Intent-oriented Dynamic Interest Modeling for Personalized Web Search

YUTONG BAI and YUJIA ZHOU, School of Information, Renmin University of China, China
ZHICHENG DOU*, Gaoling School of Artificial Intelligence, Renmin University of China, China
JI-RONG WEN, Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, China and Gaoling School of Artificial Intelligence, Renmin University of China, China

Given a user, a personalized search model relies on her historical behaviors, such as issued queries and their clicked documents, to generate an interest profile and personalize search results accordingly. In interest profiling, most existing personalized search approaches use "static" document representations as the inputs, which do not change with the current search. However, a document is usually long and contains multiple pieces of information, a static fix-length document vector is usually insufficient to represent the important information related to the original query or the current query, and makes the profile noisy and ambiguous. To tackle this problem, we propose building dynamic and intent-oriented document representations which highlight important parts of a document rather than simply encode the entire text. Specifically, we divide each document into multiple passages, and then separately use the original query and the current query to interact with the passages. Thereafter we generate two "dynamic" document representations containing the key information around the historical and the current user intent, respectively. We then profile interest by capturing the interactions between these document representations, the historical queries, and the current query. Experimental results on a real-world search log dataset demonstrate that our model significantly outperforms state-of-the-art personalization methods.

CCS Concepts: • **Information systems → Personalization**.

Additional Key Words and Phrases: Personalized Search, User Interest, Document Representation

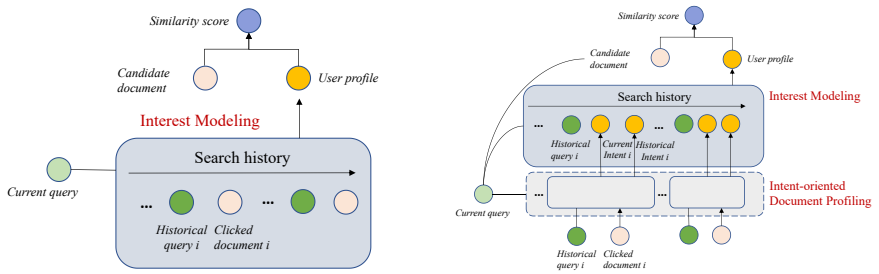
1 INTRODUCTION

Web search has become an activity most of us engage in by issuing a query to a search engine to get interested content. However, studies have shown that the query could be ambiguous, and different users have different interests under the same query [12, 28]. By taking different users' information needs into consideration, personalized search is an effective way to address the above problems. It utilizes the user's historical search behaviors to build and update an user profile reflecting her interest, so that the search engine could return an adapted result list according to the profile. Many traditional methods of personalized search focus on analyzing user query logs to extract personalized features. They mainly extract click-based features and topic-based features to model user interest [2–4, 8, 17, 20, 43]. With the emergence of deep learning, many researchers attempted to better build user interest profiles by automatically learning the representations of queries, documents, and users to

*Zhicheng Dou is the corresponding author

Authors' addresses: Yutong Bai, ; Yujia Zhou, zhouyujia@ruc.edu.cn, School of Information, Renmin University of China, Beijing, China; Zhicheng Dou, dou@ruc.edu.cn, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; Ji-Rong Wen, jrwen@ruc.edu.cn, Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, Beijing, China and Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China.

ACM 1046-8188/2024/1-ART
https://doi.org/10.1145/3639817



(a) The general user interest modeling paradigm in previous studies

(b) The simplified paradigm of our intent-oriented user interest modeling

Fig. 1. The comparison of previous user interest modeling strategy and ours. i refers to the ith search behavior in the user history. Existing methods ignore the impact of either the current search or the corresponding search during the document representation stage.

capture search interests from sequential or contextual perspective [19, 29, 31, 60, 65]. These works have shown significant improvement in search quality.

Most previous studies aim at building a user interest profile by modeling the history and highlighting relevant behaviors based on the current query. They then re-rank the search results by calculating similarities between the user profile and candidate documents. Each document (i.e., each historically clicked document and candidate document) is represented in a static fix-length vector. For example, Ge et al. [19] used the weighted average of the word representations to represent a document. Zhou et al. [66] used transformers to further encode the document together with the corresponding query to build the past interest, but they still simply took the fixed document title to represent a document. As the general user interest modeling paradigm illustrated in Figure 1(a), no matter which encoding method is used, the document representation is only dependent on the document itself, regardless of the corresponding historical query leading to the click, and also the current query. We argue such a static representation is problematic in personalized search. Since a document is usually long and contains multiple pieces of diverse information, the static fix-length document vectors used in the above methods would inevitably include broad and noisy information about the document, and make the final user profile cluttered and inaccurate. Because of this, we need to represent a document in a more effective way – paying more attention to the important parts and de-emphasizing other parts.

In fact, the important pieces of the same document when modeling historical and current interests are rather different. When a historical query is issued, the user will most probably pay attention to the passages relevant to the query, other than all content in the document she clicked. In the meanwhile, in the future when the user looks for new information, the historical information pieces that are relevant to her new information need are also important. An example is shown in Figure 2. At one search in the history, the user enters the query "iPhone" and clicks the document with a description of the phone in the first lines shown in this example. In the current search, the user would like to know information about "Apple chip" perhaps because the content in the latter part of the clicked document triggers her interest or simply reveals another aspect of her interest. In either case, the clicked document i should be considered different for its contributions to the depicting of the current intent. These observations of intent variations on documents inspire us to build intent-oriented document representations for better user interest profiling. In this paper, we attempt to explicitly extract the two sides of information during the document representation stage. The general idea is illustrated in Figure 1(b). Unlike previous studies, we build dynamic user interest by building two intent-oriented document representations for each document: (1) a document representation concentrated on the corresponding user intent when it was clicked.

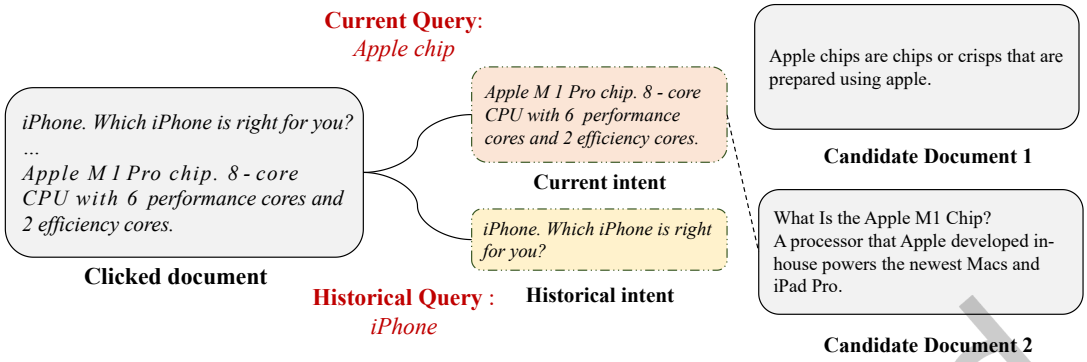


Fig. 2. Example user search behavior. The clicked document contains pieces related to the corresponding intent and pieces related to the current intent. The current intent-related pieces extracted from the clicked document provide supplemental information to clarify the current needs, as well as boost the relevance of candidate document 2.

