Incorporating Explicit Subtopics in Personalized Search

Shuting Wang
Zhicheng Dou*
wangshuting@ruc.edu.cn
dou@ruc.edu.cn
Renmin University of China
Beijing, China

Jing Yao
jingyao@microsoft.com
Social Computing Group
Microsoft Research Asia
Beijing, China

Yujia Zhou
Ji-Rong Wen†
zhouyujia@ruc.edu.cn
jrwen@ruc.edu.cn
Renmin University of China
Beijing, China

ABSTRACT

The key to personalized search is modeling user intents to tailor returned results for different users. Existing personalized methods mainly focus on learning implicit user interest vectors. In this paper, we propose ExpliPS, a personalized search model that explicitly incorporates query subtopics into personalization. It models the user's current intent by estimating the user's preference over the subtopics of the current query and personalizes the results over the weighted subtopics. We think that in such a way, personalized search could be more explainable and stable. Specifically, we first employ a semantic encoder to learn the representations of the user's historical behaviours. Then with the historical behaviour representations, a subtopic preference encoder is devised to predict the user's subtopic preferences on the current query. Finally, we rerank the candidates via a subtopic-aware ranker that prioritizes the documents relevant to the user-preferred subtopics. Experimental results show our model ExpliPS outperforms the state-of-the-art personalized web search models with explainable and stable results.

CCS CONCEPTS

• **Information systems** → **Personalization**.

KEYWORDS

personalized search, explicit query subtopics

ACM Reference Format:
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 2023 (WWW '23), May 1–5, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583488

1 INTRODUCTION

Nowadays, search engine is an important way to obtain information from the Web. To whoever enters the same query, existing search engines usually return the same ranked document list, solely based on the relevance of the candidate documents to the query. However, queries are often short and ambiguous (*e.g.*"java"), and may contain multiple specific meanings (*e.g.*"java language programming" and "java island"). Different users issuing the same query may have varying intents due to their diverse backgrounds and interests. Returning the same results to all users under such a condition will make some users unsatisfied. Personalized search is one of the mainstream solutions. Its central idea is to model a user's interest from her search histories, and tailor result rankings based on how satisfied the candidate document is with the user's interest.

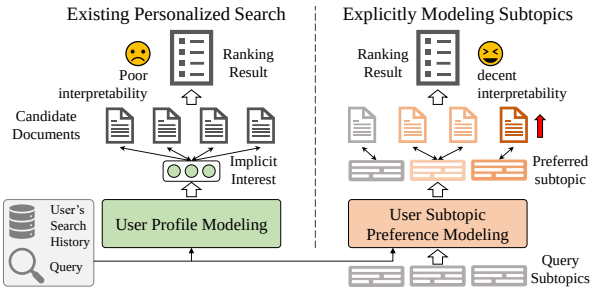


Figure 1: Comparison of previous personalized search models and our method incorporating explicit subtopics.

Till now, many personalized search models have been proposed. They employ either unsupervised methods [3, 9, 10, 12, 15, 30, 39, 40] or neural networks [17, 23, 24, 44, 48–50] to learn an implicit representation of user interests from her search histories. Then the original search results are reranked by comparing the similarity between the learned user interest vector and each candidate document. Although these methods are proved to be effective, they have a common problem: the inference process likes a "black box", which is reflected from two angles. **First**, learning user interests by simply aggregating user histories makes it hard to control the selection of valuable historical behaviours. Some noisy historical behaviours are irrelevant to either the user's long-term interests or her current intent. Although some recent studies [17] use the current query to weight historical behaviours via the attention mechanism, deciding what information should be kept or removed based on the ambiguous issued query remains problematic. **Second**, most studies model user interest in the form of representation vectors. With this implicit interest representation, it is hard to explain the real intent the user has, and which specific information the user is looking for.

In this paper, we want to go beyond the implicit user interest representation, and propose to leverage query intents in an explicit manner. As there might be multiple different information needs under the same query, we adopt the term "subtopics", which is widely used in search result diversification [1, 22, 28], to stand for

*Zhicheng Dou is the corresponding author.
†Also with Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education.

WWW '23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583488

various user intents by the query. We propose introducing subtopics into personalized search, and explicitly model the user's subtopic preferences of the current query to identify her real intent. Results are thereby ranked over the personalized subtopics: the documents relevant to the subtopics the user prefers will be ranked higher. This will make the personalization more explainable because we can know why some results are ranked higher and which subtopics they are relevant to. It also potentially improves the ranking stability since results are personalized based on the query subtopics, and the noise in the historical behaviours is harder to influence the ranking. We present the comparison of our method with existing personalized methods that employ implicit interests in Fig. 1.

Specifically, we propose a personalized search framework, namely ExpliPS, to explicitly incorporate subtopics to understand user intents and tailor search results. To exploit query subtopics reasonably, we further provide a derivation function to compute subtopic-based ranking scores. This derivation guides the construction of our model. Concretely, ExpliPS contains three core modules, *i.e.*, a semantic encoder, a subtopic preference encoder, and a subtopic-aware personalized ranker. Its workflow is as follows: **Firstly**, the **semantic encoder** is employed to produce the basic representation of a query, a candidate document, or a subtopic. **Secondly**, given the user's histories, the **subtopic preference encoder**, a transformer-encoder-based module, is applied to predict the user's preference over subtopics of the current query. Based on our derivation, we devise a subtopic-aware attention masking for this module to capture reliable preference signals. **Finally**, the **subtopic-aware personalized ranker** estimates the personalized ranking score of each document. It computes the similarity scores of documents with each subtopic and weights them up by subtopic preferences.

Nevertheless, the above pipeline is challenging to predict precise subtopic preferences, since we have no ground truth for this task. To solve this problem, a pseudo label model is devised to extract the pseudo labels of users' subtopic preferences from their click feedback for training data. We design an auxiliary task to narrow the gap between predicted preferences and pseudo labels. It helps estimate the user intents more accurately, and boosts the final ranking quality. Thus, the training objectives of our model consist of the personalized ranking task and the preference prediction task.

Note that the core idea of the paper is explicitly modeling user-preferred subtopics in personalized search. The primary goal is improving ranking quality, and subtopic mining is beyond our scope. We simply take google suggestions, which are widely used in search result diversification [19, 20, 22, 25, 28], as query subtopics. Nevertheless, our work has nothing to do with diversification: we optimize solely for personalized ranking. It is the first time that explicit subtopics are used together with the state-of-the-art Transformer based user profiling and personalized web search models.

We experiment with the AOL search log dataset. Experimental results show that our model can significantly outperform the existing personalized models with stable and explainable results.

Our main contributions in the paper are:

(1) We propose to exploit subtopics to explicitly model user intents for personalized search and develop a personalized model, ExpliPS. Though subtopics are widely used in search result diversification, this is the first time to be employed in personalized search.

(2) According to our derivation of calculating personalized ranking scores based on subtopics, we adopt a subtopic-aware attention masking mechanism to the subtopic preference encoder for modeling reliable preference signals.

(3) We design an auxiliary model to take intent signals from user feedback as pseudo labels, which guides the prediction of the subtopics preferences.

