%	.7247	−12.91%	.5738	−20.87%
	xQuAD	.7280	−11.45%	.7388	−11.21%	.5942	−18.05%
	R-LTR-NTN	.6881	−16.30%	.7036	−15.44%	.5824	−19.68%
	DESA	.6128	−25.46%	.6235	−25.07%	.4383	−39.55%
Integrated Methods	P-IA-Select	.7374	−10.30%	.7478	−10.13%	.6102	−15.85%
	P-xQuAD	.7386	−10.16%	.7491	−9.97%	.6123	−15.56%
	PnD	**.8279**†	+0.72%	**.8379**†	+0.71%	**7343**†	+1.27%

"†" indicates the model outperforms all baselines significantly with a paired t-test at the $p < 0.05$ level. The best results are shown in bold.

by previous results, then α-NDCG will restrain the contribution of the kth document in the aspects of the sth subtopic. α-NDCG is normalized in the same way as NDCG by

$$\alpha\text{-NDCG} = \frac{\widetilde{\text{DCG}}(R,k)}{\widetilde{\text{DCG}}(R^*,k)}.$$

5 EXPERIMENT RESULTS

5.1 Overall Performance

Compared to all the baselines, the overall evaluation results are shown in Tables 3 and 4. From the tables, we can observe that:

(1) **On the personalized metrics, our model significantly outperforms all baselines in terms of satisfaction with a paired t-test at the $p < 0.05$ level.** Different from baselines, our model first obtains the user's initial profiles and learns their virtual subtopics. Then, the RRNN component models the selected document's coverage of the aspects that the user is interested in. It prompts the model to capture more comprehensive user intents, which enhances the relevance of results while maintaining diversity. Thereby, more satisfactory results can be produced. And this result also implies that personalization and diversification are not contrary but can complement each other and prompt the satisfaction of the results in a reasonable way.

(2) **Our model outperforms all personalized methods on diversified metrics.** Meanwhile, our model shows comparable performance with some diversification approaches. Obviously, the results of diversified methods have high diversity but show a sharp decline in satisfaction. **Therefore, we deem that higher diversity is not always better, offering unwanted documents will negatively impact the users' satisfaction.** Different from pure diversification methods, our model incorporates the users' click data, which provides accurate information about users' satisfaction. Consequently, it performs well on both. The above results verify that personalization and diversification can be integrated to provide more favorable results and also confirm that human behavior is complex, thus, the click-through reflects the user's preferences and diversity needs simultaneously. Note that those super-

Table 4. Overall Performances of Models on Diversification

Task	Model	ERR-IA[1]	α-NDCG[1]	ERR-IA[2]	α-NDCG[2]
Ad hoc Search	Original	.5196	.3269	.3197	.1915
Personalized Search	P-Click	.5197	.3269	.3201	.1917
	SLTB	.5198	.3270	.3202	.1917
	HRNN	.5184	.3258	.3201	.1916
	HTPS	.5200	.3274	.3199	.1916
	PEPS	.5201	.3271	.3007	.1832
Search Result Diversification	MMR	.5353	.3325	.3085	.1865
	IA-Select	**.5516**	**.3403**	**.3377**	**.1984**
	xQuAD	.5515	.3403	.3309	.1957
	R-LTR-NTN	.5435	.3361	.3203	.1917
	DESA	.5456	.3369	.3110	.1877
Integrated Methods	P-IA-Select	.5216*	.3278*	.3229*	.1931*
	P-xQuAD	.5216*	.3278*	.3230*	.1931*
	PnD	.5233*	.3284*	.3216*	.1923*

Superscript 1 of the diversity metrics denotes the first subtopic mining method, and 2 denotes the second one. "$\star$" means the integrated model outperforms all the personalized baselines on diversification. The best results are shown in bold.

vised diversification methods perform worse than heuristic ones; the reason may be that they are highly dependent on data.

(3) **Compared with other integrated methods, our model significantly outperforms in satisfaction and shows comparable performance in diversity.** Also, the results show that all integrated methods perform better in diversity than personalized ones and in satisfaction than diversified ones. This result suggests that these integrated approaches have a certain ability to balance personality and diversity, but they cannot perform better on both aspects, which means that they lack the ability to capture the in-depth relationship between personalization and diversification. However, the thorough analysis of these two approaches that went into building our model allowed it to integrate them more dependably and produce results that were more satisfying.

5.2 Ablation Analysis

To verify the effectiveness of the components in our framework, we conduct an ablation analysis on these structures. Specifically, we remove one component once to produce an incomplete model and train it on the same dataset to compare with the integral model.