We use the corresponding query leading to the click to guide the generation of this representation. (2) a document representation focusing on the user's current search intent. Accordingly, we use the current query to highlight the relevant parts of the document. We argue that by using these types of intent-oriented representations, we can better model historical behaviors and consequently generate better user profiles.

More specifically, we propose a personalized model DIMPS (Dynamic Interest Modeling for Personalized Search), which builds dynamic document representations with the influence of the user's historical intent and the current intent. It consists of two modules: a dynamic intent-oriented document encoder and a dynamic interest modeling module. (1) The **dynamic intent-oriented document encoder** module generates two types of dynamic document representations separately centered on the historical intent and the current intent. By using the historical query and the current query as guidance, the relevant information pieces at the passage level in the document are dynamically captured. (2) After the dynamic document representations are generated, the **dynamic interest modeling** module then uses transformers to capture the full interactions between these historical queries and their clicked documents, the current query, and also the candidate document. In this way, the current query plays a key role in extracting features from the behavior sequence for the interest profile depicting, while the candidate document can directly attend to the intents revealed in the historical representations for its relevance estimation. (3) **At last**, we learn a dynamic user profile and calculate a ranking score for the candidate document based on this.

Experimental results on a real-world search log dataset demonstrate that our model significantly outperforms the state-of-the-art personalization methods. With the dynamic document representations, we can model user interest more accurately.

In conclusion, our main contributions lie in the three aspects:

(1) We propose building dynamic intent-oriented interests to model more accurate user profiles for personalized web search.

(2) We devise a dynamic intent-oriented document encoding module to generate dynamic representations separately centered on the corresponding historical intent and the current intent.

(3) We apply a dynamic interest modeling module that attends the interactions among the search sequence, the current query, and the candidate document to dynamically depict the user's current search interest.

(4) We design a model-free document pruning algorithm that extracts interest-related features from documents. It facilitates the learning of user interest for traditional methods using entire document text.

The rest of paper is organized as follows. In Section 2 we summarize related works. In Section 3 we describe the proposed dynamic interest modeling strategy in detail. The experimental setup is discussed in Section 4. In Section 5 and Section 6 we present and analyze the results. The conclusion is drawn in Section 7.

2 RELATED WORK

2.1 Personalized Search

Some traditional personalized works focus on modeling click behaviors in query logs [17, 48]. For example, Dou et al. [17] proposed the P-Click model to calculate the click probability by counting historical click numbers on documents. Moreover, some works [8, 20, 49, 55] view topic-based features as their essential features, and employ models as Latent Dirichlet Allocation LDA [6, 55] to build user profile in topic space. Later, some researchers utilized learning to rank algorithms to combine these features. Great improvement has been shown in search results with the utilization of the advanced ranking algorithm LambdaMART [4, 57].

However, features adopted in most traditional personalized methods suffer from the heavy burden of manually extracting and the limitation in scope and variety. The emergence of deep learning enables us to automatically learn distributed representations from query logs. For personalized search, it facilitates the modeling of potential user interest [15, 31, 52, 60, 62–64]. A line of works aimed at modeling interest by excavating sequential information. Ge et al. [19] devised query-aware hierarchical recurrent neural networks to model sequential information and generated a dynamic user profile with query-aware attention to highlight related interests. Ma et al. [32] focused on leveraging fine-grained time information associated with user actions. They designed time-aware LSTM architectures for short-term interest where subtle interest evolution of users are captured, and calculated re-finding influences for long-term interest. Some works concentrated on leveraging context information in user interactions to build interests. Li et al. [29] generated semantic features from in-session contextual information through deep-learning models, and incorporate those features to current re-ranking models. Zhou et al.[65] proposed clarifying users' information needs by encoding history with context-aware representation learning.

The key of the personalized search task is to depict user interest from the history sequence, which evolves over each time of search behaviors. Properly tackling such dynamic features would benefit the improvement of ranking results. There exists a group of works that has shown good ability at dealing with dynamic features. Rossi et al. [38] proposed a Temporal Graph Network for continuous-time graphs represented as sequences of timed events. With the help of memory modules and graph-based operators, it successfully produces the embedding of the graph nodes at each time. This network has been experimented on future edge prediction tasks and dynamic node classification tasks and has obtained adorable results. Nonetheless, similar attempts have been made in the field of community-level information pathway prediction, whose goal is to predict the transmission trajectory of content across online communities. Jin et al. [22] devised a dynamic graph to capture the temporal variability across communities and model the time-aware propagation patterns of content information.

However, directly applying these advanced dynamic approaches to personalized web search would suffer a great performance drop. These approaches focused on capturing the temporal variability, while in personalized search, clicked documents contain multiple information pieces and cannot reflect accurate user intents. Failure in intent identification will limit the improvement of these time-aware dynamic approaches. These approaches do not explicitly address such intent variations and expect the model to automatically extract useful features for the current search from the whole updated sequence.

In this work, we observe the intents revealed by documents change between historical search and current search. As a result, we focus on dealing with such dynamics in user intents, rather than the dynamics in temporal information, to present historical representations with accurate reflections of corresponding intents at each time of search behaviors.

In fact, many personalized search tasks have also paid attention to the dynamics of intents over the history sequence. For example, Ge et al. [19] built dynamic user profiles by using the current query-aware attention to emphasize important historical behaviors. Zhou et al. [65] joined the current query into transformers to better capture its correlations with user history. But all the aforementioned personalized search methods merely impose the impact of the current search upon fixed document representations. This prevents the model from further identification of the user intents at different searches. Nonetheless, Bi et al.'s study [5] on the personalized product search task has shown the effectiveness of using the current query to attend user historical interacted items and current candidate items at a fine-grained level. Although simply migrating such an approach into personalized web search would limit the performance improvements, the clicked documents not only are too long but also cannot reflect accurate intents, and letting the current query attend to each word from documents is rather ineffective. It still encourages us to pose the impact of varying intents to the details of documents, rather than the whole content of them.

