# Hierarchical Document-Aware Interest Profiling in Personalized Search

Yutong Bai, Yujia Zhou, Zhicheng Dou, *Member, IEEE*, and Ji-Rong Wen, *Senior Member, IEEE*

***Abstract*—Personalized search has been proven to be an effective method to improve ranking quality by tailoring result lists according to the user's search history. Previous studies achieve personalization by learning a user interest profile from the search log, and decide the candidate document's ranking score by calculating its relevance with the learned profile vector. However, existing approaches overlook fine-grained interaction signals by treating the candidate document separately from the user's search history, relying solely on comparisons with a unified interest vector for re-ranking. Leveraging history-document interactions is not trivial due to the challenge of assessing the contributions of fine-grained matching signals within complex evolving interest patterns. In this paper, we address this challenge by helping the model understand these interactions within the evolving interest process through their integration into the interest profiling procedure. Specifically, we hierarchically incorporate these interaction signals as document-aware interests into behavior representations, employing explicit balancing and differentiation mechanisms, while jointly learning the interest pattern from both actual clues derived from original interests and potential insights provided by document-aware interests. Experimental results show that our model obtains substantial improvements over existing methods.**

***Index Terms*—Personalized search, user profiling, interest modeling.**

## I. Introduction

In RECENT years, using search engines to obtain information from the web has become increasingly popular. A search engine is designed to deliver results that align with the user's specific interests. However, the queries entered can be ambiguous or broad [1], sometimes making it challenging to accurately capture the user's needs, resulting in the engine failing to provide a satisfactory list of results. To address the problem, many personalized search methods are proposed [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. With the utilization of the user's search history, these methods aim to build an appropriate interest profile for the better modeling of the user's information needs and provide a re-ranked document list based on the profile. Traditional personalization studies use click-based or topic-based features from the query logs to build interest for each user [2], [3], [4], [5], [6], [7]. Recently deep-learning based strategies have achieved better results by capturing the sophisticated dynamics of user interests through neural networks [8], [9], [10], [11], [12], [13], [14]. These methods have shown an adorable ability to model user interests through their efficient excavation of search histories.

Typically, the personalized web search process in most existing works [8], [9], [10], [11], [12], [13] can be divided into two key stages:

- *Stage 1: Interest Profiling* – In this phase, user interest representations are learned based on historical user behaviors and the query, without incorporating the candidate document. This stage focuses on forming a general understanding of user preferences that are independent of specific candidate documents.
- *Stage 2: Re-Ranking* – In this stage, candidate documents are scored and ranked based on the user interest representations developed in Stage 1, mostly using similarity-based approaches such as cosine similarity, dot-product operations, or other sophisticated scoring mechanisms.

Clearly, this strategy fails to capture the fine-grained relationships between candidate documents and individual historical behaviors within the evolving patterns of user interests. Instead, it relies solely on the relatedness between candidate documents and an interest profile aggregated from the entire history. While this simplifies computation, it overlooks critical interest clues embedded in historical behaviors that can only be revealed through interactions with candidate documents.

As illustrated in Table I, the traditional strategy is more suitable for candidate document lists that contain only documents like $c_1$, where all relevant interest clues can be effectively captured within the unified interest profile generated by history (and current query) modeling. However, if the document list includes documents like $c_2$, the traditional approach fails to identify and enhance the specific correlated interest parts between the historical search $h_2$ and the document (highlighted in blue in the table), as these relationships are not incorporated during the interest profiling stage. Consequently, this limitation may lead to the underestimation of $c_2$.

Obviously, enabling direct interactions between candidate documents and individual historical behaviors provides more evidence for document ranking and can significantly boost performance. This approach has been validated in other information retrieval tasks that leverage user history. For instance, a context-aware document ranking method [15] models the connections

Received 11 March 2024; revised 26 January 2025; accepted 16 January 2026. Date of publication 27 January 2026; date of current version 9 May 2026. This work was supported by the National Natural Science Foundation of China under Grant 62272467. Recommended for acceptance by Y. Gao. *(Corresponding author: Zhicheng Dou.)*

Yutong Bai and Yujia Zhou are with the School of Information, Renmin University of China, Beijing 100872, China (e-mail: ytbai@ruc.edu.cn; zhouyujia@ruc.edu.cn).

Zhicheng Dou and Ji-Rong Wen are with the Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China (e-mail: dou@ruc.edu.cn; jrwen@ruc.edu.cn).

Digital Object Identifier 10.1109/TKDE.2026.3658355

TABLE I
EXAMPLE USER SEARCH BEHAVIORS. RELEVANT INFORMATION PIECES BETWEEN HISTORICAL BEHAVIORS $h_1$, $h_2$, $h_3$, AND THE CANDIDATE DOCUMENTS $c_1$, $c_2$ ARE MARKED IN BLUE. IT IS CLEAR THAT $c_2$ CONTAINS INFORMATION ABOUT "ZEN GARDEN," WHICH ALIGNS WITH THE USER'S INTERESTS REVEALED IN $h_2$. EXISTING METHODS ISOLATE CANDIDATE DOCUMENTS FROM THE INTEREST PROFILING OF HISTORICAL BEHAVIORS, LEADING TO A LOSS OF SUCH CORRELATIONS IN THE INTEREST PROFILE AND, CONSEQUENTLY, AN UNDERESTIMATION OF $c_2$

| Item | Value |
|---|---|
| Historical search $h_1$ | 32 Cheap Beach Vacations for Travelers on a Budget |
| Historical search $h_2$ | The 10 Most Inspiring *Zen* stone *Gardens* |
| Historical search $h_3$ | Capital International Airport |
| Current query $q$ | **Tourist sites in Japan** |
| Candidate document $c_1$ | Japan Cultural Center Library |
| Candidate document $c_2$ | Ryoan-ji *Zen Garden*: the most famous and most austere *Zen* Buddhist *garden*. |

between documents and historical behaviors by jointly encoding them with BERT [16]. However, in personalized web search tasks, researchers have not yet adopted direct document-history matching at the behavior level due to **the challenge of assessing the contributions of fine-grained history-document matching signals within complex evolving interest patterns**. Unlike context-aware document ranking tasks that typically rely on short-term history, personalized web search necessitates learning interest patterns from long-term histories that span multiple sessions with complex and evolving interest variations. As the example in Table I, a greater **quantity** of document-related interest correlations (highlighted in blue) does not necessarily indicate that $c_2$ aligns better with user preferences than $c_1$. It is crucial to understand the **role** of document-related interests within the evolving interest pattern to determine their importance in re-ranking, while preserving the original history interest clues (in black text) to accurately capture the actual interest evolution.

*To better leverage history-document interaction signals, we aim to help the model understand these interactions within the evolving interest process by integrating them into the interest profiling procedure:* This ensures that interaction-based clues are organized and learned effectively under the evolving interest pattern. Since document-related interaction signals and original history information reveal different aspects—potential vs. actual interests—their contributions must be distinguished. Moreover, as potential interests evolve in alignment with actual interests, we do not separate them but fuse them into a single vector, while balancing and preserving their distinct roles within the learning process.

The Transformer [17] excels in semantic learning, making it ideal for encoding the semantic meaning of fused interests. Additionally, we assess the semantic matching signals to balance the contributions from both aspects. Specifically, we propose a Transformer-based joint learning framework with multi-faceted representation, which follows three key steps: fusion, balancing, and differentiation. In our scenario, we perform the following steps: 1) **Fusion**: We combine the document and behavior data, allowing the Transformer to model their dependencies. The behavior outputs, derived from the dominant aspect (original interests), serve as the preliminary interest fusion results. 2) **Balancing**: We assess the semantic matching level by comparing the independent Transformer outputs from both aspects: behavior and document. A high matching level indicates a greater influence from the joint learning process on the original interests. As a result, we reduce the contribution of fused interest outputs and increase the reliance on the independent original interest representations, forming the document-aware, behavior-level interest representation 3) **Differentiation**: We refine the document-related interest clues in the behavior-level interest representation using the candidate document during history-level Transformer encoding, with attention weights ensuring stronger alignment to document-specific features. This allows the two aspects to remain distinct while jointly modeling.