Pnd w/o. SL. For verifying the effectiveness of virtual subtopics, We abandon the **subtopic learning layer (SL)** and generate matching scores by cosine similarity function.

PnD w/o. RRNN. To test the effectiveness of RRNN. We deactivate it and calculate novelty scores based on the original representations of the query and profiles.

PnD w/o. DPM. We discard the **diversified personalization modules (DPM)** to analyze their importance.

PnD w/o. GDM. To verify the usefulness of general diversification, we remove the **general diversification module (GDM)** and construct the PnD w/o. GDM.

PnD w/o. WP. We abandon the **weighting process (WP)** to evaluate the benefit of dynamical fusion.

The results of ablation are shown in Table 5. We observe that all ablation strategies underperform the complete framework in satisfaction and diversity. It reveals that each component plays

Table 5. Ablation Results

Model	MAP		ERR-IA[1]	
PnD	.8279	–	.5233	–
w/o. SL	.8208	−0.86%	.5225	−0.15%
w/o. RRNN	.8222	−0.70%	.5207	−0.50%
w/o. DPM	.8240	−0.48%	.5223	−0.19%
w/o. GDM	.8250	−0.36%	.5228	−0.10%
w/o. WP	.8223	−0.69%	.5220	−0.25%

an indispensable role in our model to return more satisfactory results for users. Specifically, we give the following analysis:

(1) We find that the deficiency of the SL leads to a significant drop in MAP compared with the complete model. The reason is that the subtopic learning layer can capture a fine-grained representation of the virtual subtopics from multiple aspects and consider them explicitly. Based on this, we can learn more exact user intents, which contribute to improving the effectiveness, especially in personalization. However, the removal of SL has little impact on diversification. We infer that RRNN has good expression ability in implicit representation even if there is no virtual subtopic. Therefore, the function of the virtual subtopic is mainly to accurately capture user interests.
(2) With regard to w/o. RRNN, the experimental results show that it has a significant influence on both personalization and diversification, especially on the latter. A reasonable explanation for this phenomenon is that after clicking a related document, the user's information need on this aspect is satisfied; if she decides to continue browsing, then she will be inclined to click the documents related to other preferences. Without RRNN, our model cannot mine out the various satisfactory documents from diverse aspects, which will harm performance and lead to redundancy. This result confirms our assumption that there is inherent diversity in user interests, thus, it is insufficient to consider relevance purely.
(3) The performance of w/o. PDM and w/o. GDM agrees with the assumption that personalized diversification and general diversification jointly promote our model to capture user intents. Once we discard the PDM, our model cannot capture accurate user intents. It may cause documents that are unrelated to users' interests to be ranked at the top. If we abandon the GDM, then the probability of meeting the user's intents will be affected while the degree of refinding is low, influencing the performance of our model.
(4) From the ablation studies, we find that the deficiency of the weighting process is also harmful to user satisfaction with ranking results. This experimental result verifies our assumption that diversification and personalization show different importance in different scenarios, thus, they should be dynamically integrated based on the degree of refinding to provide more satisfactory results.

5.3 Visualization Experiments

The above ablation studies elucidate the indispensability of our modules. Furthermore, to analyze the process of our key modules in-depth, we devise some visualization experiments and demonstrate our analysis as follows:

5.3.1 Effect of Virtual Subtopics. The subtopic learning layer is designed to learn fine-grained and diverse intents; we hope our selected documents can cover more important virtual subtopics to improve satisfaction. We randomly select a query, "best mpg cars" and visualize the coverage

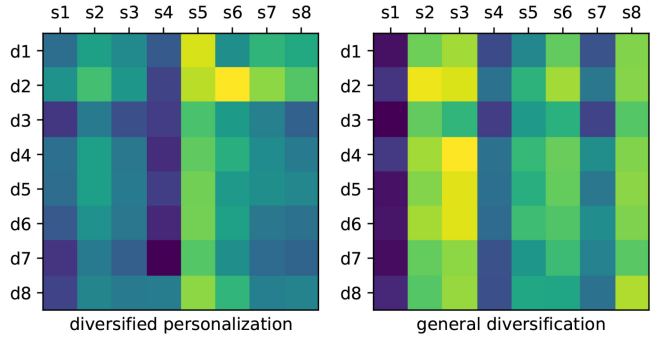

Fig. 5. The visualization of virtual subtopics coverage.