Based on the idea of capturing dynamics of user intents from more detailed evidence of documents, we devise an intent-oriented document profiling approach to build a more accurate user profile. In this work, we pose the influence of changing queries on document passages which has not been studied by previous personalized web search works.

2.2 Intent Identification

Query intent identification has long been actively researched in ad-hoc IR diversity tasks to address query ambiguity. Diversification approaches can be broadly categorized by whether or not explicit query intent representations are used [40]. The intent-explicit approaches formulate the query intents as explicit query aspect spaces, which are either directly given as genres or formed through techniques like matrix factorization [1, 39, 45]. While the intent-implicit approaches are typically based on inter-document similarity assuming dissimilar documents address diverse tastes [51, 61].

The notion of user intent was also introduced by Vargas et al. [46] in the recommender systems. It describes the uncertainty of user interests under the assumption that interests are associated with multiple sides and subareas. Exploring user intents has been a popular topic in a set of IR tasks. Early studies [46, 54] relies on item features to model user intents. Vargas et al. [46] considered two scenarios in which item feature data is explicitly known or can only be obtained through matrix factorization from the user-item preference data. Wasilewski et al. [54] injected a user perspective into items by incorporating a personalized intent-aware framework into the item-based recommendation algorithm, which is implemented by applying personalized covariance into the item similarity measure. Kaya et al. [26] proposed an alternative approach to model user intents without the use of item features. Instead, they used subsets of liked items, defined as subprofiles, to represent aspects of user tastes. The subprofiles are detected based on the nearest neighbors of like items and then sent to a subprofile-aware diversification framework.

Recent recommendation studies [9, 10, 34, 53] have observed that user demands change as search contexts evolve and have been making an effort to tackle such dynamics. For session-based recommendation, such dynamics are captured by exploiting users' recent behaviors with time information. MCPRN [53] proposed mixture-channel purpose routing networks where distinct purposes underlying items are learned dynamically in different channels. ICM-SR [34] designed an intent-guided neighbor detector to better select and leverage collaborative information from correct neighbor sessions. Further, in sequential recommendations where longer user history is considered, modeling user intents has also been an effective approach. IDSR [9] addressed the diversity in sequential recommendation with an intent-aware diversified sequential recommendation model. It introduced an implicit intent mining module, which includes a multi-intent attention mechanism where attention functions corresponding to particular intents are employed in parallel, to extract multiple user intents from

the user behavior sequence. Further, it designed an intent-aware diversity-promoting loss function to update the model regarding the diversity task. ICL [10] proposed a general learning paradigm called intent contrastive learning to leverage latent intents in the sequential recommendation. It models the user intent through a latent variable and learns the intent representation by learning the distribution function via clustering. Then, it fuses the learned intents into sequential recommendation models with contrastive self-supervised learning, which maximizes the mutual information between a view of the sequence and its corresponding intent.

To solve the personalized search problem, we intend to capture the user's intents in terms of the corresponding historical and current ones, with the dynamics across time taken into consideration. To be more specific, we excavate user intents from clicked and candidate documents by intent-oriented representation. The extracted intents are attached with temporal information to form an intent sequence for the final interest modeling.

2.3 Long Document Modeling

As for document modeling in ad-hoc search, dividing documents into fix-sized windows is a regular way to handle contents with varying lengths. This passage-level evidence makes it possible to greatly outperform traditional IR systems [11, 30]. Despite the effectiveness of traditional per-passage models, their significant computational cost remains a heavy burden. To address this problem, Hofstätter et al. [21] proposed an Intra-Document Cascading Model (IDCM) which first prunes passages of a candidate document with a less expensive model before running a slower scoring module. In news recommendation, precisely representing the news serves as a core way for user interest modeling and matching. Gao et al. [18] regarded titles, bodies, and topic categories as different views of news documents, and designed an attentive multi-view learning model to learn unified news representations. Wang et al. [50] hierarchically constructed multi-level representations for each news via stacked dilated convolutions, as an attempt to perform fine-grained interest matching from browsed news and candidate news. Nonetheless, in personalized search, most studies represent the documents by simply summarizing the word embeddings of titles or bodies.

In this work, we introduce long document modeling into personalized web search. To reduce the noise and computational cost, the documents are first divided into fixed-sized passages and then pruned according to the relatedness with historical and current intents. We then enhance the intent-related features in the two views to generate intent-oriented document profiles.

2.4 Attention Mechanisms

Attention mechanisms enable neural networks to model the dependencies in sequences regardless of distances. Recently many of the IR methods have actively integrated attention mechanisms to extract important features from long sequences. Wu et al. [56] employed personalized attention in news recommendations to model the informativeness of each news in terms of different user tastes. At the word level, after generating a preference query from the user ID, they computed the attention weights between news word representations and the preference query to determine the informativeness of each word. For the whole clicked news sequence, they again applied personalized attention to determine the informativeness of each piece of news. In personalized product search, whose object is to deliver adjusted product lists given user history, attention mechanisms could serve as an effective technique when building interests upon history. For example, Shen et al. [42] calculated the attention weights of historical behaviors in terms of the current query and the target item respectively, to extract accurate current interests. Similarly, in personalized web search, Ge et al. [19] assigned query-aware attention weights for the behavior representations, which are computed from recurrent neural networks, to build dynamic user profiles.

The self-attention mechanism, proposed by Vaswani et al. [47], captures the contributions of each part in the input sequence for generating the output sequence. It is formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \tag{1}$$

where Q, K, and V represent the query keys and values in the input sequence, while $\frac{1}{\sqrt{d}}$ is the scaling factor. Without the conjunction of recurrent networks and convolutions that are commonly used in attention mechanisms, it can draw global dependencies with less computational cost. Such natures make it suitable for most sequence modeling problems in IR tasks. A group of studies utilize the self-attention techniques in text encoding. Dai et al. [13] leveraged the BERT [16], a self-attention-based language model, to understand the text content of queries and documents with contextual information. In personalized tasks, Zhou et al. [65] used the transformer architecture [47], which is built solely upon self-attention mechanisms, to clarify information needs by learning context-aware representations for queries. Another line of studies models users' historical behaviors with self-attention mechanisms. In sequential recommendation, [24] aimed at capturing long-term semantics of user history even with sparse recent records. The self-attention mechanism perfectly served this goal by identifying relevant actions from a user's history. For personalized item search, authors in [27] proposed multi-resolution attention where relations between queries and past interactions are captured across different temporal subspaces. Through this way, the self-attention mechanism effectively retrieves historical information that is relevant to users' current search intents.

In this article, we integrate attention mechanisms in an end-to-end intent-oriented interest modeling framework. We propose a document pruning module with query-aware attention to extract corresponding user intents from documents. We also design an interest modeling module where past behaviors are modeled with self-attention so that the contributions of each behavior representation for generating the final interest profile are well captured.