More specifically, we propose a **H**ierarchical **D**ocument-aware **I**nterest **P**rofiling framework (**HDIP**) for personalized search, which progressively leverages history-document matching information while respecting the evolving interest patterns. It consists of three modules: (1) **Behavior-level Document-aware Interest Profiling**. This module learns a behavior representation by **fusing** and **balancing** history-document interaction-based (document-aware) interests with original interests. (2) **History-level Document-aware Interest Profiling**. It constructs a document-aware interest profile by **differentiation** and jointly learning document-aware and original interests from the behavior sequence. (3) **Re-ranking**. It calculates the similarity between the document-aware interest profile and the candidate document to decide the document's final ranking score.

To summarize, the main contributions of the paper are as follows:

(1) We first integrate the candidate document into the interest profiling process for personalized web search, enabling the model to better leverage the history-document relatedness information by learning their contributions within the evolving interest pattern.

(2) We propose a behavior-level document-aware interest profiling module, which effectively includes history-document correlations from individual behaviors at the word level, while balancing the original interest clues within the behaviors.

(3)We design a history-level document-aware interest profiling module that jointly learns the contributions of original and document-related interests, effectively capturing their roles in the interest profiling process.

(4) We leverage the transformer's ability for joint learning with multi-faceted representation, consisting of three key steps: fusion, balancing, and differentiation, to effectively model the contributions of different interest aspects.

## II. RELATED WORK

### A. Personalized Web Search

In recent years, personalized search has been provided as an effective technique for improving ranking quality by leveraging users' search history. The key to personalized methods of achieving satisfactory results is understanding user preferences from the rich historical information. Thus, multiple works have

endeavored to mine useful features within the query log. Generally, such features could be classified into two categories: click-based features and topical-based features. There are many early studies show great attention to click-based features since they are both accessible and informative. The P-Click model, proposed by Dou et al. [3], predicts the click probability by counting the historical click number on documents. Similar strategies can be found in [7] when figuring out the personal navigation problem. As for the topical-based features, early studies such as Open Directory Project (ODP) [7], [18], [19] build topical profiles from clicked documents through a manual online ontology. However, these methods suffer from the problem that not all clicked documents appear in the online ontology, which prevents the model from tackling new documents. In recent years, some latent-topical models have been proposed and have effectively alleviated this problem [4], [20], [21]. Then, some researchers [19], [22] attempt to combine these features through ranking algorithms like LambdaMART [23] with greatly improved results.

Deep learning based methods can extract features automatically and thus are widely adopted in recent personalized search tasks. Song et al. [24] utilize individual evidence when adapting the general ranking model. Ge et al. [9] propose a hierarchical recurrent neural network to enhance sequential information and use a query-aware mechanism to extract interest-related features. Zhou et al. [12] focus on better query reformulation by encoding queries with the history as contextual information. Yao et al. [25] employ personal word embeddings where the representations are mainly decided by each user's data.

However, in these works the candidate document does not join the procedure of interest modeling. This prevents the deeper excavation of history-document correlations. In this paper, we explicitly model history-document relatedness in interest profiling to address this problem.

### B. Document Matching in Information Retrieval

The goal of document-aware interest modeling proposed in this paper is achieved by exploiting the matching features between local historical behaviors and candidate documents. Actually, in information retrieval, a group of studies can be formalized as a matching problem.

Typically, the ad-hoc retrieval task is a matching problem between the search query and the candidate document. A great number of ad-hoc methods [26], [27], [28] perform semantic matching by independently representing the query and the document and then comparing the two representations. Another popular paradigm in ad-hoc search [29], [30], [31], [32], [33] performs relevance matching by jointly modeling the query-document pairs to explore sophisticated interaction signals. For instance, Yu et al. [34] explore the different grain-sized hierarchical matching signals between the document and the query. They explicitly model the document-level word relationship through the graph structure, where subtle long-distance information and different grain-sized hierarchical matching signals are well captured. Likewise, Fu et al. [35] devise a mult-view inter-passage interaction-based ranking model, which successfully utilizes longer-range relationships in passage with the influence of the query.

Similarly, the personalized task can be formalized as a matching problem between the user's search history and the candidate document. A popular approach for personalized task [36], [37], [38], [39], [40], [41], [42], [43] is modeling and comparing the interest representation learned from user history and the candidate document representation, which is similar to the semantic matching approach in ad-hoc search. Recently, some researchers have attempted to solve the personalized search problem by capturing more detailed matching signals through jointly modeling the user history and the candidate document. For example, Wang et al. [44] choose to perform fine-grained interest matching between each pair of browsed news and the candidate news for personalized news recommendation. Similarly, for personalized product search, Bi et al. [45] jointly encode historical reviews and candidate item reviews with a transformer architecture.

Whereas, for personalized web search, excavating fine-grained history-document matching features could be a complicated problem. It is mainly because the user does not reveal explicit signals, like reviews in product search, about real user interests. Either the history or the document could contain much noise to the user's actual interest profiling. Simply matching the behavior-document pair as [44] or encoding the behavior sequence and the document together as [45] would impede the model from capturing real interests and lead to low ranking quality.

Due to the complexity of user interests mentioned above, existing personalized web search approaches consider the candidate document only at the re-ranking stage, excluding it from the interest profiling process. Most works [8], [9], [10], [11], [12], [13], [46], [47] completely overlook the relevance between the candidate document and individual behaviors, simply comparing the document with a unified user representation vector derived from the entire user history. While a recent work [15] models the connections between documents and behaviors by jointly encoding them with BERT, it bypasses the interest profiling stage and fails to address how the candidate document influences interest modeling.

To leverage the potential of candidate documents while preserving user interests, we propose a model that performs fine-grained history-document matching from an interest profiling perspective. In this approach, interest clues from the original user history and the candidate document are carefully fused and balanced.

## III. Proposed Method

As mentioned in Section I, most personalized methods overlook fine-grained interactions between candidate documents and behaviors, relying solely on a unified interest profile, which hinders re-ranking performance. To address this, we progressively capture history-document interaction clues in interest profiling through fusion, balancing, and differentiation of document-aware and original interests. To begin with, the problem is