of reranked documents on virtual subtopics in Figure 5. sj denotes the jth virtual subtopic, and $d1$ denotes the document ranked first. The darker the cell, the lower the relevance between the document and subtopic. It can be seen that each document has different levels of coverage for different subtopics. When earlier documents adequately cover a subtopic, the later ones will focus more on other subtopics to meet the user's other intents. For example, in the left figure of Figure 5, the first document, d_1, is most relevant to s5, so for the second document, our model selects the one that pays more attention to s6, which can reflect the diversity of the ranking results. This phenomenon is in line with our assumption. We notice that some subtopics are highly relevant to most documents, since the corresponding cells are bright, while some subtopics are irrelevant to most documents. We infer three reasons: (1) The subtopic number is the general value for all queries, while the user intent may contain fewer subtopics, resulting in the candidate documents focusing on some main subtopics. These bright subtopics may represent the user's main interests, and the dark subtopics are the aspects that the current query is not related to.

5.3.2 Effect of RRNN. The RRNN module is devised to update the representations of user interests and query aspects after selecting a local optimal document. To visualize the change of representations at each step, we concatenate the virtual subtopics into a vector and project it on 2-dimensional space, which is shown in Figure 6. The number on the figure indicates the number of steps. It illustrates that in both parts, the representation of the hidden state in each step is different from the others. In other words, after each selection, the hidden state will be updated, which prompts the model to focus on the other subtopics that have not been covered. We notice that the distance between adjacent points falls as the number of steps increases. Our analysis is that with the growth of the step number, less information can be forgotten, thus the spacing gets smaller overall. This phenomenon is consistent with our expectation of diversification. In the case of occasional larger spacing, the reason may be that the document selected at this step covers less important but more diverse subtopics; as our model considers relevance and diversity simultaneously, it cannot be selected ahead.

5.4 Quantity Setting of Virtual Subtopics

The previous results show that the application of virtual subtopics is in favor of both satisfaction and diversity. To further analyze the impact of its quantity setting on effectiveness, we conduct multiple experiments with different quantity settings and display their performance on personalization and diversification. The results are presented in Figure 7. The Y-axis represents the percentage of performance decline compared to the optimal setting. It can be found that with the growth of the number of virtual subtopics, the overall performance presents a trend of first increasing and then

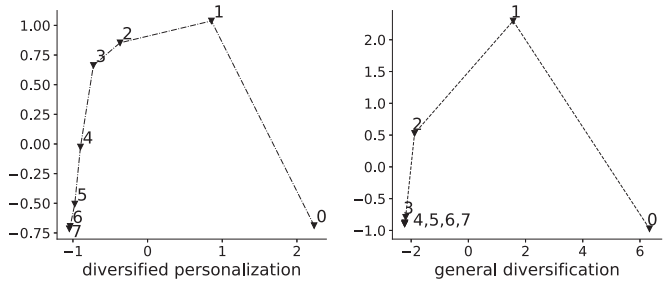

Fig. 6. The visualization of user interests at each step.

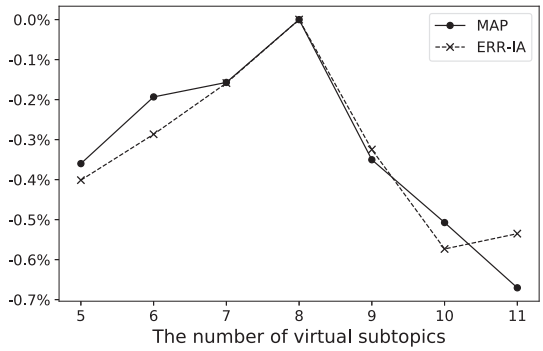

Fig. 7. Performance trends for subtopic quantity settings.

Table 6. Two Ranking Results Produced by Our Model

query	distracted driving accident deid	blood vessel broken in eye
d_1	distracted driving nhtsa distracted driving nhtsa skip main (click)	subconjunctival hemorrhage broken blood vessel eye overview mayo (click)
d_2	many crashes minimal fines distracted driving new york (click)	causes broken eye blood vessels livestrong com broken (click)
d_3	yakima car accident lawyer mariano morales law encourages (non-click)	treatment broken blood vessels eye ehow treatment broken (non-click)
d_4	distracted driving motor vehicle safety cdc injury center (non-click)	ohio lions eye research foundation category broken blood (non-click)

decreasing. The reason may be that when the number is small, the model could capture more beneficial information as the number of subtopics increases, while once the number exceeds a threshold, too many subtopics will introduce noisy information, which will negatively affect performance.