3 METHODOLOGY

Personalized search has significantly improved users' search results by capturing their real information needs for document re-ranking. As we stated in Section 1, most of the existing methods learn user interests by extracting features from user behaviors, but their static representations do not distinguish the broad and noisy information pieces in documents. This further hinders the procedure of learning accurate user interests. In this paper, we focus on building dynamic intent-oriented document representations to better model user profiles. Specifically, with the corresponding query and the current query as guidance, intent-oriented document representations regarding the historical intent and the current intent are generated by the dynamic document encoder. Then, in the dynamic interest modeling part, we organize these query and document representations into a sequence and capture their interactions between the current query and the candidate document to depict an interest pattern to judge the documents' relevance.

The main notations in this paper are summarized in Table 1. To begin with, the problem is formulated as follows. Suppose that a user has a search history defined as a sequence of queries and their clicked document sets. $H = \{(q_1, d_1), \ldots, (q_N, d_N)\}$, where N is the number of queries. The document set d_i consists of a title list and a body list. Note that a query may have multiple clicked documents, we respectively concatenate their titles and bodies to form the two lists. Considering a user's current query q and a candidate document list returned by the search engine, our objective is to re-rank each document c in the list according to the search history H and the current query q.

As shown in Figure 3, our proposed model DIMPS consists of three components: (1) dynamic intent-oriented document encoder; (2) dynamic interest modeling; (3) re-ranking. In the following sections, we will elaborate on the structure details.

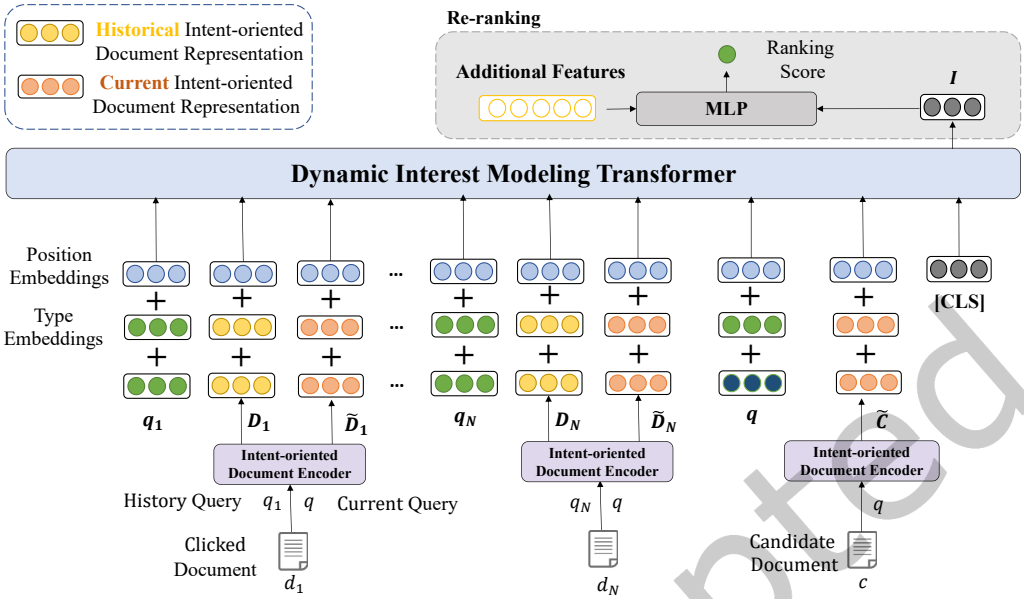


Fig. 3. The architecture of our proposed model is named DIMPS. Given each query-document pair in the search log, we first send it to the intent-oriented document encoder. Passage-level evidence is selected and modeled with the guidance of the corresponding query and the current query respectively to generate historical intent-oriented representations and current intent-oriented representations. The details of the document encoder are illustrated in Figure 4. Together with the historical query representation, the three vectors compose the dynamic behavior representation. We then send them to a sequence-level transformer to capture their interactions with the current query the and candidate document and learn a dynamic user interest profile. Finally, we can calculate the candidate document's ranking score from the learned profile and additional features.

3.1 Dynamic Intent-oriented Document Encoder

As we stated in Section 1, the fix-length static document representations include much noise hindering us from capturing real user interests. To solve this problem, we intend to build intent-aware document representations. Based on the observation that some search behaviors encompass information not only about the corresponding search but also the future search, for each document we extract features centered on historical and current intents in two separate parts. Each part is composed of two modules: (1) document pruning, where the body of each document is divided and further pruned by passage according to the relevance with the historical or the current intents; (2) document encoding. It aggregates the selected passage-level evidence to reflect the corresponding search intent. At last, we obtain one historical intent-oriented document representation and one current intent-oriented document representation. The structure of this module is illustrated in Figure 4.

3.1.1 Document Pruning. Documents returned by search engines contain abundant information, but could have too much noise for the learning of interest-related features. Also, they can be too long to deploy in real applications. Thus, aimed at extracting intent-oriented parts and reducing the query latency, we devise an offline document pruning mechanism to extract parts most related to the user's historical intent and the current intent respectively. Some previous studies [7, 25] have shown the effectiveness of organizing documents by fix-sized passages. Moreover, such passage-level evidence has been widely used in ad-hoc search [21, 41]. Similarly, we divide the document bodies into these passage-level information pieces.

Table 1. Summary of the Main Notations

Notation	Description
H	user's historical search sequence
N	the number of search behaviors in the user history
q	user's current query
c	the candidate document
q_i	the query of the ith search in the user history
d_i	the clicked document of the ith search in the user history corresponding to the query q_i
T_i	the title of document d_i
$P_{i,j}$	the jth passage of document d_i selected by our **historical** intent-oriented document pruning part
$\widetilde{P}_{i,j}$	the jth passage of document d_i selected by our **current** intent-oriented document pruning part
K	the number of passages in each intent-oriented pruned document
$\mathbf{q}$	the representation vector of the current query q
$\widetilde{\mathbf{C}}$	the representation vector of candidate document c
$\mathbf{q}_i$	the representation vector of the historical query q_i
$\mathbf{T}_i$	the representation vector of title T_i
$\mathbf{P}_{i,j}$	the representation vector of passage $P_{i,j}$
$\widetilde{\mathbf{P}}_{i,j}$	the representation vector of passage $\widetilde{P}_{i,j}$
$\mathbf{D}_i$	the **historical** intent-oriented representation vector of document d_i
$\widetilde{\mathbf{D}}_i$	the **current** intent-oriented representation vector of document d_i
$\mathbf{I}_H$	the historical interest sequence formed by historical query and document representations
$\mathbf{I}$	the learned user profile