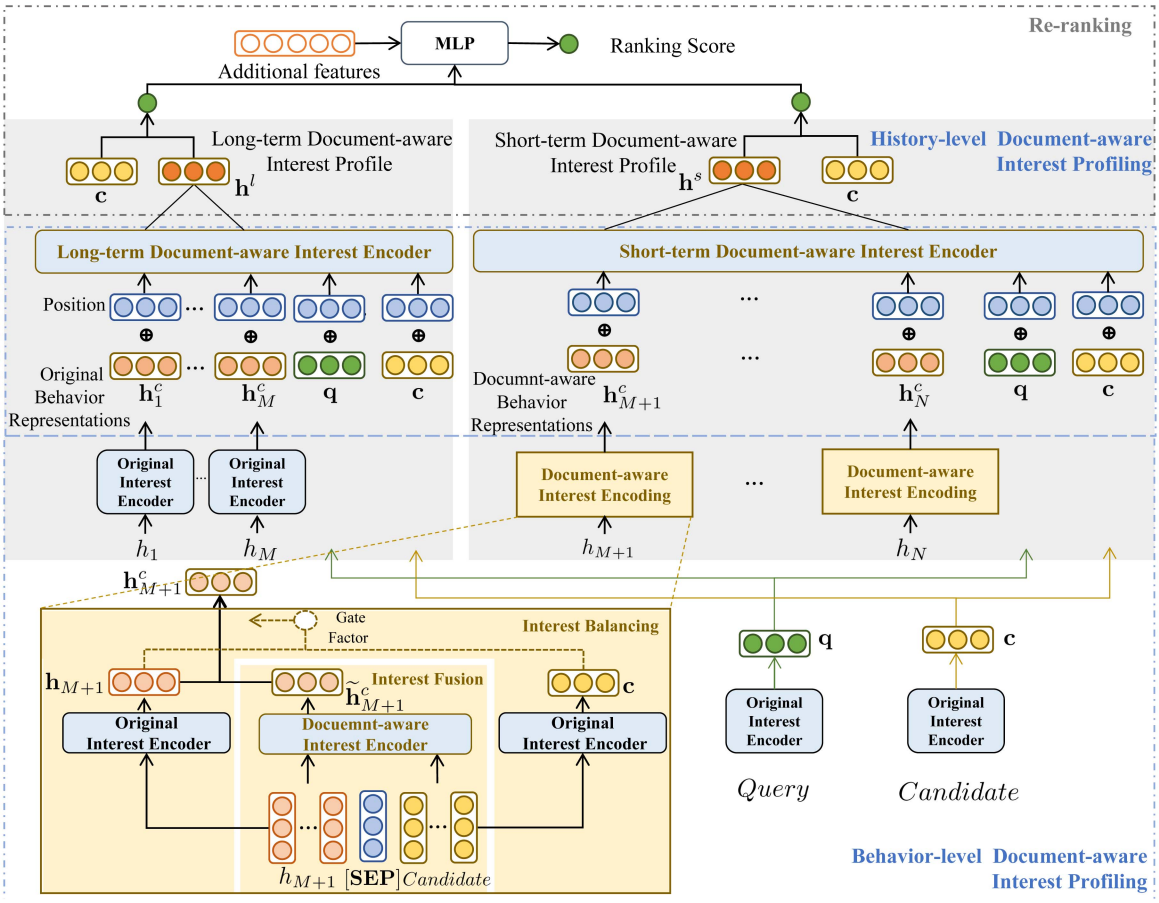


Fig. 1. **The overview of the proposed HDIP.** The model profiles document-aware interests by progressively fusing, balancing, and differentiating interests from both the history and candidate document. In **behavior-level document-aware interest profiling**, it applies **document-aware interest encoding** to short-term behaviors through the **interest fusion** and **interest balancing** modules. For each behavior, first, the model fuses interests by integrating the influence of the candidate document based on word-level history-document relatedness. Then, it refines the fused representation by filtering out noise according to the behavior-level history-document relatedness (computed from independent encoded word sequences of the behavior and the candidate document). For long-term behaviors, the model directly employs **original interest encoders** due to weaker history-document correlations. After obtaining the current query and candidate document representations by original interest encoders, the model proceeds to **history-level document-aware interest profiling**, enhancing document-related interests from the behavior sequence guided by the candidate document representation. In this way, during the **re-ranking** stage, the generated profiles preserve both document-aware and original interests distinctly, enabling the comparison with the candidate document to capture matching signals from both aspects without interference.

formulated as follows. Suppose that for each user, his historical query log $H$ is composed of a long-term history $H^l$ and a short-term history $H^s$. The former consists of the user's former $M$ past behaviors, $H^l = \{h_1, \ldots, h_M\}$, where $h_i$ represents the concatenation of issued queries and their corresponding clicked document titles. The latter includes $N - M$ searched behaviors after the $M$ behaviors, $H^s = \{h_{M+1}, \ldots, h_N\}$. Given the current query $q$ and the candidate document list $C = \{c_1, c_2, \ldots\}$ returned by the search engine, our target is to calculate a ranking score $score(c)$ for each candidate document $c$ in $C$ according to the current query $q$ and historical behaviors $H$.

As illustrated in Fig. 1, we devise a hierarchical document-aware interest profiling network to tailor an optimal ranking list by attending the history-document dependency based on the interest evolvement in a progressive way. Next, we will elaborate on the construction details of our proposed HDIP in the following three stages: (1) behavior-level document-aware interest profiling, and (2) history-level document-aware interest profiling. (3) re-ranking.

### A. *Behavior-Level Document-Aware Interest Profiling*

In this stage, we capture the interactions between the candidate document and individual behaviors at the word level, considering their roles in interest profiling. To achieve this, we integrate document-aware interests into behavior representations while preserving the original interests based on word-level history-document relatedness. To further minimize noise and reduce computational costs, document-aware behavior-level interest encoding is applied only to short-term behaviors, where history-document relatedness is more prominent.

*1) Interest Fusion:* In this module, transformer encoders are employed to capture word-level dependencies between the behavior and the candidate document. The behavior's output words incorporate informational signals from the candidate document, guided by their relatedness. By combining these words, we obtain preliminary document-aware behavior representations that retain the original interests while appropriately integrating document-aware interests. To be specific, in the beginning, we

get the word embedding sequence $\mathbf{h}_i^w$ for historical behavior $h_i$, as well as the word embedding sequence $\mathbf{c}^w$ for the candidate document $c$. Following existing personalized search works [9], [10], [12], the word embeddings are initialized by training a word2vec [48] model.

To enhance the effects from the candidate document at word-level, we then encode the two word sequences with a separate token [**sep**] to a word-level **document-aware interest encoder** noted as Trm$(\cdot)$:

$$\widetilde{\mathbf{h}}_i^c = \text{Trm}^{\text{history}}(h_i^w + h_i^{w,p}, [\mathbf{sep}] + [\mathbf{sep}]^p, c^w + c^{w,p}]), \quad (1)$$

where $\widetilde{\mathbf{h}}_i^c$ represents the preliminary document-aware behavior representation, summarized from the corresponding word outputs. Trm$^{history}$ denotes the transformer encoder's word-level outputs of the historical behavior. $h_i^{w,p}$, $[\mathbf{sep}]^p$and $c^{w,p}$ is the corresponding word position embeddings sequence of $h_i^w$, [**sep**] and $c^w$. The encoder would automatically learn the relationship between the historical words and candidate words. For historical words that are more related to the candidate words, the corresponding outputs would be assigned with larger attention weights by the related candidate words. In this way, the preliminary document-aware behavior representation $\widetilde{\mathbf{h}}_i^c$ can include document-related information along with the behavior's original information.

Within the Transformer process used in our document-aware interest encoder, we apply the standard *multi-head self-attention* mechanism. The input is first projected into multiple heads by applying different learned weight matrices for the query ($\mathbf{q}$), key ($\mathbf{k}$), and value ($\mathbf{v}$) for each head $h$:

$$\mathbf{q}_h = \mathbf{W}_Q^h \cdot h_i, \quad \mathbf{k}_h = \mathbf{W}_K^h \cdot h_i, \quad \mathbf{v}_h = \mathbf{W}_V^h \cdot h_i \quad (2)$$

where $\mathbf{W}_Q^h$, $\mathbf{W}_K^h$, $\mathbf{W}_V^h$ are the learned weight matrices for the $h$-th head, and $h_i$ represents the input word embeddings for the entire behavior sequence. Next, the attention weights for each head are computed based on the query and key:

$$\text{Attention}_h(\mathbf{q}_h, \mathbf{k}_h, \mathbf{v}_h) = \text{Softmax}\left(\frac{\mathbf{q}_h \cdot \mathbf{k}_h^T}{\sqrt{d_k}}\right) \cdot \mathbf{v}_h. \quad (3)$$

The outputs from all heads are then concatenated and projected into a single vector. This process is followed by a position-wise feed-forward network, which applies a fully connected layer with ReLU activation, and layer normalization to stabilize training.