5.5 Case Study

To confirm that our model is able to provide ranking results meeting diversified user intents, we illustrate two case studies in Table 6.

We present the issued query and the top-four ranking documents. It can be found that for the query "distracted driving accident deid," the first document (clicked) is the description of "distracted driving" provided by the **National Highway Traffic Safety Administration (NHTSA)**, which is an official website. The second document (click) deals with fines for traffic accidents. These two documents are both related to distracted-driving accidents while showing the fine-grained di-

versity in the perspective of subtopics. Similarly, for the query "blood vessel broken in eye," the first document is related to an overview of blood vessel broken eyes, while the second one is about the cause of this disease. These two case studies suggest that when the user issues a query, she may be interested in various subtopics of this query; in other words, the user's interests are also diverse. The ranking results of our model reveal the ability of our method to model the diversity of user interests and provide satisfactory results for the user.

6 CONCLUSION

In this article, we propose a novel framework that integrates search result personalization and diversification dynamically based on the refinding degree. We further consider the diversity of users' interests in personalization to satisfy the user's needs comprehensively. To model the influence of selected documents on novelty, we devise the RRNN model based on the "reset" gate. Moreover, we design the "subtopic learning layer" to learn the virtual subtopics and consider the subtopics explicitly. Experiment results verify the effectiveness of our model for both personalization and diversification. The flexibility of our framework is pretty high, and the profile generator, embedding matrix, and weighting process are all replaceable. In the future, we can replace these components with more advanced structures to improve the performance of our framework.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China No. 62272467, No. 61832017 and No. 62102421, Beijing Outstanding Young Scientist Program No. BJJWZYJH012019100020098, Public Computing Cloud, Renmin University of China, the fund for building world-class universities (disciplines) of Renmin University of China.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. Association for Computing Machinery, New York, NY, 5–14. DOI: https://doi.org/10.1145/1498759.1498766

[2] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. Association for Computing Machinery, New York, NY, 185–194. DOI: https://doi.org/10.1145/2348283.2348312

[3] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. Association for Computing Machinery, New York, NY, 185–194. DOI: https://doi.org/10.1145/2348283.2348312

[4] David Blei, Andrew Ng, and Michael Jordan. 2001. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 601–608.

[5] Chris J. C. Burges, Krysta M. Svore, Qiang Wu, and Jianfeng Gao. 2008. *Ranking, Boosting, and Model Adaptation.* Microsoft Research, Technical Report MSR-TR-2008-109. Retrieved from https://www.microsoft.com/en-us/research/publication/ranking-boosting-and-model-adaptation/

[6] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*. Association for Computing Machinery, New York, NY, 335–336. DOI: https://doi.org/10.1145/290941.291025

[7] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. Association for Computing Machinery, New York, NY, 621–630. DOI: https://doi.org/10.1145/1645953.1646033

[8] Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. 2017. Personalized query suggestion diversification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. Association for Computing Machinery, New York, NY, 817–820. DOI: https://doi.org/10.1145/3077136.3080652

[9] Paul Alexandru Chirita, Wolfgang Nejdl, Raluca Paiu, and Christian Kohlschütter. 2005. Using ODP metadata to personalize search. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development*

in Information Retrieval (SIGIR'05). Association for Computing Machinery, New York, NY, 178–185. DOI: https://doi.org/10.1145/1076034.1076067

[10] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. Association for Computing Machinery, New York, NY, 659–666. DOI: https://doi.org/10.1145/1390334.1390446

[11] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 web track. In *Proceedings of the 18th Text REtrieval Conference (TREC'09) (NIST Special Publication)*, Ellen M. Voorhees and Lori P. Buckland (Eds.), Vol. 500-278. National Institute of Standards and Technology (NIST). Retrieved from http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf

[12] Steve Cronen-Townsend and W. Bruce Croft. 2002. Quantifying query ambiguity. In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT'02)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 104–109.