To be specific, considering a user-issued query q_i, we first join the titles and bodies of its clicked documents, which may be more than one. For document titles, the first words of a certain number in the title lists are retained, denoted as T_i. For document bodies, we divide them into multiple partially overlapping windows with a fixed size, which results in a passage set. As described in Section 1, in this module we aim to build intent-oriented document representations. Intuitively corresponding queries will be good evidence for selecting intent-related passages. Thus, the pruning part is implemented by selecting the top K most relevant passages with the corresponding query. For each passage, we apply TF-IDF and calculate two relevance scores. The procedure for calculating the relevance score regarding the historical intent can be formulated as follows:

$$\text{score}(P_{i,j}, q_i) = \sum_{t \in q_i} P\left(t | P_{i,j}\right) = \sum_{t \in q_i} \frac{tf_{t,P_{i,j}}}{M} \lg(\frac{|C|}{cf_t + 1}), \tag{2}$$

where $P_{i,j}$ denotes the j-th passage of the document d_i, while $\text{tf}_{t,P_{i,j}}$ is the count of term t in it. cf_t is the corpus frequency of t. $|C|$ is the length of the corpus, and q_i symbolize the corresponding historical query leading to the document. When calculating the score of the current intent just substitute q_i with the current query q. Note when pruning candidate documents q_i is q.

For each document d_i, after the pruning stage as we get its title T_i and two sets of top K passages: $\left\{P_{i,1}, \ldots, P_{i,K}, \right\}$ and $\left\{\widetilde{P}_{i,1}, \ldots, \widetilde{P}_{i,K}\right\}$, respectively relevant to the historical intent and the current intent. After encoding them with

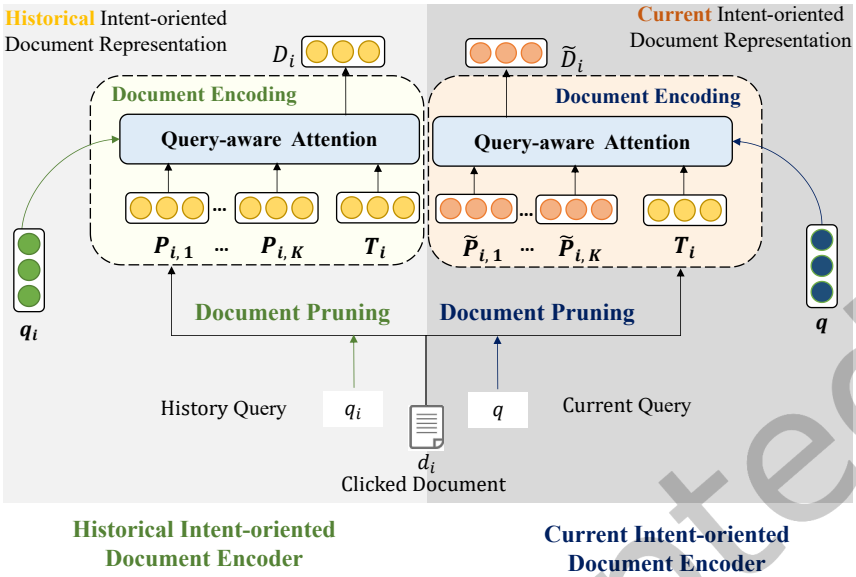


Fig. 4. The architecture of the dynamic intent-oriented document encoder. It is composed of two parallel parts where historical intent and current intent are respectively modeled. Given a document, we first send it to two separated pruning modules where passages are selected according to their relevance with the historical query or the current query. Then the selected evidence is further encoded with the guidance of the corresponding query to reflect the specific search intent. Finally, we obtain a historical intent-oriented document representation and a current intent-oriented document representation for each document.

the Sentence-BERT model [37], we have the pruned document formulated as:

$$d_i^p = \left\{\mathbf{T}_i, \left\{\mathbf{P}_{i,1}, \ldots, \mathbf{P}_{i,K}, \right\}, \left\{\widetilde{\mathbf{P}}_{i,1}, \ldots, \widetilde{\mathbf{P}}_{i,K}\right\}\right\}. \tag{3}$$

The pruned results of the candidate document c are similar to that of historical documents. Note that to cut down the computational cost, during the model implementation, the whole procedure of document pruning is performed offline.

3.1.2 Document Encoding. As previously discussed, documents have many irrelevant parts that may hinder us from learning the real search interest. Intuitively, the corresponding queries can serve as guidance for intent-related feature extraction. Hence we aggregated the passage-level evidence with the influence of the corresponding query.

Take the current intent-oriented behavior encoding part as an example. Considering the document d_i, we assign query-aware attention weights based on the current query vector $\mathbf{q}$, which is generated by encoding the q with the Sentence-BERT model, each current intent-oriented passage-level information pieces (e.g., the current-intent oriented passages set $\left\{\widetilde{\mathbf{P}}_{i,1}, \ldots, \widetilde{\mathbf{P}}_{i,K}\right\}$ and the title $\mathbf{T}_i$ denoted as $\widetilde{\mathbf{P}}_{i,K+1}$ under this scenario) to further extract current interests. The weights for each passage-level evidence of the document d_i are calculated as follows:

$$\mathbf{e}_{i,j} = \phi(\mathbf{q}, \widetilde{\mathbf{P}}_{i,j}), \tag{4}$$

$$\alpha_{i,j} = \frac{e_{i,j}}{\sum_{m=1}^{K+1} e_{i,m}}, \tag{5}$$

where $\phi(\cdot)$ indicates the multilayer perceptron (MLP). Then the current intent-oriented document representation $\widetilde{\mathbf{D}}_i$ is computed by summarizing the weighted passage-level vectors:

$$\widetilde{\mathbf{D}}_i = \sum_{j=1}^{K+1} \alpha_{i,j} \mathbf{e}_{i,j}. \tag{6}$$

The procedure of computing the historical intent-oriented document representation $\mathbf{D}_i$ is merely the same as computing the current intent-oriented representation. Just substitute the current query vector $\mathbf{q}_i$and current intent-oriented passage set $\left\{\mathbf{P}_{i,1}, \ldots, \mathbf{P}_{i,K}\right\}$ for the corresponding historical query vector $\mathbf{q}$ and the historical intent-oriented passage set $\left\{\widetilde{\mathbf{P}}_{i,1}, \ldots, \widetilde{\mathbf{P}}_{i,K}\right\}$.

Finally, we have generated two dynamic intent-aware representations $\mathbf{D}_i$ and $\widetilde{\mathbf{D}}_i$ for every document d_i, and one dynamic current-intent oriented representation $\widetilde{\mathbf{C}}$ for the candidate document c. Next, we will organize these documents with their corresponding queries into a behavior sequence to learn a search pattern and re-rank documents.