2 RELATED WORK

2.1 Personalized Search Models

Personalized search has been a popular topic due to the effectiveness of customizing results for different users. Traditional models mainly rely on heuristic rules [15, 35] and manual features [4, 38, 40]. For example, Thanh et al. [39] employed Latent Dirichlet Allocation (LDA) [6] to construct user profiles in the topic space. With the appearance of the advanced learning to rank model LambdaMART [8], some supervised personalized search models [5, 38, 41] achieved significant improvement over the previous unsupervised ones.

Recently, many neural personalized search models [17, 23, 24, 44, 47–50] are proposed. They have shown a superior ability to model implicit user interests in high-dimension feature space. The earliest one among them, HRNN was proposed to build dynamic user profiles by sequentially modeling the user histories via a hierarchical RNN with Query-aware Attention [17]. After that, researchers have employed many different techniques, such as generative adversarial network [23], reinforcement learning [45], contrastive learning [50], etc., to improve the quality of user profiling.

Although these models are proved to be effective in facilitating the search results personalization, they commonly pay attention to modeling the dense vector of user interests. Most of them neglected the user's preference on explicit query subtopics, which could improve the robustness and interpretability of personalized search. In this paper, we make a preliminary exploration of this direction.

2.2 Application of Subtopics

Previous studies [14, 15, 32] revealed that an ambiguous query can be decomposed into multiple intents or subtopics. A user is usually interested in some subtopics but dislikes the left. Thus, understanding query intent and mining subtopics play a crucial role in solving the query ambiguity problem. Numerous works have developed various ways to mine query subtopics [11, 16, 18, 34, 43, 46]. However, mining subtopics is not the focus of our paper. We simply use google suggestions as subtopics, which is a common way in other researches [19, 20, 22, 25, 28].

Subtopics have been extensively used in search result diversification [1, 19, 20, 25, 28], which aims to diversify ranking results by covering aspects of issued queries as much as possible. However, the goal of personalization is modeling user interests to customize results for each user. These two tasks have different learning targets.

At present, there is a lack of methods that consider explicit query subtopics to enhance personalization. Note that there exists some traditional personalized methods [29, 31, 33, 41] that learn the user's profiles based on topical category or ontology. Nevertheless, these are not the subtopics we discussed in this paper. Several studies [26, 36] introduce explicit diversification into personalized search, while they mainly aim at enhancing the diversity of personalized results,

which is different from the target of our work: we optimize solely for personalized ranking.

3 PRELIMINARIES

Personalized search has been an effective way to provide users with desired search results. However, existing works mostly focus on modeling implicit user interests, which may cause unstable and unexplainable results. As we stated in Section 1, explicitly considering the subtopics is conducive to clarifying the user's specific intents, hence improving the quality and interpretability of the ranking results. Consequently, we propose a personalized framework that explicitly incorporates query subtopics to comprehend user intent and personalize results. We first define the problem and provide some preliminary derivations as below.

3.1 Problem Definition

When a user u enters a query q in the search box, the search engine will first return a batch of candidate documents D. A personalized search model is supposed to rerank the candidates based on current query q and the user's historical search behaviours H. Specifically, $H = \{(q_i, D_i)|i \in [1, n]\}$, where n is the amount of historical issued queries, q_i is the i-th query and D_i denotes its clicked document set. Previous studies [17, 48, 50] calculate the candidate d's final score, where $d \in D$, by using an aggregator agg() to combine the personalized score $\mathrm{Ps}(d|q,u)$ and the ad-hoc score $\mathrm{Rs}(d|q)$:

$$\mathrm{score}(d|q,u) = \mathrm{agg}\left(\mathrm{Ps}\left(d|q,u\right), \mathrm{Rs}(d|q)\right), \tag{1}$$

The personalized score is mostly computed by the similarity between the document and implicit user interests. We call it implicit score, and denote it as s^{im}, *i.e.*, $\mathrm{Ps}\left(d|q,u\right) = \mathrm{s}^{\mathrm{im}}$.

Generally, the personalized score can be viewed as the approximation of the probability that document d is relevant given the current user and query, *i.e.*, $\mathrm{Ps}\left(d|q,u\right) \propto p(d|q,u)$.[1] In previous, s^{im} is an **implicit approximation** as its production involves no subtopic. In this work, we introduce an explicit score s^{ex}, which **explicitly approximates** $p(d|q,u)$ by considering query subtopics. **Thus, the s^{ex} and s^{im} approximate $p(d|q,u)$ from explicit and implicit perspectives**, *i.e.*, $\mathrm{s}^{\mathrm{ex}} \propto p(d|q,u)$ and $\mathrm{s}^{\mathrm{im}} \propto p(d|q,u)$.

Therefore, suppose $Q_s = \{s_i|k \in [1, k]\}$ is the current query's subtopic set obtained by google suggestion, we convert Eq. (1) into:

$$\mathrm{score}(d|q,u) = \mathrm{agg}(\mathrm{Ps}(d|q,u,Q_s), \mathrm{Rs}(d|q)), \tag{2}$$

where $\mathrm{Ps}(d|q,u,Q_s)$ is the personalized score that explicitly considers query subtopics. We use two learnable parameters α, β to linearly combine $\mathrm{s}^{\mathrm{ex}}, \mathrm{s}^{\mathrm{im}}$ for ensuring $\mathrm{Ps}(d|q,u,Q_s) \propto p(d|q,u)$.

$$\mathrm{Ps}(d|q,u,Q_s) = \alpha \mathrm{s}^{\mathrm{ex}} + \beta \mathrm{s}^{\mathrm{im}}. \tag{3}$$

Next, we will provide the explicit derivation of $p(d|q,u)$, *i.e.*, the estimation methods of s^{ex}, by introducing subtopic variables.

[1]Since the ranking task emphasizes relative values between scores rather than absolute values, it's unnecessary to limit ranking scores to the interval of $[0, 1]$, but proportional to the probability. Thus, the approximation of this paper is proportionate.

3.2 Explicit Derivation

Without loss of generality, $p(d|q,u)$ can be derived by introducing variables of query subtopics, s_i, as follows:

$$p(d|q,u) = \sum_{s_i \in Q_s} p(d, s_i|q,u) = \sum_{s_i \in Q_s} p(d|s_i,q,u)p(s_i|q,u). \tag{4}$$

We view $p(s_i|q,u)$ as the user u's preference degree on subtopic s_i given query q. $p(d|s_i,q,u)$ is the relevance of d to the subtopic s_i when u issues q. Moreover, the semantic range of query subtopics are contained by the one of current query, which means that $p(s_i,q) = p(s_i)$, hence $p(d|s_i,q,u) = p(d|s_i,u)$.[2] Furthermore, assuming that the user u is independent of the query q, the $p(s_i|q,u)$ can be simplified as below,

$$p(s_i|q,u) = \frac{p(s_i,q,u)}{p(q,u)} = \frac{p(s_i,u)}{p(q)p(u)} = \frac{p(s_i|u)}{p(q)} \propto p(s_i|u), \tag{5}$$

where $p(q)$ is constant for a certain search scenario. Therefore, the explicit derivation of $p(d|q,u)$ can be represented as,

$$p(d|q,u) \propto \sum_{s_i \in Q_s} p(d|s_i,u)p(s_i|u). \tag{6}$$

This derivation can provide guidance for our model construction, which will be demonstrated in Section 4.