*2) Interest Balancing:* While Transformers excel in capturing semantic relationships, they may still introduce noise when candidate documents contain a substantial amount of irrelevant content to the user's behavior. This issue is prevalent, as candidate documents are often imperfect. To address this, we introduce an interest balancing gate that effectively filters out misleading signals, ensuring that document-aware interests are integrated without compromising the original interest representation.

To achieve this, we regulate the document's influence by evaluating the semantic similarity between the independently encoded behavior and the candidate document. More precisely, we generate the original interest behavior representation $\mathbf{h}_i$ by summarizing the outputs from the **original interest encoder** as following:

$$\mathbf{h}_i = \text{Trm}(h_i^w + h_i^{w,p}). \quad (4)$$

Similarly, we obtain the candidate document representation $\mathbf{c}$. After that, we send $\mathbf{h}_i$ and its corresponding document-aware behavior representation $\widetilde{\mathbf{h}}_i^c$ to a gate mechanism where the gate factor is the cosine similarity between $\mathbf{h}_i$ and $\mathbf{c}$. The procedure can be formulated as follows:

$$z_i = \text{sim}(\mathbf{h}_i, \mathbf{c}), \quad (5)$$

$$\mathbf{h}_i^c = z_i * \widetilde{\mathbf{h}}_i^c + (1 - z_i) * \mathbf{h}_i, \quad (6)$$

where $\mathbf{h}_i^c$ is the final document-aware behavior representation. sim$(\cdot)$ refers to the cosine similarity. We believe that if $\mathbf{h}_i$ and $\mathbf{c}$ have larger similarities in the semantic representations, they will carry closer information and the participation of the document in behavior representation would cause less damage to the original interest.

### B. History-Level Document-Aware Interest Profiling

Noticing that document-aware interests only reflect potential interests—since the user's satisfaction with the document remains uncertain—while historical interests represent certain preferences, it is crucial to ensure that the final profile retains distinct cues from both aspects. This distinction helps the model determine the importance of each interest cue based on its characteristics. Therefore, we further differentiate the original interests from the document-aware interests by leveraging the candidate document to enhance document-related signals. Considering the density of history-document relatedness, we separately perform history-level document-aware encoding for short-term history and long-term history.

*1) Short-Term Document-Aware History Encoding:* For short-term history, we join the document-aware behavior representations from the previous module as the history sequence and use the candidate document representation $\mathbf{c}$ to enhance the document-related behaviors. Moreover, we also use the current query representation $\mathbf{q}$, which is generated as $\mathbf{c}$ in (4), to emphasize behaviors related to the current search.

To be specific, for each document-aware behavior representation $\mathbf{h}_i^c$ and the corresponding position embedding in the historical sequence $\mathbf{h}_i^{c,p}$, we get the short-term behavior input $\mathbf{h'}_i^c$. Likewise, we obtain the query input $\mathbf{q}'$ and the candidate document input $\mathbf{c}'$ :

$$\mathbf{h'}_i^c = \mathbf{h}_i^c + \mathbf{h}_i^{c,p}, \quad (7)$$

$$\mathbf{q}' = \mathbf{q} + \mathbf{q}^p, \quad (8)$$

$$\mathbf{c}' = \mathbf{c} + \mathbf{c}^p, \quad (9)$$

where $\mathbf{q}^p$ and $\mathbf{c}^p$ is the position embedding for the current position.

Next, we join the behavior sequence with the current query and the candidate document and send them to a history-level

transformer encoder. This operation can be formulated as follows:

$$\mathbf{h}^s = \text{Trm}^{\text{history}}([\mathbf{h'}^c_{M+1}, \ldots, \mathbf{h'}^c_N, \mathbf{q}', \mathbf{c}']). \tag{10}$$

$\mathbf{h}^s$ is the short-term document-aware interest profile, which is the summarization of the corresponding historical behavior representations.

*2) Long-Term Document-Aware History Encoding:* For long-term history where word-level correlations are less frequent, incorporating the document into behavior representations would cause more damage to original interest representing. Hence, we abandon the document-aware encoding and perform document-aware interest profiling directly at the behavior level.

To be specific, we add the position embedding and type token to the original behavior representation $h_i$ and combine the long-term behaviors with the candidate document. The original behavior representation is generated as (4). The long-term document-aware interest modeling is shown as follows:

$$\mathbf{h}^l = \text{Trm}^{\text{history}}([\mathbf{h'}^c_1, \ldots, \mathbf{h'}^c_M, \mathbf{c}']), \tag{11}$$

where $\mathbf{h}^l$ is the long-term document-aware interest profile, summarized by the corresponding outputs of the historical behavior representations.

### C. Re-Ranking

By maintaining distinct original and document-aware interest features, the comparison between the candidate document and the profile enables the model to evaluate both established preferences and potential interests, ensuring a more accurate alignment assessment.

The matching quality of the short-term and long-term history is calculated by the cosine similarity of the corresponding interest profile and the document. This procedure can be formulated as follows:

$$s^l = \text{sim}(\mathbf{h}^l, \mathbf{c}), \tag{12}$$

$$s^s = \text{sim}(\mathbf{h}^s, \mathbf{c}), \tag{13}$$

where $s^l$ refers to the long-term matching score, while $s^s$ refers to the short-term matching score. $\text{sim}(\cdot)$ is the cosine similarity, computed by the dot product of the vectors divided by their magnitudes (norms).

At the final re-ranking stage, we send the two matching scores into the multilayer perceptron (MLP) $\phi(\cdot)$, which learns weights on the input scores, allowing it to capture the relative importance of each score in the final ranking for the candidate document $c$:

$$\text{score}\,(c|H,\ q) = \phi[s^l, s^s]. \tag{14}$$

The MLP is defined as:

$$\phi(x) = \sigma(Wx + b), \tag{15}$$

with $x = [s^l, s^s]$ being the input vector consisting of the two matching scores, $W$ being the weight matrix, b the bias term, and $\sigma$ the activation function (such as ReLU or sigmoid).

Furthermore, following [5] we extract additional features $f_{(q,\ d)}$, which is popularly used in many state-of-the-art methods (e.g., HTPS [12], PSSL [13]), and send them into MLP to generate the ad-hoc relevance score:

$$\text{score}\,(c|\ q) = \phi\left(f_{(q,\ d)}\right). \tag{16}$$

At last, the final ranking score of the candidate document is calculated based on the two scores:

$$\text{score}(c) = \text{score}\,(c|H,\ q) + \text{score}\,(c|\ q)\,. \tag{17}$$

During the training procedure, we utilize a basic ranking algorithm, LambdaRank [49], to update the network parameters. Training pairs are generated from the search log by taking SAT-clicked documents as positive samples and others as negative samples. For example, given a positive sample $c_i$ and a negative sample $c_j$, the loss function is defined as the product of cross entropy between real probabilities and predicted probabilities:

$$loss\ = -\left|\lambda_{ij}\right|\left(\bar{p_{ij}}\log\left(p_{ij}\right)\ +\ \bar{p_{ji}}\log\left(p_{ji}\right)\right), \tag{18}$$

where $|\lambda_{ij}|$ is the change of metrics while swapping the positions of the two documents. $p_{ij}$ symbolizes the predicted probability that $d_i$ is more relevant than $d_j$, and $\bar{p_{ij}}$ is the real probability. The predicted probability is calculated by a logistic function:

$$p_{ij} = \frac{1}{1 + \exp\left(\text{score}(c_j) - \ \text{score}(c_i)\right)}. \tag{19}$$

## IV. Experiment Setup

### A. Dataset

We conduct experiments on a real-world dataset from a commercial search engine, which is referred as the "B dataset" in the following of this paper. The basic statistics are shown in Table II. This large-scale search log has the click-through data for two months from 1st January 2013 to 28th February 2013. Each piece of data contains a user ID, a query string, the query issued time, a session identifier, the top 20 retrieved URLs, their titles, click labels, and dwelling times. The documents with dwelling times longer than 30 seconds are regarded as clicked documents. As personalized search is based on historical data, we divide the dataset into historical data and experimental data. The first six weeks are taken as historical data and the last two weeks are experimental data. Users with less than 4 sessions in the experiment data are discarded. For experimental data, we further divide it into a training set, validation set, and testing set with a ratio of 4:1:1 by sessions.