3.2 Dynamic Interest Modeling

So far, we have one query representation and two intent-oriented document representations for each behavior. Next, we are going to model the interactions among these behaviors through a sequence-level user transformer to model the user interest profile. Additionally, we regard the combination of the current query and the candidate document as the predicted current behavior and append it to the behavior sequence. Including the current query facilitates the learning of features related to the current intent and the building of a more accurate user profile. In the meantime, directly capturing the relevance between the candidate document and intent-oriented history representations also benefits the judgment of the candidate document's relevance with the user interest. Specifically, first, to provide the search sequence with sequential information, we assign position embeddings for each vector. The positions indicate the number of behaviors appearing in the historical sequence. The larger the number the latter the behavior. In dynamic modeling, the positions of behaviors are used to indicate the sequential order within the history. Note that the three vectors from one behavior share the same position embeddings since they are from the same search. At last, we join the historical interest sequence $I_H = \left\{\left(\mathbf{q}_1, \widetilde{\mathbf{D}}_1, \mathbf{D}_1\right), \ldots, \left(\mathbf{q}_N, \widetilde{\mathbf{D}}_N, \mathbf{D}_N\right)\right\}$ generated from the historical click data with the current query q and the intent-oriented candidate document representation $\widetilde{\mathbf{C}}$ and then send them to a transformer encoder. Similarly, we take the output of the "[CLS]" token as the learned dynamic user profile I:

$$\mathbf{I} = \mathrm{Trm}^{\mathrm{CLS}}\left(\left[I_H, \mathbf{q}, \widetilde{\mathbf{C}}, [\mathrm{CLS}]\right]\right). \tag{7}$$

3.3 Re-ranking

Furthermore, similar to existing personalization methods [19, 59, 65, 66], we follow the idea of SLTB [4] and extract traditional click and topic features $f_{q,c}$ for every candidate document. We send the additional features to a multilayer perceptron (MLP) to compute the ad-hoc relevance score for the document. Next, to generate the final ranking score $p(c|q, H)$ we aggregate the ad-hoc relevance score and the learned user profile $\mathbf{I}$ with a MLP $\phi(\cdot)$ operation. Specifically, our output is:

$$p(c|q, H) = \phi\left(\mathbf{I}, \phi\left(f_{q,c}\right)\right). \tag{8}$$

Table 2. Statistics of the dataset.

Item	Statistics	Item	Statistics
number of days	58	average query length	3.25
number of users	33,204	number of sessions	97,858
number of queries	267,479	average click number per query	1.19

Until now, we have obtained the final personalized score $p(c|q, H)$. In training, we adopt the ranking algorithm LambdaRank, which is also popular used in modern personalized models like our baselines [19, 59, 65, 66], in a pair-wise manner. First, we generate training document pairs from query logs with satisfactory clicked documents as positive samples and skipped documents as negative samples. Our objective is to maximize the distance between the positive score and the negative score. Hence, we compute the final loss with the weighted cross-entropy between the true probability $\bar{p_{ij}}$ and the predicted probability p_{ij}:

$$loss = -\left|\lambda_{ij}\right| \left(\bar{p_{ij}} \log(p_{ij}) + \bar{p_{ji}} \log(p_{ji})\right), \tag{9}$$

where the weight λ_{ij} is the change of metric when swapping the positions of the candidate document c_i and the candidate document c_j. The predicted probability p_{ij} is calculated as follows:

$$p_{ij} = \frac{1}{1 + \exp\left(p\left(c_j|q, H\right) - p\left(c_i|q, H\right)\right)}. \tag{10}$$

4 EXPERIMENT SETUP

4.1 Dataset

The search log AOL [35] does not provide the documents' full content. The URLs in the log are from 2006 which are too old to crawl from the current Web. Besides, as the dataset is not published for academic search, another reason for its inaccessibility stems from a combination of confidentiality concerns and potential ethical implications associated with its use. We have prioritized ethical considerations and the protection of individual privacy, which align with our research principles and ethical guidelines. Hence, we evaluated our model on a real-world dataset sampled from a commercial search engine, referred to as 'B dataset' in the remainder of this paper. The basic statistics are shown in Table 2. It is a large-scale query log containing two months of click-through data from 1st January 2013 to 28th February 2013. Each query record is composed of a user ID, a query string, a query issued time, a session identifier, the top 20 retrieved URLs, their click labels, and dwelling times. We regard documents with a dwelling time longer than 30 seconds as clicked documents. As for dataset partitioning, we take the first six weeks as history and the last two weeks as experimental data. For each user, the split ratio of training and test set is 4:1 under the measurement of session number in the experiment data, while the last one-fifth sessions of the training set are taken as validation data. Users with less than 4 sessions in experiment data are abandoned to ensure an effective division of training and test datasets. Additionally, different from previous studies, we extract the body data of clicked and candidate documents to enrich the behavior information.

4.2 Baselines

To evaluate the performance of our model, we select several state-of-the-art ad-hoc search models and personalized search models as baselines. They are as follows:

KNRM [58]. It is an ad-hoc model that matches queries and documents based on their interactions by utilizing kernel-pooling to extract soft-match features.

Table 3. A Summary of the Baseline Models Discussed in the Article

Model	Description	Document Representation
Adhoc Search Model		
KNRM [58]	An ad-hoc model that matches queries and documents based on their interactions by utilizing kernel-pooling to extract soft match features.	An embedding layer to map each word in the document titles.
Conv-KNRM [14]	Based on KNRM to model n-gram soft matches with an additional convolutional layer.	An embedding layer followed by convolutional layers where filters compose n-grams from document titles.
BERT [36]	It concatenates query-document sequence and feeds it into the pre-trained BERT model.	An embedding layer to map each word in the document titles.
Personalized Search Model		
SLTB [4]	Based on learning to rank algorithms with click features, topical features, time features, and position features.	A representation function to map document URLs to sparse vectors, and a text-based classifier on titles and bodies to obtain topical features.
HRNN [19]	It uses hierarchical RNN and query-aware attention to exploit sequential information and generate a dynamic user profile.	Calculated as the weighted average of the title word representations, which are mapped by an embedding matrix, multiplied by TF-IDF weights.
PEPS [59]	It enhances personal word embeddings from the global word embeddings by taking her individual search history as the training data.	A global embedding matrix and a personal embedding matrix to map each word in the document titles.
HTPS [65]	It uses hierarchical transformers to encode history with context-aware representation learning to disambiguate the current query.	An embedding layer to map each word in the document titles.
	It applies a self-supervised learning framework with contrastive sampling to reduce the dependency of sufficient data.	An embedding layer to map each word in the document titles and bodies followed by a transformer-based encoder to generate document vectors.