4 SUBTOPIC-AWARE PERSONALIZATION

In this section, we provide the details of our model, which is guided by the *derivation* to provide reliable personalized ranking results. The structure of our proposed ranking model is displayed in Fig. 2. First, the semantic encoder is applied to yield the embeddings of historical behaviours, the current query, subtopics, and candidate documents. Then we employ the subtopic preference encoder to predict the user's explicit subtopic preferences and implicit interests from her historical behaviours. Finally, the subtopic-aware personalized ranker computes the ranking score of candidates by explicitly incorporating user's preferences on query subtopics.

The details of each step are depicted as follows.

4.1 Semantic Encoder

Previous studies [44, 48] illustrated that users' historical search behaviours (*i.e.*, queries and clicked documents) are favorable for inferring their interest and current intent. Most personalized search models take the behaviour representations as input to learn user interests. In this paper, we adopt transformer encoder [37] to embed the behaviours into semantic representations, the effectiveness of which has been proved in recent works [25, 37, 48].

Specifically, take the i-th historical behaviour (q_i, D_i) for example. We concatenate the word embeddings of the query q_i and clicked documents D_i with "[SEP]" as the separator to construct the input sentence, *i.e.*, S_i.

$$S_i = q_i[\mathrm{SEP}]d_{i,1}[\mathrm{SEP}]...[\mathrm{SEP}]d_{i,p}, \tag{7}$$

where $d_{i,j} \in D_i \ \forall j \in [1, p]$ and p is the number of clicked documents. Further, type embeddings are introduced for distinguishing

[2]Though there may be a few subtopics containing the same contents under different queries, we view them as different subtopics to ensure the correctness of formulas.

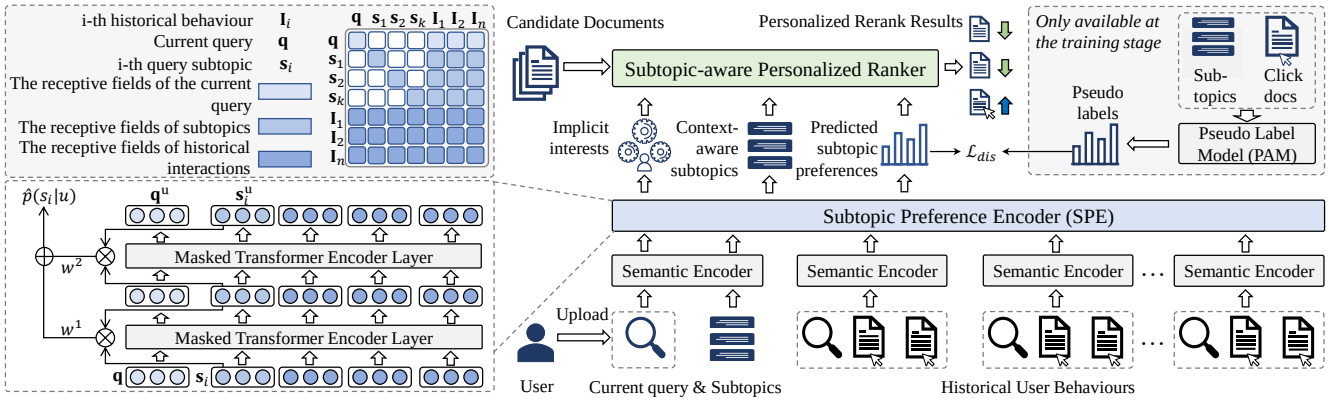


Figure 2: The architecture of our proposed framework, where candidate embeddings are also yielded by the semantic encoder, but we have no space to display it. The up-right part is used to produce pseudo subtopic labels for subtopic preferences. The up-left part depicts the attention masking employed by our model. The bottom-left shows the prediction of subtopic preference.

the different types of inputs. We represent the type embedding sentence by T_i. Thus, we encode the i-th historical behaviour by

$$\mathbf{I}_i = \text{Avg}\left(\text{Trm}\left(S_i + T_i\right)\right), \tag{8}$$

where Trm() denotes the transformer encoder with L layers and Avg() is average pooling. $\mathbf{I}_i$ denotes the representation of user intents at i-th historical behaviour.

Meanwhile, we employ the same structure to yield representations of the current query, a candidate document, and a subtopic, and get $\mathbf{q}$, $\mathbf{d}$, $\mathbf{s}_i$, respectively.

4.2 Subtopic Preference Encoder

Having representations of the user's historical behaviours, we need a higher-level structure to capture the user's preference signals from them to clarify the user's current intent. Previous works [17, 48, 50] implemented it by modeling the implicit user interest. Though it is beneficial for personalization, solely exploiting implicit interest is sensitive to noisy histories and hard to provide explainable user intents. Therefore, we devise a subtopic preference encoder (SPE) to capture the user's explicit subtopic preference and implicit interest, hence providing clear and precise user intents.

Furthermore, since we design our ranking model based on the *derivation*, the document relevance to subtopics, $p(d|s_i, u)$, dependents on the current user. Thus the $p(d|s_i, u)$ should be estimated in the context of the user. In other words, the representations of subtopics should consider the user's historical behaviors, hence estimating $p(d|s_i, u)$ by their similarity with candidate documents. We call these subtopic representations context-aware subtopics.

Consequently, the SPE module takes embeddings of the current query, subtopics, and historical behaviours as input and produces the explicit subtopic preferences, implicit interests, and context-aware subtopics for the subsequent processes.

4.2.1 Context-aware subtopics and implicit interests. Since the transformer encoder [37] can refine the input representations by relevant contextual information, we base it to construct the SPE. However, the full attention is inapposite to the SPE as there are some invisible requirements between different input terms according to our derivation of Eq. (6). Thus, we propose an attention masking mechanism, namely subtopic-aware mask, to capture more reliable representations of inputs by the following rules.

Mask between the current query and subtopics. As we introduced in Section 3.2 that Eq. (4) and (6) are both reasonable for our model to compute the explicit personalized score. The difference is that for the former, the context to learn the subtopic representations is the current query q and the user histories u, while the latter is only the user histories. In this paper, we follow the latter and introduce the attention mask on the current query for its subtopics. The reason is that the information flow from the query to its subtopics is useless because the query contains similar but more general information than subtopics. If the query is visible to subtopics, it will dominate the subtopics' attention and restrain the attention of user histories. Similarly, we mask subtopics for the query to capture the implicit user interests that involve no subtopics.

Mask within subtopics. Note that we estimate the relevance of document d to subtopic s_i, *i.e.*, $p(d|s_i, u)$, via the similarity between the document and the subtopic's context-aware representation. Since $p(d|s_i, u)$ involves none of other subtopics, subtopics should be invisible to each other when learning their context-aware representations. Our qualitative analysis of this rule is that subtopics represent distinct aspects of the current query, while interactions within them may fuzzy the discrepancy between their representations, which is harmful to subtopics' representativeness.