[13] Van Dang and Bruce W. Croft. 2013. Term level search result diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. Association for Computing Machinery, New York, NY, 603–612. DOI: https://doi.org/10.1145/2484028.2484095

[14] Van Dang and W. Bruce Croft. 2012. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. Association for Computing Machinery, New York, NY, 65–74. DOI: https://doi.org/10.1145/2348283.2348296

[15] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. Association for Computing Machinery, New York, NY, 581–590. DOI: https://doi.org/10.1145/1242572.1242651

[16] Zhicheng Dou, Xue Yang, Diya Li, Ji-Rong Wen, and Tetsuya Sakai. 2020. Low-cost, bottom-up measures for evaluating search result diversification. *Inf. Retr. J.* 23, 1 (2020), 86–113. DOI: https://doi.org/10.1007/s10791-019-09356-x

[17] Songwei Ge, Zhicheng Dou, Zhengbao Jiang, Jian-Yun Nie, and Ji-Rong Wen. 2018. Personalizing search results using hierarchical RNN with query-aware attention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18)*. Association for Computing Machinery, New York, NY, 347–356. DOI: https://doi.org/10.1145/3269206.3271728

[18] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR'16)*, Yoshua Bengio and Yann LeCun (Eds.). Retrieved from http://arxiv.org/abs/1511.06939

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computat.* 9 (12 1997), 1735–80. DOI: https://doi.org/10.1162/neco.1997.9.8.1735

[20] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM'15)*. Association for Computing Machinery, New York, NY, 63–72. DOI: https://doi.org/10.1145/2806416.2806455

[21] Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou, Wayne Xin Zhao, Jian-Yun Nie, and Ming Yue. 2017. Learning to diversify search results via subtopic attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. Association for Computing Machinery, New York, NY, 545–554. DOI: https://doi.org/10.1145/3077136.3080805

[22] Shangsong Liang, Zhaochun Ren, and Maarten de Rijke. 2014. Personalized search result diversification via structured learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. Association for Computing Machinery, New York, NY, 751–760. DOI: https://doi.org/10.1145/2623330.2623650

[23] Jiongnan Liu, Zhicheng Dou, Xiaojie Wang, Shuqi Lu, and Ji-Rong Wen. 2020. DVGAN: A minimax game for search result diversification combining explicit and implicit features. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20)*. Association for Computing Machinery, New York, NY, 479–488. DOI: https://doi.org/10.1145/3397271.3401084

[24] Jiongnan Liu, Zhicheng Dou, Qiannan Zhu, and Ji-Rong Wen. 2022. A category-aware multi-interest model for personalized product search. In *Proceedings of the ACM Web Conference (WWW'22)*. Association for Computing Machinery, New York, NY, 360–368. DOI: https://doi.org/10.1145/3485447.3511964

[25] Shuqi Lu, Zhicheng Dou, Xu Jun, Jian-Yun Nie, and Ji-Rong Wen. 2019. PSGAN: A minimax game for personalized search with limited and noisy click data. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, 555–564. DOI: https://doi.org/10.1145/3331184.3331218

[26] Zhengyi Ma, Zhicheng Dou, Guanyue Bian, and Ji-Rong Wen. 2020. *PSTIE: Time Information Enhanced Personalized Search.* Association for Computing Machinery, New York, NY, 1075–1084. https://doi.org/10.1145/3340531.3411877

[27] Ahmet Murat Ozdemiray and Ismail Sengor Altingovde. 2015. Explicit search result diversification using score and rank aggregation methods. *J. Assoc. Inf. Sci. Technol.* 66, 6 (2015), 1212–1228. DOI: https://doi.org/10.1002/asi.23259

[28] Xubo Qin, Zhicheng Dou, and Ji-Rong Wen. 2020. *Diversifying Search Results Using Self-attention Network.* Association for Computing Machinery, New York, NY, 1265–1274. DOI: https://doi.org/10.1145/3340531.3411914

[29] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys'17).* Association for Computing Machinery, New York, NY, 130–137. DOI: https://doi.org/10.1145/3109859.3109896

[30] Filip Radlinski and Susan Dumais. 2006. Improving personalized web search using result diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06).* Association for Computing Machinery, New York, NY, 691–692. DOI: https://doi.org/10.1145/1148170.1148320

[31] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10).* Association for Computing Machinery, New York, NY, 881–890. DOI: https://doi.org/10.1145/1772690.1772780

[32] Procheta Sen, Debasis Ganguly, and Gareth J. F. Jones. 2021. I know what you need: Investigating document retrieval effectiveness with partial session contexts. *ACM Trans. Inf. Syst.* 40, 3, Article 53 (Nov. 2021), 30 pages. DOI: https://doi.org/10.1145/3488667

[33] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum* 33, 1 (Sept. 1999), 6–12. DOI: https://doi.org/10.1145/331403.331405

[34] Zhan Su, Zhicheng Dou, Yutao Zhu, Xubo Qin, and Ji-Rong Wen. 2021. Modeling intent graph for search result diversification. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'21).* Association for Computing Machinery, New York, NY, 736–746. DOI: https://doi.org/10.1145/3404835.3462872

[35] Bin Tan, Xuehua Shen, and ChengXiang Zhai. 2006. Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06).* Association for Computing Machinery, New York, NY, 718–723. DOI: https://doi.org/10.1145/1150402.1150493

[36] David Vallet and Pablo Castells. 2012. Personalized diversification of search results. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12).* Association for Computing Machinery, New York, NY, 841–850.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).* Curran Associates Inc., Red Hook, NY, 6000–6010.