Conv-KNRM [14]. It is devised based on KNRM to model n-gram soft matches with an additional convolutional layer. This model boosts the matching accuracy by learning contextual information of surrounding words.

BERT [36]. It concatenates query-document sequence and feeds it into the pre-trained BERT model. We take the representation of the "[CLS]" token from the last layer as the matching features.

SLTB [4]. It uses learning to rank algorithms to integrate click features, topical features, time features, and position features into the personalized task.

HRNN [19]. It uses hierarchical RNN and query-aware attention to exploit sequential information and generates a dynamic user profile based on the current query. This work highlights the more important sessions according to the present information need.

PEPS [59]. This work solves the problem of personalized search without building a user interest profile. It enhances personal word embeddings from the global word embeddings by taking her individual search history as the training data.

HTPS [65]. This model focuses on clarifying the user's information need by disambiguating the current query. It uses hierarchical transformers to encode history with context-aware representation learning to complete this idea.

PSSL [66]. This work aims at reducing the dependency of sufficient data in many personalized works through data representation enhancement. It applies a self-supervised learning framework with the technology of contrastive sampling. Note that it also utilizes document body contents to cover more details.

DIMPS. (**D**ynamic **I**nterest **M**odeling for **P**ersonalized **S**earch) Our proposed model with a detailed description in Section 3.

DIMPS w/o. c: As the baseline models do not model candidate documents in their personalization models, for a fair comparison, we exclude the candidate documents' participation in interest profiling. In this way, the benefits of DIMPS w/o. c only come from its personalization model rather than its more complex relevance modeling. Specifically, we discard the input of the candidate document at the user-level transformer and summarize the outputs of the history sequence I_H as the learned user profile. Then we compute the cosine similarity between the user profile and the candidate document representation as the personalized ranking score. Note that the current query is still sent to the dynamic interest modeling transformer. As the "w/o. c" model is designed to explore the effects of the candidate document in interest modeling, we reserve the influence of the current query.

A summary of the baseline models is shown in Table 3. Descriptions of their document representing approaches are also attached. Most works simply map documents through a word embedding matrix before sending them for user interest profiling. Besides, a large percentage of works, except for SLTB and PSSL, solely use titles to represent documents. We assume that it is the large content and cluttered information that impedes the further exploitation of documents. In this paper, we address this problem by dynamically extracting corresponding intents and current intents from the documents.

4.3 Model Settings and Evaluation Metrics

To achieve a balance between effectiveness and efficiency, we performed multiple experiments and set the final parameters as follows: The search history includes 10 search behaviors. The passage length is 31. The first 15 passages of each clicked document list are sent to the document pruning parts. The number of top-selected passages fed into each intent-oriented document encoding part is 3. The dimension of document and query vectors is 384. The transformer encoder is one layer with a hidden size 150. The number of heads in multi-head self-attention is 2. We train the model for 2 epochs to get a satisfactory result. The learning rate is 1e-4 for the first epoch and 1e-5 for the second epoch. To compare the performance of all baselines and DIMPS, we use Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Precision@1 (P@1) to evaluate the ranking results. Besides, to alleviate the position bias problem, we calculate the percentage of improved pairs (P-improve) on inverse document pairs following previous works [19, 31].

- **MAP:** The value describes the percentage of correctly predicted candidate documents, measured between 0 and 1. Higher values indicate more accurate results.
- **MRR:** The value measures the positions of the correct answers in the ranking list. Larger values indicate better ranking quality.

Table 4. Overall performance of all models. '†' indicates the model outperforms all baselines significantly with paired t-test at $p < 0.05$ level. Best results are shown in bold.

Type	Model	MAP	MRR	P@1	P-improve
Adhoc Search	Ori.	.7399	.7506	.6162	-
	KNRM	.4916	.5001	.2849	.0655
	Conv-KNRM	.5872	.5977	.4188	.1442
	BERT	.6232	.6326	.4475	.1778
Personalized	SLTB	.7921	.7998	.6901	.1177
	HRNN	.8065	.8191	.7127	.2404
	PEPS	.8221	.8321	.7251	.2545
	HTPS	.8224	.8324	.7286	.2552
	PSSL	.8301	.8398	.7338	.2688
Our	DIMPS	**.8421**†	**.8512**†	**.7532**†	**.2939**†
	DIMPS w/o. c	.8370	.8467	.7461	.2647

- **P@1:** The value measures the precision of the top 1 item in the ranking list. The higher the value, the better the ranking.
- **P-improve:** It is designed under the observation that the user tends to click documents higher in the list, which may leave relevant but low-ranked documents not examined. We measure the actual improvements on inverse document pairs [23] to alleviate such positive bias. Higher values represent larger improvements.

5 RESULTS

We evaluate the overall performances of all baselines and our DIMPS model on the B dataset. The results are reported in Table 4. We can observe that:

(1) **Our proposed dynamic interest modeling model DIMPS outperforms all the baseline models, including ad-hoc models and personalized models, with paired t-test at $p<0.05$ level on the B dataset.** Especially for state-of-the-art model PSSL, we gain significant improvements in terms of all metrics. Our model improves the results by 1.11% in MAP and 1.14% in MRR. Besides, it outperforms PSSL by 2.51% in the more objective metric P-improve. These results prove that building dynamic intent-oriented behaviors is a more effective way for personalized search.

(2) The "DIMPS w/o. c" also reduces the accuracy. We believe it verifies our assumption that learning current intents from candidate documents benefits re-ranking. Further, preventing the candidate documents from drawing connections to historical intent representations may lead to great information loss. However, it still outperforms the state-of-the-art method, proving the adorable ability of our dynamic document profiling approach to build user interest.

(3) Among all the personalized models that adopt the body content of documents, models with dynamic representation strategy (i.e., our methods) bring better performance gains than the model with static representation strategy (i.e., PSSL). Although the SOTA model PSSL also leverages the body content, it simply encodes the entire text as final document representations, which includes too much noise for the model to extract accurate user interests. This indicates that without the dynamic adaption to user intents, just incorporating bodies is not that efficient in improving personalized results.

(4) In general, all personalized search models outperform ad-hoc search models, which demonstrates the contribution of user behaviors for reflecting the real information need. In addition, it is noted that among all evaluation metrics, the improvement of P@1 is more significant than others. One possible reason is that by learning from users' search history, personalized models are more effective in handling re-finding behaviors, while in terms of other behaviors, their performance is limited due to the lack of relevant logs. Unlike most personalized search models which adopt static behavior representations to profile user interests, we focus on building dynamic interests by exploring the potential in user behaviors of providing information about both current and future interests.