The masked attention matrix is visualized in the up-right part of Fig. 2. We build an input sequence based on the representations of the current query, its subtopics, and historical behaviours, then fed it into the transformer encoder with this mask strategy, which is called masked transformer encoder. The output of the query is the user's implicit interest, $\mathbf{q}^u$, *i.e.*,

$$\mathbf{q}^u = \text{MaskTrm}^{n+1,}\left(\left[\mathbf{I}_1, ..., \mathbf{I}_n, \mathbf{q}, \mathbf{s}_1, ..., \mathbf{s}_k\right]\right), \tag{9}$$

where MaskTrm() denotes the masked transformer encoder with L layers, the superscript, *e.g.*, $n+1$, is the output index. The outputs of subtopics are their context-aware representations, $\mathbf{s}_i^u, i \in [1, k]$.

$$\mathbf{s}_i^u = \text{MaskTrm}^{n+1+i}\left(\left[\mathbf{I}_1, ..., \mathbf{I}_n, \mathbf{q}, \mathbf{s}_1, ..., \mathbf{s}_k\right]\right), \tag{10}$$

4.2.2 Explicit subtopic preferences. To generate the user's specific intent, we need to predict the user's preference distribution on the subtopics of the current query, *i.e.*, $\hat{p}(s_i|u)$. With the assumption that the user's preference on a subtopic is reflected by whether the user has searched the information related to the subtopic, we propose a preference predictor based on the feature of the transformer encoder. Specifically, if the user has no relevant histories for the subtopic s_i, the transformer layer will gather a lot of noisy information from the user histories and change the semantic representation of the subtopic; otherwise, it will aggregate related histories and keep semantic. Therefore, the semantic difference between subtopic representations before and after masked transformer layers can be used to derive the user's subtopic preference, *i.e.*,

$$\hat{p}(s_i|u) = \text{softmax}_i \left(\sum_{l=1}^{L-1} w^l f \left(\mathbf{s}_i^{u,l}, \mathbf{s}_i^{u,l+1} \right) \right), \tag{11}$$

where $\mathbf{s}_i^{u,l}$ is the representation of the subtopic s_i at the l-th layer, w^l is the trainable parameter denoting the weight of the l-th layer, and $f()$ is the similarity function, which is implemented based on a layer-specific matrix $\mathbf{W}^l$, *i.e.*, $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{W}^l \mathbf{y}$. We show this process in the bottom-left part of Fig. 2, which takes the $L = 2$ as an example and visualizes the subtopic s_i as a representative.

4.3 Subtopic-aware Personalized Ranker.

Based on Eq. (6), we construct a subtopic-aware personalized ranker to produce the final scores of candidate documents by explicitly incorporating query subtopics. Specifically, the s^{ex} is computed by,

$$\mathbf{s}^{\text{ex}} = \sum_{i=1}^{k} \hat{p}(s_i|u) \text{sim}(\mathbf{d}, \mathbf{s}_i^u), \tag{12}$$

where sim() is the cosine similarity function. $\text{sim}(\mathbf{d}, \mathbf{s}_i^u) \propto p(d, s_i, u)$ as the generation of $\mathbf{s}_i^u$ involves the user histories. Moreover, considering $p(d|s_i, u) = \frac{p(d,s_i,u)}{p(s_i,u)}$, where $p(s_i, u)$ is same for all documents of the current query q, we can assume that $\text{sim}(\mathbf{d}, \mathbf{s}_i^l) \propto p(d|s_i, u)$. This assumption is also applicable to later calculations.

To avoid the incomplete coverage of subtopics on query aspects, we retain the implicit score, which has been verified to be effective in personalized search. We produce the s^{im} by the following function,

$$\mathrm{s}^{\text{im}} = \text{sim}(\mathbf{q}^u, \mathbf{d}) \tag{13}$$

The personalized score $\text{Ps}(d|q, u, Q_s)$ is yieled by Eq. (3).

Following previous personalized search methods [44, 48–50], we calculate ad-hoc score, $\text{Rs}(d|q)$ by

$$\text{Rs}(d|q) = \mathbf{w}^a \left[\text{KNRM}(q^w, d^w); \text{sim}(\mathbf{q}, \mathbf{d}); \phi(f_{qd}) \right]^T, \tag{14}$$

where q^w, d^w denote the word embeddings of the current query q and candidate d. KNRM() is the word-level cross-matching function proposed by [42]. f_{qd} are the relevance features between the document and the query extracted following previous studies [50]. We utilize a trainable weight $\mathbf{w}^a \in \mathbb{R}^{1\times 3}$ to combine them.

Eventually, an MLP layer $\phi()$ is used to generate the final score,

$$\text{score}(d|q, u, Q_s) = \phi \left([\text{Ps}(d|q, u, Q_s); \text{Rs}(d|q)] \right). \tag{15}$$

Based on the final scores, we can rerank the candidate documents by explicitly considering the user's subtopic preference, and enhancing the quality and interpretability of the ranking results.

4.4 Auxiliary Subtopic Preference Prediction

The above sections described the architecture of our ranking model which explicitly considers query subtopics for personalized search. Nevertheless, the only label data available for training the ranking model is the user's implicit feedback (*e.g.*, the candidates' click label of the current query), without signals of the user's explicit subtopic preferences. Optimizing our model solely based on the learning-to-rank (LTR) task would make it difficult to ensure the accurate prediction of the user's subtopic preferences. Motivated by this, we construct a pseudo label model (PAM) to extract the user's preferences signals from her click feedback. We view the extracted preferences as the pseudo label to assist the training of our ranking model, which will be introduced in Section 4.5.

Considering that unsupervised methods are hard to capture the high-order semantic information of the user's feedback, which is detrimental to the accuracy of extracted pseudo labels, we construct the PAM in a supervised manner. Next, we will demonstrate the extracting process and the training task of PAM.

4.4.1 Extracting process of pseudo labels. To learn the semantic information from click feedback, we devise the PAM based on the transformer encoder. Specifically, the input of PAM consists of the clicked documents of the training query. We denote the word embedding sentence as S, and adopt the transformer encoder Trm() to learn the user's current implicit intent, $\mathbf{I}$ as below,

$$\mathbf{I} = \text{Avg}(\text{Trm}(S)). \tag{16}$$

We view $\mathbf{I}$ as the accurate user intent representation since it is derived from her click feedback. Thus, user's subtopic preferences can be directly decoded by,

$$p(s_i|u) = \text{softmax}_i \left(\text{sim} \left(W^i \mathbf{I}, W^s \mathbf{s}_i \right) \right). \tag{17}$$

W^i and W^s are learnable parameters and $p(s_i|u)$ is the pseudo label.

4.4.2 Training of the PAM. Since final scores computed based on accurate user preferences can yield high-quality ranking results, the quality of the derived ranking results is evidence of the accuracy of the user preferences. Thus, we rerank candidate documents based on extracted preference signals and optimize the PAM by LTR task. Considering the click feedback maximum reflects the user intent, the PAM can capture users' subtopic preferences more accurately than SPE. Thus, candidates can be ranked based on their relevance with preferred subtopics, which denote the user's actual intents. In this ideal scenario, the relevance of d is conditionally independent of u given the s_i, *i.e.*, $p(d|s_i, u) \sim p(d|s_i)$. By removing the demand for context-aware subtopics, this derivation allows for a reduction in model parameters and an increase in training convergence. Therefore, we compute the explicit score s^{ex} by,

$$\mathrm{s}^{\text{ex}} = \sum_{i=1}^{k} p(s_i|u) \text{sim}(\mathbf{d}, \mathbf{s}_i), \tag{18}$$

where $\text{sim}(\mathbf{d}, \mathbf{s}_i) \propto p(d|s_i)$. We yield the implicit score by $\mathrm{s}^{\text{im}} = \text{sim}(\mathbf{I}, \mathbf{d})$, and produce the personalized score by Eq. (3). The ad-hoc score is generated in the same way as Eq. (14). Then, we calculate the final score of the candidate document by Eq. (15).