[38] Maksims Volkovs. 2015. Context models for web search personalization. *CoRR* abs/1502.00527 (2015).

[39] Thanh Tien Vu, Alistair Willis, Son Ngoc Tran, and Dawei Song. 2015. Temporal latent topic user profiles for search personalisation. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29–April 2, 2015. Proceedings (Lecture Notes in Computer Science)*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.), Vol. 9022. 605–616. DOI: https://doi.org/10.1007/978-3-319-16354-3_67

[40] Hongning Wang, Xiaodong He, Ming-Wei Chang, Yang Song, Ryen W. White, and Wei Chu. 2013. Personalized ranking model adaptation for web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13).* Association for Computing Machinery, New York, NY, 323–332. DOI: https://doi.org/10.1145/2484028.2484068

[41] Shuting Wang, Zhicheng Dou, Jing Yao, Yujia Zhou, and Ji-Rong Wen. 2023. Incorporating explicit subtopics in personalized search. In *Proceedings of the ACM Web Conference (WWW'23).* Association for Computing Machinery, New York, NY, 3364–3374. DOI: https://doi.org/10.1145/3543507.3583488

[42] Shuting Wang, Zhicheng Dou, and Yutao Zhu. 2023. Heterogeneous graph-based context-aware document ranking. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining (WSDM'23).* Association for Computing Machinery, New York, NY, 724–732. DOI: https://doi.org/10.1145/3539597.3570390

[43] Ryen W. White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13).* Association for Computing Machinery, New York, NY, 1411–1420. DOI: https://doi.org/10.1145/2488388.2488511

[44] Shengli Wu, Chunlan Huang, Liang Li, and Fabio Crestani. 2019. Fusion-based methods for result diversification in web search. *Inf. Fusion* 45 (2019), 16–26. DOI: https://doi.org/10.1016/j.inffus.2018.01.006

[45] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *Proceedings of the 38th International ACM SIGIR Conference on*

Research and Development in Information Retrieval (SIGIR'15). Association for Computing Machinery, New York, NY, 113–122. DOI: https://doi.org/10.1145/2766462.2767710

[46] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling document novelty with neural tensor network for search result diversification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. Association for Computing Machinery, New York, NY, 395–404. DOI: https://doi.org/10.1145/2911451.2911498

[47] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural Ad-Hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. Association for Computing Machinery, New York, NY, 55–64.

[48] Jing Yao, Zhicheng Dou, and Ji-Rong Wen. 2020. Employing personal word embeddings for personalized search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, 1359–1368. DOI: https://doi.org/10.1145/3397271.3401153

[49] Sevgi Yigit-Sert, Ismail Sengor Altingovde, Craig Macdonald, Iadh Ounis, and Özgür Ulusoy. 2020. Supervised approaches for explicit search result diversification. *Inf. Process. Manag.* 57, 6 (2020), 102356. DOI: https://doi.org/10.1016/j.ipm.2020.102356

[50] Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2020. Encoding history with context-aware representation learning for personalized search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, 1111–1120. DOI: https://doi.org/10.1145/3397271.3401175

[51] Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2020. Enhancing re-finding behavior with external memories for personalized search. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM'20)*. Association for Computing Machinery, New York, NY, 789–797. DOI: https://doi.org/10.1145/3336191.3371794

[52] Yujia Zhou, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2021. *PSSL: Self-supervised Learning for Personalized Search with Contrastive Sampling*. Association for Computing Machinery, New York, NY, 2749–2758. DOI: https://doi.org/10.1145/3459637.3482379

[53] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu. 2014. Learning for search result diversification. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)*. Association for Computing Machinery, New York, NY, 293–302. DOI: https://doi.org/10.1145/2600428.2609634

[54] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web (WWW'05)*. Association for Computing Machinery, New York, NY, 22–32. DOI: https://doi.org/10.1145/1060745.1060754

Received 18 June 2022; revised 8 July 2023; accepted 17 October 2023