TABLE II
Basic Statistics of the Dataset

| Item | Statistics | Item | Statistics |
|---|---|---|---|
| #days | 58 | #distinct queries | 1,624,496 |
| #users | 33204 | #sessions | 654,776 |
| #queries | 2665625 | #SAT-clicks | 1,228,028 |

### B. Model Settings and Evaluation Metrics

With the clicked documents treated as relevant ones and others as irrelevant ones, we use the following three common metrics to evaluate the quality of different models: mean average

precise(MAP), mean reciprocal rank (MRR), and precision@1 (P@1). In addition, we generate inverse document pairs according to [9], [10] to measure a reliable re-ranking improvement. The main reason for that operation lies in the phenomenon that a relevant document may be overlooked due to its lower position. The P-improve is adopted as a more credible method to further eliminate the influence of position bias. To achieve a balance between effectiveness and efficiency, we conducted multiple experiments to determine the construction details of our model. The final parameters are set as follows: The number of behaviors is 20 for short-term history and 30 for long-term history. The embedding dimension is 100. The word-level transformer encoder for document-aware behavior encoding and original behavior encoding is one layer with 6 heads. Besides, it is also the same as the transform encoder for query encoding and document encoding. The history-level transformer encoder for long-term is 4 heads with one layer, while the history-level transformer encoder for the short-term is 4 heads with two layers.The transformers in the original interest encoders, document-aware interest encoders, long-term document-aware interest encoders, and short-term document-aware interest encoders do not share parameters between these four types. We train the model for 2 epochs to achieve a satisfactory result. The learning rate is 1e-4 for the first epoch and 1e-5 for the second epoch.

### C. Baselines

Besides the original ranking returned by the commercial search engine, we compare our model with several state-of-the-art ad-hoc search models and personalized search models. They are listed as follows:

*KNRM [31]:* It is a neural ranking model designed for ad-hoc search. It models the interactions between queries and documents with the kernel-pooling to extract soft match features.

*Conv-KNRM [50]:* It is an upgraded version of KNRM which models n-gram soft matches with an additional convolutional layer. Features Of surrounding words are taken as contextual information to learn context-aware word embeddings.

*BERT [16]:* It uses the pre-trained BERT model to encode the concatenated query-document sequence. The output of the "[CLS]" token from the last layer is taken as the final matching feature.

*P-Click [3]:* It ranks the documents simply based on the number of clicks on the same document under the same query in history.

*SLTB [5]:* It uses LambdaMART to train personalized learning to rank models with more than 100 features.

*HRNN [9]:* It uses a hierarchical RNN with query-aware attention to generate dynamic user profiles.

*PEPS [25]:* This work tackles the personalized search problem through an alternative approach without building a user interest profile. It trains personal word embeddings for each user where the representations are mainly decided by individual user data.

*HTPS [12]:* This model attempts to enhance the query representation for a better reflection of users' informative needs. It regards historical data as contextual information for the current query and devises a query disambiguation sub-model and a personalized language sub-model to refine the query under different scenarios.

*PSSL [13]:* This work designs a self-supervised learning framework with a contrastive sampling technique to enhance data representation. It alleviates the data sparsity problem when trying to learn high-quality representations.

*DIMPS [51]:* This work constructs dynamic document representations by incorporating the influence of both the user's historical and current intent. It leverages passage-level evidence encoded by a pre-trained language model from the full document content to achieve fine-grained interest modeling.

*HDIP (**H**ierarchical **D**ocument-aware **I**ntertest **P**rofiling Model):* Our proposed personalization model with a detailed description in Section III.

*HDIP(PLM):* To ensure a fair comparison, we implement a model variation that employs the same **P**re-trained **L**anguage **M**odel—Sentence-BERT [52]—to replace our original interest encoder, as described in Section III-A2. Similar to DIMPS, the pre-trained model remains frozen during training.

TABLE III
OVERALL PERFORMANCE. THE BEST RESULTS ARE SHOWN IN BOLD. '†' INDICATES THE MODEL SIGNIFICANTLY OUTPERFORMS ALL BASELINE MODELS WITH PAIRED T-TESTS AT $p<0.05$ LEVEL

| Model | MAP | MRR | P@1 | P-improve |
|---|---|---|---|---|
| Adhoc Search Model | | | | |
| Ori. | .7399 | .7506 | .6162 | - |
| KNRM | .4916 | .5001 | .2849 | .0655 |
| Conv-KNRM | .5872 | .5977 | .4188 | .1442 |
| BERT | .6232 | .6326 | .4475 | .1778 |
| Personalized Search Model | | | | |
| P-Click | .7509 | .7634 | .6260 | .0611 |
| SLTB | .7921 | .7998 | .6901 | .1117 |
| HRNN | .8065 | .8191 | .7127 | .2404 |
| PEPS | .8221 | .8321 | .7251 | .2545 |
| HTPS | .8224 | .8324 | .7286 | .2552 |
| PSSL | .8301 | .8398 | .7338 | .2688 |
| DIMPS | .8421 | .8512 | .7532 | .2939 |
| Our Model | | | | |
| HDIP | $0.8424^{\dagger}$ | $0.8517^{\dagger}$ | $0.7534^{\dagger}$ | $\mathbf{0.2973}^{\dagger}$ |
| HDIP(PLM) | $\mathbf{.8436}^{\dagger}$ | $\mathbf{.8529}^{\dagger}$ | $\mathbf{.7559}^{\dagger}$ | $.2965^{\dagger}$ |

## V. RESULTS AND ANALYSIS

### A. Overall Performance

We conduct experiments for all the baseline models and our HDIP on the commercial search log dataset to evaluate our model's performance. The overall performance is shown in Table III. It can be observed that:

(1) **Compared to all the baseline models, including ad-hoc models and personalized models, our proposed HDIP shows significant improvements with paired t-test at $p<0.05$ level on the dataset.** Especially for the best personalization model DIMPS, HDIP outperforms it in terms of all evaluation metrics. Especially when using the same pre-trained language models, our HDIP(PLM) improves the ranking results by 0.15% in terms of MAP and 0.17% in terms of MRR. Furthermore, in terms of the more convincing metric P-improve, our model outperforms

DIMPS by 0.34% . Note that the recent baselines, PSSL and DIMPS, leverage the full document body, while our approach only utilizes document titles. These results demonstrate that incorporating fine-grained history-document matching features is more effective for personalization than solely matching the candidate document with an independent user profile learned from the history, even when the latter contains more concrete interest clues.

(2)The transformer-based methods (e.g., BERT, HTPS, PSSL, DIMPS, HDIP) achieve advanced performance in both ad-hoc search and personalized search tasks. In our HDIP, we further extend the capability of transformers beyond sequential semantic encoding by assessing matching signals to regulate the interaction between document-aware and original interests, ensuring a more accurate user profile.