We apply a pairwise LTR algorithm, LambdaRank [7] to optimize the PAM. The training sample is constructed as follows:

$x = \{q, Q_s, d_i, d_j, D^+\}$, $y_{ij} \in \{0, 1\}$. d_i and d_j denote the pair of positive and negative documents, y_{ij} is the ground truth that d_i is more relevant than d_j. Q_s is the query's subtopics and D^+ represents the clicked documents. The loss function is computed as below:

$$\mathcal{L}(d_i, d_j) = -y_{ij} \log(\overline{p_{ij}}) + (1 - y_{ij})(\log(\overline{p_{ij}})), \tag{19}$$

where $\overline{p_{ij}} = \text{sigmoid}(\text{score}(d_i) - \text{score}(d_j))$ is the predicted probability that d_i is better than d_j. $\text{score}(d_i)$ and $\text{score}(d_j)$ are the final score of d_i and d_j generated by PAM.

4.4.3 Auxiliary Subtopic Prediction Task. With the well-trained PAM introduced above, we can extract the pseudo labels of users' subtopic preferences for assisting us in ranking model training. We devise an auxiliary task for narrowing the gap between the predicted user's subtopic preferences produced by Eq. (11), $\hat{p}(s_i|u)$, and the pseudo label. Due to the symmetry of JS divergence, we adopt it to devise the following loss function:

$$\mathcal{L}_{dis} = \text{JS}\left(\hat{p}(s_i|u), p(s_i|u)\right). \tag{20}$$

4.5 Ranking Model Training

The LTR is the main objective of our model to ensure the ranking quality. For model optimization, we utilize the same pairwise loss function as PAM, *i.e.*, Eq. (19) . The difference is that the compositions of training samples are $x = \{q, Q_s, H, d_i, d_j\}$, $y_{ij} \in \{0, 1\}$, and the final scores of d_i and d_j are produced by our ranking model based on Eq. (15). We denote the LTR loss as $\mathcal{L}_{rank}$.

The final loss of our model is generated by the linear combination of $\mathcal{L}_{rank}$ and $\mathcal{L}_{dis}$ with manually defined weights λ_1 and λ_2:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rank} + \lambda_2 \mathcal{L}_{dis}, \tag{21}$$

5 EXPERIMENT

5.1 Dataset and Evaluation Metrics

The experiments are conducted on a publicly available dataset, AOL that contains a search log from 1^{st} March 2006 to 31^{st} May 2006. The basic statistic information is presented in Table 1. We construct the dataset following Wasi et al. [2], where all non-alphanumeric characters of queries are removed. The session boundaries are identified by the similarity between two consecutive queries. Based on the above process, each piece of data consists of an anonymous user ID, a session ID, a query, the uploaded time of the query, and the corresponding clicked URLs. As AOL only contains clicked documents, which are viewed as relevant ones, we follow [2] to crawl document content, dig out candidate documents and reconstruct the original rank list by BM25 algorithm [27]. 5 candidate documents are built for the queries in training and validation sets, and 50 candidates for the queries in the test set. The detailed processes can be referred to at [2]. We view clicked/unclicked documents as positive/negative samples following existing studies [44, 48, 50] to build training samples. For ensuring an adequate search log for each user, the first five weeks are reserved as a background set, and we view the latest eight weeks as an experimental set, which is divided into the training set, validation set, and test set in a 4:1:1 ratio. Following [2], we regard the titles of documents as the content.

We select three widely used metrics, MAP, MRR, and P@K to evaluate the performance of our proposed model. For P@K, we further consider P@1, P@3, and P@5 in our experiments. Meanwhile, due to the phenomenon that the original rank position of documents will impact the click behaviour of the user, namely position bias, we further select a more reliable metric P-Improve [23]. According to [13, 21, 23], we only view the skipped and next non-clicked documents as irrelevant documents and construct inverse document pairs. Thus, P-Improve is obtained by computing the ratio of the correctly ranked inverse pairs.

Table 1: Basic statistical information of the dataset.

Item	Value	Item	Value
# days	91	Avg. query length	2.87
# users	110,439	Avg. session length	2.55
# queries	736,454	Avg. #click per query	1.11
# sessions	279,930	Avg. #subtopic per query	5.22

5.2 Baselines

Except for the original ranking, we select several popular ad-hoc models and personalized models to compare with our model.

KNRM[42] is a kernel-based method to learn the soft matching between the tokens of the document and the query. **P-Click** [15] is a heuristic personalization method that reranked the candidates of the refinding queries based on the personal click number. **SLTB** [5] used LambdaMART to yield personalized rerank results with 102 features extracted from the search log. **HRNN** [17] is a deep-learning-based model which applied the hierarchical RNN and attention mechanism to learn user profiles. **PSGAN** [23] focused on exploring high-quality negative examples based on the generative adversarial network. **RPMN** [49] constructed the personalized model based on memory networks to model the multi-level refinding. **PEPS** [44] proposed to enhance personalization by constructing a personalized embedding matrix for each user. **HTPS** [48] applied the transformer encoder to integrate the search history for disambiguating the current query. **PSSL** [50] devised four contrastive tasks to pretrain the personalized model, leading to reliable representations of user profiles, queries, and documents to improve the personalized search. It is the state-of-the-art personalized search model.

Our methods includes **ExpliPS & ExpliPS-S5.** To demonstrate the robustness of the model for the methods of subtopic mining, we implement two subtopic choice approaches. One is called ExpliPS, where all the google suggestions are treated as query subtopics. And the other randomly selects five google suggestions as query subtopics, the corresponding model called ExpliPS-S5. We provide the implement details of our method in Appendix .1.

5.3 Overall Performance

We first compare our model with all baselines, the overall results are shown in Table 2. Our observations and analysis are as follows:

(1) **Comparing with all the baselines, our model ExpliPS significantly surpasses them in terms of all the metrics with paired t-test at $p < 0.05$ level.** Especially for the state-of-the-art (SOTA) personalized model, PSSL, our model achieves a 3.00% improvement on MAP, and 3.03% improvement on MRR. Additionally, 5.60% promotion on P@1 implies that our model can accurately find the document that meets the user's need and rank it atop. Further,

Table 2: Overall performances of all models. "†" indicates that the model outperforms the state-of-the-art baseline, *i.e.*, PSSL, significantly with paired t-test at $p < 0.01$ level. The best results are shown in bold.

Model	MAP	MRR	P@1	P@3	P@5	P-Imp
Adhoc search model						
Original	.2504	.2596	.1534	.2865	.3522	-
KNRM	.4298	.4399	.2718	.5130	.6089	.6633
Previous personalized search model						
P-Click	.4221	.4305	.3780	.4128	.4431	.1657
SLTB	.5113	.5237	.4693	.5244	.5507	.3374
HRNN	.5438	.5555	.4841	.5663	.6042	.5927
PSGAN	.5480	.5601	.4892	.5741	.6190	.5985
RPMN	.5926	.6049	.5322	.6333	.6858	.6586
HTPS	.7091	.7251	.6268	.7728	.8300	.7730
PEPS	.7127	.7258	.6279	.7902	.8467	.8105
PSSL	.7359	.7484	.6431	.8248	.8805	.8278
Personalized search model incorporating query subtopics						
ExpliPS-S5	.7498†	.7627†	.6649†	.8325†	.8814	.8503†
ExpliPS	**.7580**†	**.7711**†	**.6791**†	**.8384**†	**.8860**†	**.8517**†

the result that ExpliPS achieves a 2.88% improvement on P-Improve over PSSL verifies its effectiveness from a more credible perspective. These comparison results confirm that considering query subtopics and user preference distribution explicitly is beneficial for improving personalized search quality.

(2) **Another version of our model, ExpliPS-S5, significantly outperforms all the baselines on most evaluation metrics.** We find that ExpliPS-S5 achieves 1.88%, 3.39%, and 2.72% promotion on MAP, P@1, and P-Improve, respectively. It indicates the stability of our model for different subtopic selections. The improvement on P@5 is trivial, the reason might be that top-5 is a relatively relaxed condition, thus the baselines can also perform well. Furthermore, though ExpliPS-S5 obtains promising growth, ExpliPS still outperforms it overall. We analyze it because the random selection may omit some important subtopics that the user is interested in, the ranking performance will be affected as a result.

(3) **All personalized models improve the quality of original rankings significantly**, which reveals the effectiveness of personalization for promoting user satisfaction. The results of P-Click and RPMN confirm the importance of refinding behaviour, and SLTB verifies the role of manual relevance features. HRNN and PSGAN take the advantage of long- and short-term history for modeling the user profiles, while PSGAN produces greater results due to the capturing of high-quality negative samples. PEPS and HTPS disambiguate the query representation directly based on the contextual information, and PSSL enhances the personalization by devising multiple contrastive learning tasks. However, all these personalized models focus on learning implicit user interests without considering the explicit subtopic information, thus our model achieves a higher performance when incorporating them simultaneously.

5.4 Ablation Study

To investigate the effectiveness of the components of our model, we design some ablation studies and exhibit the results in Table 3. The design motivations and result discussion are presented as follows.

Table 3: Experimental results of ablation studies.

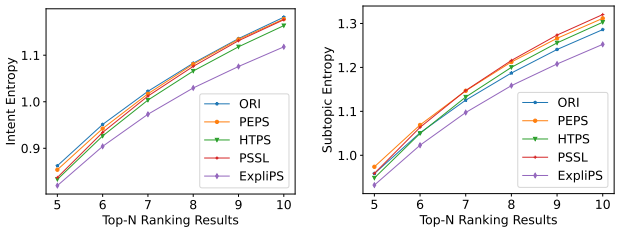
Model	MAP		MRR		P@1	
ExpliPS	.7580	-	.7711	-	.6791	-
w/o EXP	.7237	-4.53%	.7369	-4.44%	.6373	-6.16%
w/o DIS	.7457	-1.62%	.7588	-1.60%	.6640	-2.22%
w/o IMP	.7408	-2.27%	.7540	-2.22%	.6577	-3.15%
w/o PAM	.7425	-2.04%	.7558	-1.98%	.6602	-2.78%
w/o SUP	.7444	-1.79%	.7578	-1.72%	.6636	-2.28%
w/o MASK	.7415	-2.18%	.7544	-2.17%	.6587	-3.00%
PAM	.9408	-	.9514	-	.9146	-

Explicit signal. Our model explicitly considers query subtopics to clarify the user's search intent. To verify its influence on ranking performance, we conduct two variants of our model. The first one drops the explicit score and ranks candidates by aggregating remaining scores, namely "*w/o* EXP". The other one removes the subtopic preference distribution with a uniform distribution, *i.e.*, "*w/o* DIS". The results presented in the 2^{nd} and 3^{rd} lines illustrate that both variants underperform our model, even though the second one retains query subtopics but drops the user's subtopic preferences. This phenomenon proves that explicitly measuring the user's subtopic preferences can capture user intents more accurately, and provide more satisfactory results for users.

Implicit signal. Though we introduce subtopics explicitly, our model still retains the implicit user interest to achieve comprehensive modeling of user intents. To prove its effectiveness, we drop the implicit score s^{im} and construct a variant, "*w/o* IMP". We discover that the "*w/o* IMP" performs 2.14% and 2.58% worse in MAP and P@1 than our model. It implies that the user's implicit interest plays an important role in personalized search. It suggests that explicit and implicit preference signals are coherent and improve the ranking quality together.

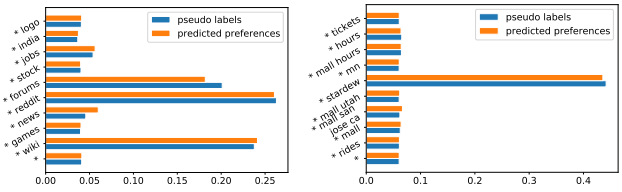
Pseudo label model. To verify the utility of the PAM in ensuring the accuracy of predicted subtopic preferences, we devise "*w/o* PAM" that optimizes the ranking model by the LTR task only. Considering the construction of PAM relies on a supervised manner, we further employ an alternative unsupervised way to extract pseudo labels, which decodes the user's subtopic preferences from the averaging of the clicked documents via cosine similarity. We call this model "*w/o* SUP". From Table 3, the results of both variants decline significantly. It confirms the importance of the PAM and indicates that a well-trained PAM can capture accurate preference signals, hence providing the correct learning direction for the training of ExpliPS. We also illustrate the performance of PAM in the last row, which demonstrates that it can achieve satisfied ranking results, which potentially ensures the accuracy of the pseudo labels.

Subtopic-aware mask. Our derivation of Eq. (6) reveals the rationality of applying the subtopic-aware mask to the transformer structure of SPE. To practically validate this mask strategy, we build a variant of our model, namely "*w/o* MASK", leveraging the full attention strategy for SPE. The degraded performance verifies the effectiveness of our mask strategy that all subtopics and the issued query should be invisible to each other to avoid paying excessive attention to redundant information. Thus, our SPE can capture more personal signals from user histories and provide more reliable representations of queries and subtopics.



(a) The intent entropy of five models (b) The subtopic entropy of five models

Figure 3: The comparison of intent & subtopic entropy.



(a) Uid:277 Qry:paradox entertainment (b) Uid:7886 Qry:valley fair

Figure 4: Visualization of two types of subtopic preferences

5.5 Experiments of Stability of Ranking Results

Since aggregating user histories solely based on the vague query cannot clarify the specific user intent but introduce noisy historical information. It will cause the top results cannot steadily meet the user's intent because the noise information will mislead the model to broadly cover unimportant subtopics rather than focusing on the user's desired subtopics. However, our model explicitly captures the user's subtopic preferences, enhancing the documents related to the preferred subtopics to be ranked ahead. Therefore, top results can cover the user's intent more stably. To confirm it quantitatively, we propose an "Intent Entroy" (IEnt) to measure the coverage stability of the ranking results on the user intent. The lower the intent entropy is, the more stable the user intent is covered. We provide the calculation of the intent entropy in the Appendix A.

The intent entropy of our model with other baselines is illustrated in Fig. 3(a). From the comparison, we find that all personalized models are under the original ranking, which means that the personalized models have the ability to prioritize the documents the user desires. Meanwhile, our model performs the highest stability of user intent coverage with the lowest intent entropy. The results reveal that our model can filter out the noise and focus on the documents that satisfy the user intents. In addition, the intent entropy of all models rises with the increase of N. It is a natural phenomenon as more documents will inevitably cover more subtopics.