(3) In general, all the personalized methods improve the ad-hoc search results greatly, indicating the contribution of search history in referring users' real information needs. In addition, the significant improvements of methods leveraging neural networks over the traditional methods (e.g., P-Click, SLTB) verify the power of the deep-leaning technique of capturing latent features. Moreover, among all evaluation metrics, the results of P@1 are more obvious than others, possibly due to the personalized methods' superior ability to tackle re-finding behaviors compared with other behaviors without sufficient relevant logs.

To summarize, the experimental results show that **by hierarchically attending history-document relevance from a fine-grained level, we successfully build a document-aware interest profile and boost the accuracy of the document ranking.** In the following sections, we are going to analyze the functionality of our major components with a set of experiments.

### B. Ablation Analysis

Our HDIP model is built upon two levels of interest profiling: the behavior level and the history level. Within the behavior-level profiling, we have an interest balancing module to balance the original interests and document-aware interests within behavior representations. For the three components(i.e., the two levels of profiling and the interest fusion module), in this section, we conduct several experiments to figure out their roles in search results personalization.

To test the functionality of the history-level document-aware profiling module, we design the following model:

*HDIP w/o. HD:* We strip off the history-level document-aware profiling (HD) part, described in Section III-B, for both the long-term history and the short-term history. Specifically, we discard the candidate document representation as transformer encoder inputs.

To test the functionality of the behavior-level document-aware profiling module, we design the following two models:

*HDIP w/o. BD:* We abandon the behavior-level document-aware profiling (BD) module, which is described in Section III-A. The short-term document-aware representations are replaced by original representations generated as (4) and then sent to the next history-level profiling module. This model is designed to demonstrate the effectiveness of incorporating document-aware interests from word-level into behavior representations, denoted as $\widetilde{\mathbf{h}}_i^c$ in (1).

TABLE IV
RESULTS OF ABLATION EXPERIMENTS

| Model | MAP | MRR | P@1 | P-improve |
|---|---|---|---|---|
| HDIP | **.8424** | **.8517** | **.7534** | **.2973** |
| w/o. HD | .8409 | .8503 | .7507 | .2953 |
| w/o. BD | .8372 | .8468 | .7446 | .2866 |
| w/o. IB | .8405 | .8500 | .7496 | .2938 |

*HDIP w/o. IB:* We discard the interest balancing (IB) module. Instead, we directly send the preliminary document-aware behavior representations from Section III-A1 as the inputs of history-level profiling. This model is designed to verify the effectiveness of balancing the contributions of document-aware and original interests in behavior representations, based on the history-document semantic matching signals.

As the results shown in Table IV, all the ablation models underperform the HDIP. Particularly, we can find that:

(1) The greatest drop is observed when stripping off the BD module. Specifically, the"HDIP w/o. BD" model significantly damages the results by 0.52% on MAP and 0.49% on MRR. This verifies the necessity of capturing the fine-grained history-document relevance. Without the matching features at the word level, the candidate document-aware profiling at the history level could not address informational history-document relevance and shows limited improvements in re-ranking.

(2) The "HDIP w/o. HD" model also damages the results by 0.15% on MAP. which highlights the importance of using the candidate document to emphasize and differentiate document-related features within the two aspects.

(3) The drop in the "HDIP w/o. IB" model by 0.19% on MAP demonstrates that effectively assessing the semantic matching signals helps prevent the document-aware features from hindering the interest pattern learning.

To conclude, all of the ablation models damage the results of the HDIP model but raise the results of all baseline models. This proves the efficiency of the three modules for improving ranking quality.

### C. Effects of the Interest Balancing

In the document-aware behavior representation module, we design an interest balancing part to prevent interference from the candidate documents and protect the original interest profiling. Within the interest fusion part, we introduce the original behavior representations to the final document-aware representations with a single-gate operation. In the ablation analysis section, the model "HDIP w/o. IB" has already verified the effectiveness of the balancing of original interests and document-aware interests within the behavior representations. In this section, we explore different balancing strategies to further investigate the influence between the two types of interest.

To be specific, we make another four variations for the interest balancing part as follows:

TABLE V
RESULTS OF DIFFERENT INTEREST BALANCING STRATEGIES

| Model | MAP | MRR | P@1 | P-improve |
|---|---|---|---|---|
| HDIP | **.8424** | **.8517** | **.7534** | **.2973** |
| IB-Add | .8328 | .8423 | .7365 | .2679 |
| IB-Concat | .8399 | .8493 | .7495 | .2892 |
| IB-DotProd | .8355 | .8451 | .7409 | .2823 |
| IB-MLP | .8354 | .8448 | .7408 | .2815 |

*HDIP IB-Add:* We simply add the original behavior representations $\mathbf{h}_i$ in (4) to their corresponding document-aware behavior representations $\widetilde{\mathbf{h}}_i^c$ in (1) to get the final document-aware representations $\mathbf{h}_i^c$.

*HDIP IB-Concat:* We concatenate the $\mathbf{h}_i$ and $\widetilde{\mathbf{h}}_i^c$ to get the final document-aware representations $\mathbf{h}_i^c$. Then, we send the document-aware representations into the MLP to project them to have the same dimension size as the query and candidate document representations. In this way, we let the model automatically decide the contributions from the two types of interest.

*HDIP IB-DotProd:* Instead of using cosine similarity, we compute the dot product of the gate inputs (original behavior representation and candidate representation) as the gate factor.

*HDIP IB-MLP:* We concatenate the gate inputs and pass them through an MLP, using the output as the gate factor.

As reported in Table V, the "HDIP IB-Add" suffers the most obvious accuracy drops of 0.96% , 0.94% , and 1.69% on MAP, MRR, and P@1 respectively. This indicates that simply summarizing the original behavior representations and document-aware representations would prevent the model from learning the actual user interests. One possible reason is that the addition operation does not distinguish the different roles played by the two types of interest, resulting in great information interference in interest profiling.

On the other hand, the "HDIP IB-Concat" also damages the results by 0.25% , 0.24% , and 0.39% on MAP, MRR, and P@1. This illustrates that the model cannot effectively address the contributions of the two types of interest from the concatenated vector. It may require a more sophisticated design to help the model further understand their dependencies and contributions.

When altering the gate factor calculation, the model drops by 0.69% on MAP for "HDIP IB-DotProd" and 0.7% for "HDIP IB-MLP". We believe this is because the gates require explicit similarity signals between the behavior and document, and cosine similarity seems to be the most effective way to measure the semantic relatedness of transformer outputs. Moreover, both models outperform "HDIP IB-Add", suggesting that integrating both behavior and document information helps adjust the influence of the document on the behavior during joint encoding.

To conclude, the gate mechanism is the most efficient structure for our HDIP model. Adjusting the information flow by the similarities between the candidate document and historical behaviors can balance the contributions of document-aware interest profiling and original interest profiling.

### D. Effects of Different History Lengths

In our proposed model, we separately model document-aware interest profiles for long-term history and short-term history, considering the density of document-related information. In this section, we would like to explore the impacts of the different lengths for short-term and long-term history. Concretely, we test our model with different history division strategies with unchanged total history length.

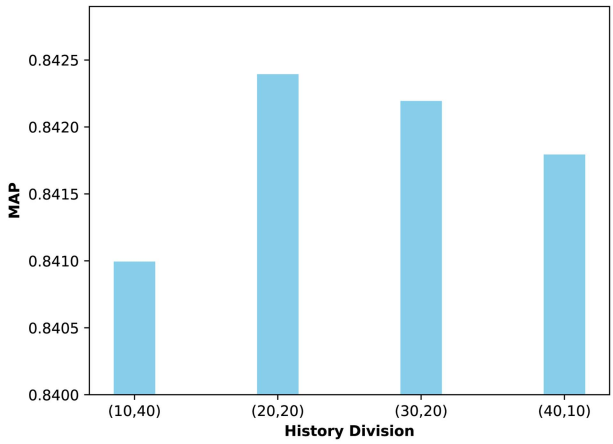


Fig. 2. Results of different history division strategies. Each division strategy is represented as $(M, K)$ where $M$ is the short-term history length and $K$ is the long-term history length.