Considering that the calculation of IEnt involves the user's subtopic preferences, which are produced by the PAM, the results may be biased toward our model. We additionally develop "subtopic entropy" (SEnt) that excludes user preferences to measure the subtopic coverage of top results. Its computation is also presented in Appendix B. From the results shown in Fig. 3(b), our model is still under other baselines, which implies that ExpliPS focuses on specific subtopics and arranges the documents based on them. The high SEnt of other results might be because their perception of user intent is confused, thus the subtopic coverage of ranking results exhibits a broad and chaotic state.

5.6 Case Study

Previous experiments have validated the performance of our model from multiple perspectives. To further test that our model can predict the user's subtopic preferences accurately, and illustrate the interpretability of the ranking results, we provide two kinds of case studies. The one visualizes some subtopics reference distributions produced from ExpliPS and PAM to prove that our model is able to predict accurate subtopic preferences. The other one provides a ranking result returned by our model to demonstrate its interpretability.

Visualization of subtopic preference distributions. Our model involves two types of subtopic preference distribution, one is the pseudo label yielded by the PAM, and the other one applied to inference is generated by the SPE. We randomly select 2 queries with 10 subtopics each from the test set and visualize their two types of distributions in Fig. 4. The query content is abbreviated as "*" on the Y-axis, which displays the query subtopics. On the X-axis is the degree of preference. From the figures, we notice that there are clear preference variations across the query subtopics, and a few subtopics dominate the user interests. This phenomenon illustrates that users do actually have certain tendencies for certain subtopics when they ask an informative question. This result supports our hypothesis that explicit consideration of the user's subtopic preferences promotes personalization.

The experiment of interpretability is presented in Appendix C.

6 CONCLUSION

In this paper, we propose a personalized model which incorporates query subtopics explicitly to promote the ranking quality. For capturing the user's subtopic preferences, we first use the semantic encoder to learn the representation of every historical behaviour, then we adopt the subtopic preference encoder to capture the user's subtopic preferences and implicit interests. Finally, we apply the subtopic-aware personalized ranker to yield the final scores of candidates by exploiting query subtopics explicitly. Endowed with the benefits of the multi-task loss function, our model could predict the user's subtopic preferences accurately and provide satisfactory results for her. The experiments confirm the effectiveness of our model on search result personalization and interpretability. Due to the time limitation, we only employ google suggestions to represent the subtopics, in future, we will consider exploring diverse subtopic mining methods for explicit search result personalization.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author. This work was supported by the National Natural Science Foundation of China No. 62272467 and No. 61832017, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China NO. 22XNKJ34, and Public Computing Cloud, Renmin University of China. The work was partially done at Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE, and Beijing Key Laboratory of Big Data Management and Analysis Methods.

REFERENCES

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

[2] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context Attentive Document Ranking and Query Suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 385–394. https://doi.org/10.1145/3331184.3331246

[3] Paul N. Bennett, Filip Radlinski, Ryen W. White, and Emine Yilmaz. 2011. Inferring and Using Location Metadata to Personalize Web Search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/2009916.2009938

[4] 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* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 185–194. https://doi.org/10.1145/2348283.2348312

[5] 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* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 185–194. https://doi.org/10.1145/2348283.2348312

[6] David Blei, Andrew Ng, and Michael Jordan. 2001. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3, 601–608.

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany) *(ICML '05)*. Association for Computing Machinery, New York, NY, USA, 89–96. https://doi.org/10.1145/1102351.1102363

[8] Chris J. C. Burges, Krysta M. Svore, Qiang Wu, and Jianfeng Gao. 2008. *Ranking, Boosting, and Model Adaptation*. Technical Report MSR-TR-2008-109. 18 pages. https://www.microsoft.com/en-us/research/publication/ranking-boosting-and-model-adaptation/

[9] Fei Cai, Shangsong Liang, and Maarten de Rijke. 2014. Personalized Document Re-Ranking Based on Bayesian Probabilistic Matrix Factorization. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 835–838. https://doi.org/10.1145/2600428.2609453

[10] Mark J. Carman, Fabio Crestani, Morgan Harvey, and Mark Baillie. 2010. Towards Query Log Based Personalization Using Topic Models. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (Toronto, ON, Canada) *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1849–1852. https://doi.org/10.1145/1871437.1871745

[11] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web Track. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009 (NIST Special Publication, Vol. 500-278)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf

[12] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. 2011. Personalizing Web Search Results by Reading Level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Glasgow, Scotland, UK) *(CIKM '11)*. Association for Computing Machinery, New York, NY, USA, 403–412. https://doi.org/10.1145/2063576.2063639

[13] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-Bias Models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (Palo Alto, California, USA) *(WSDM '08)*. Association for Computing Machinery, New York, NY, USA, 87–94. https://doi.org/10.1145/1341531.1341545

[14] Steve Cronen-Townsend and W. Bruce Croft. 2002. Quantifying Query Ambiguity. In *Proceedings of the Second International Conference on Human Language Technology Research* (San Diego, California) *(HLT '02)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 104–109.

[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* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 581–590. 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. *Information Retrieval Journal* 23 (02 2020). 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* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 347–356. https://doi.org/10.1145/3269206.3271728

[18] Gustavo Gonçalves, Flávio Martins, and João Magalhães. 2018. Analysis of Subtopic Discovery Algorithms for Real-Time Information Summarization. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1855–1856. https://doi.org/10.1145/3184558.3191651

[19] 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* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 63–72. https://doi.org/10.1145/2806416.2806455

[20] 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* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 545–554. https://doi.org/10.1145/3077136.3080805

[21] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 154–161. https://doi.org/10.1145/1076034.1076063

[22] 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* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 479–488. https://doi.org/10.1145/3397271.3401084

[23] 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* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 555–564. https://doi.org/10.1145/3331184.3331218

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

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

[26] 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* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 691–692. https://doi.org/10.1145/1148170.1148320

[27] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (apr 2009), 333–389. https://doi.org/10.1561/1500000019

[28] 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* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 881–890. https://doi.org/10.1145/1772690.1772780

[29] S. Sendhilkumar and T. V. Geetha. 2008. Personalized Ontology for Web Search Personalization. In *Proceedings of the 1st Bangalore Annual Compute Conference* (Bangalore, India) *(COMPUTE '08)*. Association for Computing Machinery, New York, NY, USA, Article 18, 7 pages. https://doi.org/10.1145/1341771.1341790

[30] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit User Modeling for Personalized Search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (Bremen, Germany) *(CIKM '05)*. Association for Computing Machinery, New York, NY, USA, 824–831. https://doi.org/10.1145/1099554.1099747

[31] Ahu Sieg, Bamshad Mobasher, and Robin Burke. 2007. Web Search Personalization with Ontological User Profiles. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (Lisbon, Portugal) *(CIKM '07)*. Association for Computing Machinery, New York, NY, USA, 525–534. https://doi.org/10.1145/1321440.1321515

[32] 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. https://doi.org/10.1145/331403.331405

[33] David Sontag, Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Susan Dumais, and Bodo Billerbeck. 2012. Probabilistic Models for Personalizing Web Search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. Association for Computing Machinery, New York, NY, USA, 433–442. https://doi.org/10.1145/2124295.2124348

[34] Toru Takaki, Atsushi Fujii, and Tetsuya Ishikawa. 2004. Associative Document Retrieval by Query Subtopic Analysis and Its Application to Invalidity Patent Search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (Washington, D.C., USA) *(CIKM '04)*. Association for Computing Machinery, New York, NY, USA, 399–405. https://doi.org/10.1145/1031171.1031251

[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* (Philadelphia, PA, USA) *(KDD '06)*. Association for Computing Machinery, New York, NY, USA, 718–723. 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* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 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* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[38] Maksims Volkovs. 2015. Context Models For Web Search Personalization. *CoRR* abs/1502.00527 (2015). arXiv:1502.00527 http://arxiv.org/abs/1502.00527

[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, Vol. 9022)*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). 605–616. 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* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 323–332. https://doi.org/10.1145/2484028.2484068

[41] 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* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 1411–1420. https://doi.org/10.1145/2488388.2488511

[42] 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* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 55–64.