In Fig. 2, we list the lengths as $(M, K)$ where $M$ is the short-term history length and $K$ is the long-term history length. It is illustrated that the accuracy significantly grows from a short-term length of 10 to 20. Meanwhile, the accuracy continues dropping when the short-term history is larger than 20. It is perhaps because the recent historical behaviors tend to have more correlations with the candidate document, and are more efficient in document-aware profiling. While longer historical behaviors are rarely related to the document. Fine-grained document-aware profiling from the word level for a longer history would bring noise for the model in capturing the accurate user interest.

In conclusion, the experimental results verify the necessity of modeling document-aware interest for short-term history and long-term history from different levels. Besides, this separate modeling strategy also indicates our document-aware profiling method does not require much more computational resources compared to the previous original profiling method.

### E. Case Study

The experimental results shown in Section V-A have demonstrated that modeling document-aware interest can improve ranking quality. Further, the results in Section V-B verify that capturing document-aware interests from word-level to history-level is effective for document-aware interest profiling. In this section, we analyze how the HDIP captures document-aware interests by studying a user log sampled from the dataset.

First, let's understand how HDIP captures user interests. Our model enhances document-related information in user behaviors using the transformer architecture. Specifically, behaviors more related to the candidate document receive higher attention weights, strengthening their influence in behavior representation and improving similarity matching. This shows that HDIP effectively enhances document-related information in user history to

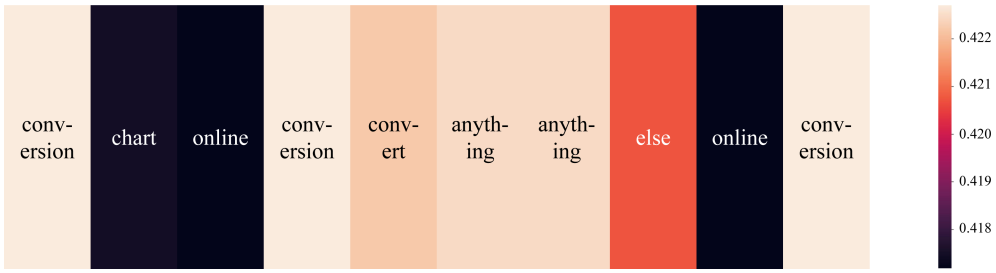


Fig. 3. The weights of the historical behavior words assigned by the candidate document. A lighter area indicates a larger weight. The current query is "grams tablespoons". The candidate document is "convert tablespoons gram conversion measurement units quickly convert". The words in the historical behavior are presented in the corresponding area.

capture document-aware interests. In this section, we visualize the document-aware attention weights for historical behavior words to demonstrate this capability.

Specifically, the document-aware weight of a word in historical behavior is the sum of the attention it receives from all words in the candidate document. Higher weights indicate stronger relevance to the document. Fig. 3 visualizes these weights at the word level. The historical query and its clicked document suggest a past interest in online conversion services, while the current query "grams tablespoons" provides limited clues about the user's intent. The candidate document, which covers grams-to-tablespoons conversion, may fit the user's needs given their past interest in "conversion".Traditional methods compress such interests into a unified profile without enhancement, potentially underestimating the document's relevance. In contrast, our model enhances the importance of "conversion" during behavior representation.

From Fig. 3, we can observe that words related to "conversion" gain larger document-aware weights. This leads to the enhancement of "conversion" interest in both the behavior representation stage and the history-level interest profiling stage. At last, the profile with such enhancement can make a more accurate match with the candidate document.

### F. Experiments on Ambiguous and Non-Ambiguous Queries

To study our model's contribution to capturing real user interests, we categorize the whole dataset into ambiguous and non-ambiguous query sets based on click entropy. Queries with the click entropy $\geq 1$ are treated as non-ambiguous queries while others are taken as ambiguous ones. The former refers to queries with more than one meaning like "Apple", and the latter describes queries with identical meanings for different users. Studies [3], [53] have shown that great ambiguity represents more potential for personalization.

From Fig. 4, we observe that all the personalized methods improve much more on ambiguous queries, which is consistent with the former conclusions. Regarding the best existing method PSSL, our models outperform it on both query sets. Moreover, the performance gap becomes larger on ambiguous queries, which demonstrates our model's superiority in clarifying the search interest. Similar results could be found between the whole model and the three ablation models. This confirms that each of our three major components is effective in excavating useful features to understand the user's real preference, which is coherent with our goal of leveraging document-related information.

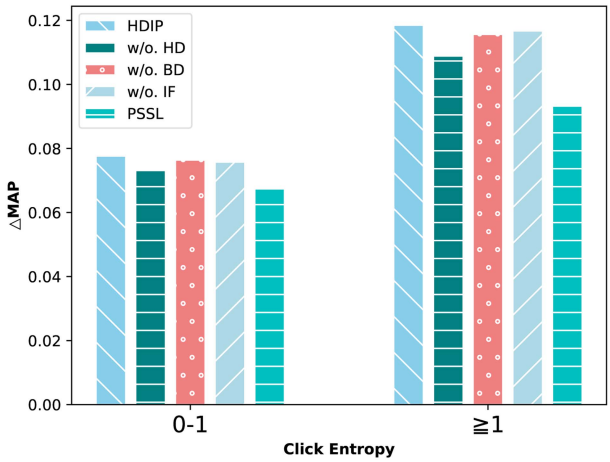


Fig. 4. Results on ambiguous and non-amibiguous queries.

## VI. Conclusion

Existing personalized works build interest profiles independently of candidate documents, which makes it difficult to capture fine-grained document-related features. To address the challenge of understanding history-document interactions for re-ranking, we propose the HDIP model, a hierarchical document-aware interest profiling approach that aligns interaction signals with the original interest pattern. By adapting transformers to jointly learn document-aware and original interests, while balancing and distinguishing them for effective pattern learning, we achieve better performance.

We conduct extensive experiments to test the effectiveness of our major components. Moreover, we explain the document-aware profiling procedure with an example. After that, we illustrate our model's performance on ambiguous and non-ambiguous queries to further verify its ability to capture real search interest.

## Acknowledgment

The work was partially done at the Engineering Research Center of Next-Generation Intelligent Search and Recommendation.

## REFERENCES


[1] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," in *Proc. ACM SIGIR Forum*, New York, NY, USA, 1999, pp. 6–12.

[2] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz, "Inferring and using location metadata to personalize web search," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 135–144.

[3] Z. Dou, R. Song, and J.-R. Wen, "A large-scale evaluation and analysis of personalized search strategies," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 581–590.

[4] M. J. Carman, F. Crestani, M. Harvey, and M. Baillie, "Towards query log based personalization using topic models," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 1849–1852.

[5] P. N. Bennett et al., "Modeling the impact of short-and long-term behavior on search personalization," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 185–194.

[6] M. Harvey, F. Crestani, and M. J. Carman, "Building user profiles from topic models for personalised search," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 2309–2314.

[7] A. Sieg, B. Mobasher, and R. Burke, "Web search personalization with ontological user profiles," in *Proc. 16th ACM Conf. Conf. Inf. Knowl. Manage.*, 2007, pp. 525–534.

[8] X. Li, C. Guo, W. Chu, Y.-Y. Wang, and J. Shavlik, "Deep learning powered in-session contextual ranking using clickthrough data," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014.