[43] Takehiro Yamamoto, Yiqun Liu, Min Zhang, Zhicheng Dou, Ke Zhou, Ilya Markov, Makoto P. Kato, Hiroaki Ohshima, and Sumio Fujita. 2016. Overview of the NTCIR-12 IMine-2 Task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, National Center of Sciences, Tokyo, Japan, June 7-10, 2016*, Noriko Kando, Tetsuya Sakai, and Mark Sanderson (Eds.). National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/ntcir/OVERVIEW/01-NTCIR12-OV-IMINE-YamamotoT.pdf

[44] 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, USA, 1359–1368. https://doi.org/10.1145/3397271.3401153

[45] Jing Yao, Zhicheng Dou, Jun Xu, and Ji-Rong Wen. 2020. RLPer: A Reinforcement Learning Model for Personalized Search. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 2298–2308. https://doi.org/10.1145/3366423.3380294

[46] Wei Zheng, Xuanhui Wang, Hui Fang, and Hong Cheng. 2011. An Exploration of Pattern-Based Subtopic Modeling for Search Result Diversification. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (Ottawa, Ontario, Canada) *(JCDL '11)*. Association for Computing Machinery, New York, NY, USA, 387–388. https://doi.org/10.1145/1998076.1998148

[47] Yujia Zhou, Zhicheng Dou, Bingzheng Wei, Ruobing Xie, and Ji-Rong Wen. 2021. *Group Based Personalized Search by Integrating Search Behaviour and Friend Network*. Association for Computing Machinery, New York, NY, USA, 92–101. https://doi.org/10.1145/3404835.3462918

[48] 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, USA, 1111–1120. https://doi.org/10.1145/3397271.3401175

[49] 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* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 789–797. https://doi.org/10.1145/3336191.3371794

[50] 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, USA, 2749–2758. https://doi.org/10.1145/3459637.3482379

.1 Implement Details

We test multiple experiments to determine the hyper-parameter settings. Eventually, The word embedding matrix is initialized with 100-dimension. The hidden size of the transformer structure is 256 and the layer L is 2. For the multi-head attention, we set the head number as 8. To trade-off the effectiveness and efficiency, we select the first relevant for each query as the behaviour information, and the quantity of the subtopic is up to 11. For the queries without google suggestions returned, we consider the query itself to be its subtopic for ensuring the generalization of our model. Then we select the latest 50 interactions for the current query as the historical search log. For the optimization of the PAM, the epoch is set as 2 with a $1e^{-3}$ learning rate. In the training of the whole model, the weights of the two objectives, λ_1, λ_2 are set to 1, 10. The training epoch number is 2 and the learning rate is $5e^{-5}$. Note that the PAM model is pre-trained separately and viewed as a labeling model, which is not trained with our ranking model. For both stages, the AdamW optimizer is used to learn the parameters of the model. We release our code in https://github.com/ShootingWong/ExpliPS.

A CALCULATION OF INTENT ENTROPY

As we introduced in Section 5.5, we need to quantitatively verify the ranking results produced by our personalized model can stably cover the user intents. Thus, intent entropy (IEnt) is devised to measure the coverage stability of the ranking results on the user's preferred subtopics. Specifically, we first determine the coverage of top-N results on each subtopic, *i.e.*, $c(s_i|T_N)$, as below,

$$c(s_i|T_N) = \sum_{j=1}^{N} c(s_i|d_j) \tag{22}$$

$$c(s_i|d_j) = \begin{cases} 1, & \arg\max_k \left(\text{sim}\left(\mathbf{d}_j, \mathbf{s}_k\right)\right) = i \\ 0, & else, \end{cases} \tag{23}$$

where T_N denotes the top-N results and d_j is one of them, *i.e.*, $d_j \in T_N$. $c(s_i|d_j)$ represents whether the document d_j covers the subtopic s_i, and we view the most relevant subtopic as the covered one by the hardmax operation.

Then, we weight it by the user's subtopic preference, $p(s_i|u)$, which is produced by the pseudo label model, resulting in a distribution, $p(s_i|u, T_N)$, whose information entropy represents the stability of user intent coverage. Therefore, the intent entropy, $\text{IEnt}(T_N)$, is calculated as follows,

$$p(s_i|u, T_N) = \text{softmax}_i p(s_i|u) c(s_i|T_N), \tag{24}$$

$$\text{IEnt}(T_N) = -\sum_{i=1}^{k} p(s_i|u, T_N) \log\left(p(s_i|u, T_N)\right), \tag{25}$$

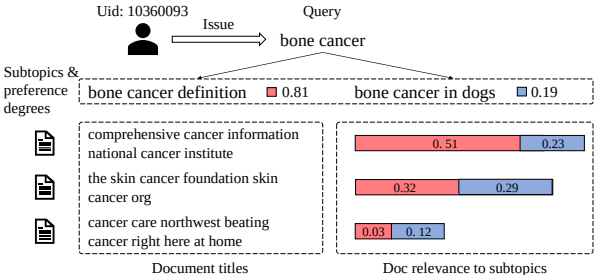


Figure 5: A case for interpretability of the search result.

The lower the intention entropy is, the more stable the user intent is covered.

B CALCULATION OF SUBTOPIC ENTROPY

Considering that the user's subtopic preferences are generated by the PAM, the results may be biased in favor of our model. Consequently, we additionally design "subtopic entropy" (SEnt) that excludes user preferences to measure the subtopic coverage of top results. The following provides a presentation of its computation:

$$\mathrm{SEnt}(T_N) = -\sum_{i=1}^{k} p(s_i|T_N) \log\left(p\left(s_i|T_N\right)\right) \tag{26}$$

$$p(s_i|T_N) = \mathrm{softmax}_i\left(\sum_{j=1}^{N} c(s_i|d_j)\right). \tag{27}$$

The subtopic entropy can measure the coverage stability of the top-N results on certain subtopics. The stability of query subtopic coverage increases with decreasing subtopic entropy.

C CASE OF SEARCH RESULT INTERPRETABILITY

As we introduced in Section 1, considering the user's preference distribution on query subtopics is also in favor of the search result interpretability. Thus, we provide an example to illustrate it, which is displayed in Fig. 5. We randomly select a behaviour in the test dataset, and visualize the user preferences and ranking results from our model. We normalize and present the most preferred two subtopics to save space. Our model finds that the user want to understand the definition when she searches for an unknown disease, thus it predicts that the user prefers the subtopic "bone cancer definition" . As a result, the model will prioritize the documents related to the "bone cancer definition". In this case, we can provide a reasonable interpretability of the returned list, hence improving the user's search experience.