[9] S. Ge, Z. Dou, Z. Jiang, J.-Y. Nie, and J.-R. Wen, "Personalizing search results using hierarchical RNN with query-aware attention," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 347–356.

[10] S. Lu, Z. Dou, X. Jun, J.-Y. Nie, and J.-R. Wen, "PSGAN: A minimax game for personalized search with limited and noisy click data," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 555–564.

[11] J. Yao, Z. Dou, J. Xu, and J.-R. Wen, "RLPer: A reinforcement learning model for personalized search," in *Proc. Web Conf.*, 2020, pp. 2298–2308.

[12] Y. Zhou, Z. Dou, and J.-R. Wen, "Encoding history with context-aware representation learning for personalized search," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1111–1120.

[13] Y. Zhou, Z. Dou, Y. Zhu, and J.-R. Wen, "PSSL: Self-supervised learning for personalized search with contrastive sampling," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 2749–2758.

[14] C. Deng, Y. Zhou, and Z. Dou, "Improving personalized search with dual-feedback network," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, Tempe, AZ, USA, 2022, pp. 210–218, doi: 10.1145/3488560.3498447.

[15] Y. Zhu et al., "Contrastive learning of user behavior sequence for context-aware document ranking," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Queensland, Australia, 2021, pp. 2780–2791, doi: 10.1145/3459637.3482243.

[16] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, "Understanding the behaviors of BERT in ranking," 2019, *arXiv:1904.07531*.

[17] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[18] P. N. Bennett, K. Svore, and S. T. Dumais, "Classification-enhanced ranking," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 111–120.

[19] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang, "Enhancing personalized search by mining and modeling task behavior," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 1411–1420.

[20] T. Vu, D. Q. Nguyen, M. Johnson, D. Song, and A. Willis, "Search personalization with embeddings," in *Proc. Eur. Conf. Inf. Retrieval*, 2017, pp. 598–604.

[21] T. Vu, A. Willis, S. N. Tran, and D. Song, "Temporal latent topic user profiles for search personalisation," in *Proc. Eur. Conf. Inf. Retrieval*, 2015, pp. 605–616.

[22] M. Volkovs, "Context models for web search personalization," 2015, *arXiv:1502.00527*.

[23] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao, "Ranking, boosting, and model adaptation," Microsoft Res., Redmond, WA, USA, Tech. Rep. MSR-TR-2008-109, 2008.

[24] Y. Song, H. Wang, and X. He, "Adapting deep ranknet for personalized search," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, 2014, pp. 83–92.

[25] J. Yao, Z. Dou, and J.-R. Wen, "Employing personal word embeddings for personalized search," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1359–1368.

[26] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 2333–2338.

[27] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 101–110.

[28] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2042–2050.

[29] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 55–64.

[30] K. Hui, A. Yates, K. Berberich, and G. De Melo, "PACRR: A position-aware neural ir model for relevance matching," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Assoc. Comput. Linguistics, 2017, p. 1.

[31] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power, "End-to-end neural ad-hoc ranking with kernel pooling," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 55–64.

[32] K. Hui, A. Yates, K. Berberich, and G. De Melo, "Co-PACRR: A context-aware neural IR model for ad-hoc retrieval," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 279–287.

[33] S. Hofstätter, B. Mitra, H. Zamani, N. Craswell, and A. Hanbury, "Intra-document cascading: Learning to select passages for neural document ranking," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Canada, 2021, pp. 1349–1358, doi: 10.1145/3404835.3462889.

[34] X. Yu, W. Xu, Z. Cui, S. Wu, and L. Wang, "Graph-based hierarchical relevance matching signals for ad-hoc retrieval," in *Proc. Web Conf.*, 2021, pp. 778–787.

[35] C. Fu et al., "Leveraging multi-view inter-passage interactions for neural document ranking," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, 2022, pp. 298–306.

[36] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, and X. Xie, "NPA: Neural news recommendation with personalized attention," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 2576–2584.

[37] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, and X. Xie, "Neural news recommendation with attentive multi-view learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, S. Kraus, Ed., Macao, China, 2019, pp. 3863–3869, doi: 10.24963/ijcai.2019/536.

[38] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu, and X. Xie, "Neural news recommendation with long-and short-term user representations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 336–345.

[39] S. Okura, Y. Tagami, S. Ono, and A. Tajima, "Embedding-based news recommendation for millions of users," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Halifax, NS, Canada, 2017, pp. 1933–1942, doi: 10.1145/3097983.3098108.

[40] Q. Ai, D. N. Hill, S. V. N. Vishwanathan, and W. B. Croft, "A zero attention model for personalized product search," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Beijing, China, 2019, pp. 379–388, doi: 10.1145/3357384.3357980.

[41] K. Bi, Q. Ai, and W. B. Croft, "A transformer-based embedding model for personalized product search," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, China, 2020, pp. 1521–1524, doi: 10.1145/3397271.3401192.

[42] Q. Ai, Y. Zhang, K. Bi, X. Chen, and W. B. Croft, "Learning a hierarchical embedding model for personalized product search," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Shinjuku, Tokyo, Japan, 2017, pp. 645–654, doi: 10.1145/3077136.3080813.

[43] Y. Guo, Z. Cheng, L. Nie, Y. Wang, J. Ma, and M. S. Kankanhalli, "Attentive long short-term preference modeling for personalized product search," *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 19:1–19:27, 2019, doi: 10.1145/3295822.

[44] H. Wang, F. Wu, Z. Liu, and X. Xie, "Fine-grained interest matching for neural news recommendation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 836–845.

[45] K. Bi, Q. Ai, and W. B. Croft, "Learning a fine-grained review-based transformer model for personalized product search," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Canada, 2021, pp. 123–132, doi: 10.1145/3404835.3462911.

[46] F. Cai, S. Liang, and M. de Rijke, "Personalized document re-ranking based on Bayesian probabilistic matrix factorization," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Gold Coast, QLD, Australia, 2014, pp. 835–838, doi: 10.1145/2600428.2609453.

[47] S. Wang, Z. Dou, and Y. Zhu, "Heterogeneous graph-based context-aware document ranking," in *Proc. 16th ACM Int. Conf. Web Search Data Mining*, 2023, pp. 724–732, doi: 10.1145/3539597.3570390.
[48] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," 2013, *arXiv:1309.4168*.
[49] C. Burges et al., "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 89–96.
[50] Z. Dai, C. Xiong, J. Callan, and Z. Liu, "Convolutional neural networks for soft-matching N-grams in ad-hoc search," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 126–134.
[51] Y. Bai, Y. Zhou, Z. Dou, and J. Wen, "Intent-oriented dynamic interest modeling for personalized web search," *ACM Trans. Inf. Syst.*, vol. 42, no. 4, pp. 96:1–96:30, 2024, doi: 10.1145/3639817.
[52] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 3982–3992.
[53] J. Teevan, S. T. Dumais, and D. J. Liebling, "To personalize or not to personalize: Modeling queries with variation in user intent," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 163–170.

**Yutong Bai** is currently working toward the doctor's degree with the School of Information, Renmin University of China. Her research interests include information retrieval and data mining.

**Yujia Zhou** is currently working toward the PhD degree with the School of Information, Renmin University of China. His research interests include information retrieval and data mining.

**Zhicheng Dou** (Member, IEEE) is currently a professor with the School of Artificial Intelligence, Renmin University of China. His research interests include information retrieval and data mining.

**Ji-Rong Wen** (Senior Member, IEEE) is currently a professor with the School of Artificial Intelligence, Renmin University of China. His research interests include information retrieval and data mining.