In a word, it is proved that **dynamically modeling interests by building historical intent-oriented and current intent-oriented document representations is helpful for understanding user preferences and re-ranking documents**. To further analyze the functions of the main components in our model, we will conduct an ablation analysis and show a visualized example.

6 ANALYSIS

In this section, we conduct sets of experiments to deeply investigate the functionality of the major components in our DIMPS. To be more specific, we try to understand the following topics:

- The necessity and performance of the three intent-oriented components: document pruning, document encoding, and interest modeling.
- The effectiveness of the query-aware attention mechanism in the document encoding part for capturing corresponding intents, and the possibility of employing alternative passage aggregation patterns, like attending the inter-passage correlations as well.
- The effectiveness of capturing interactions among all the intent representations in the dynamic interest modeling stage. We wonder if the accuracy will deteriorate when abandoning or indirectly capturing such interactions.
- The effects of the number of selected passages in the document pruning part.
- The effects of the history length.
- The ability of our intent-oriented method to model interests when the queries are too ambiguous to express real user intents.
- If there exists a case supporting our assumption that dynamically representing documents according to historical and current intents will help build accurate user interests, and the DIMPS are capable of implementing this idea.
- The efficiency-effectiveness performance of the DIMPS.
- The possibility of applying the document pruning module on other personalized models.
- The possibility of applying the candidate document-aware interest modeling on other personalized models.

6.1 Ablation Analysis

Our dynamic interest modeling strategy is based on the intent-oriented document encoder, which includes the following parts: the historical intent-oriented document encoding and the current intent-oriented document encoding. To verify the contribution of each part, we conduct several ablation experiments. The results are illustrated in Table 5.

DIMPS w/o. DP. We delete the DP (document pruning) part in the document encoder.

DIMPS w/o. HIE. We strip off the HIE (historical intent-oriented document encoding) part in the document encoder, which results in the discard of the historical intent-oriented document representations in the dynamic interest modeling module.

Table 5. Performance of ablation models.

Model	MAP	MRR	P@1	P-improve
DIMPS w/o. DP	.8390	.8483	.7481	.2838
DIMPS w/o. HIE	.8400	.8496	.7514	.2857
DIMPS w/o. CIE	.8403	.8496	.7510	.2830
DIMPS	**.8421**	**.8512**	**.7532**	**.2939**

DIMPS w/o. CIE. We strip off the CIE (current intent-oriented encoding) part in the document encoder, which results in the discard of the current intent-oriented document representations in the dynamic interest modeling module. Note that for the candidate documents, we reserve its CIE part to reserve information.

The results of the ablation analysis are presented in Table 5. It is clear to be seen that all ablation models damage the results of the original DIMPS. When deleting the document pruning part, the model shows the worst performance. This proves that our passage selection algorithm does provide relatively clear and useful intent-oriented information for the model to learn. Whereas, DIMPS w/o. DP still outperforms baseline models with fixed document representations like PSSL, HRNN, and PSGAN with great improvements, which indicates the effectiveness of our dynamic interest modeling structure. As we prune the documents to make them adapt to the intents, the model DIMPS further promotes the re-ranking results. This illustrates that documents indeed contain information related to both the corresponding and future interests, and our document pruning mechanism offers a practical way to extract useful features.

As for the document encoding sub-module, without the historical intent-oriented encoding, the MAP, MRR, P@1, and P-improve metrics drop 0.21%, 0.16%, 0.18%, and 0.82%. This demonstrates the usefulness of extracting historical intent-related features in building user profiles. Moreover, the performance also loses significantly when the current intent-oriented encoding part is abandoned, with a decline of 0.18%, 0.16%, 0.22%, and 1.09% in MAP, MRR, P@1, and P-improve. This shows that clicked documents contain many features directly related to current intents, while explicitly modeling them from the passage level could support the model to better understand search needs and re-rank documents. It is notable that the current intent-related features, which are overlooked by previous studies, are proved to be as informational as the historical ones. Moreover, even when stripping one of the two components in our model, the results still obviously outperform the baseline model PSSL, demonstrating the effectiveness of our methods for modeling intent-oriented interest.

To sum up, stripping off any main component in the document representation procedure will impede the learning of user interest. Further, it is observed that extracting current intent clicked documents is as effective as solely extracting corresponding historical ones, which accords with our assumption: there exists rich information within clicked documents related to users' current intent but overlooked by previous studies.

6.2 Effects of the Query-aware Aggregation in the Document Encoder

To explore the function of the query-aware attention mechanism in the document encoder, for passage-level evidence we test our model with three different aggregation patterns:

(1) **w/o q**: the passage vectors and the title vector are simply summarized without the query-aware attention.

(2) **self-attention**: the passage vectors and the title vector are fed into a self-attention layer together with the corresponding query. The output vectors of the passages and the title are summarized to represent the whole document.

Table 6. Results with different aggregation patterns for passage-level evidence.

Model	MAP	MRR	P@1	P-improve
w/o. q	.8411	.8502	.7515	.2877
self-attention	**.8425**	**.8516**	**.7544**	**.2920**
self-attention w/o. q	.8407	.8502	.7520	.2883
DIMPS	.8421	.8512	.7532	.2939

Table 7. Results of different sequence-level interest modeling strategies

Model	MAP	MRR	P@1	P-improve
independent HC	.8358	.8449	.7426	.2895
compared	.8372	.8475	.7482	.2572
DIMPS	**.8421**	**.8512**	**.7532**	**.2939**

(3) **self-attention w/o. q**: the passage vectors and the title vector are fed into a self-attention without the corresponding query. The output vectors of the passages and the title are summarized to represent the whole document. The results are reported in Table 6.

As expected, models with query-aware attention obtain better accuracy, which shows the effectiveness of using corresponding queries to extract intent-related features in documents. The self-attention model achieves the best results. We believe this proves the interactions among passages are useful clues for the model to understand search interests and the corresponding query still provides critical information for the organization of such clues. To balance efficiency and effectiveness, we abandon the relatively expensive self-attention models in our DIMPS.

6.3 Effects of the Dynamic Interest Modeling

After the intent-oriented document representation, we have extracted historical and current intents from the history, the current query, and the candidate document. Next, we model user interests by attending to the dependencies among those intent representations. The reason for directly attending dependencies among the history, the current query, and the candidate document lies in two aspects:

- The contribution of different historical representations for interest profiling can be addressed with the help of the current representations (i.e., the current query and its potential clicked documents: candidate documents).
- The candidate documents may include complementary information for current intents which are useful for interest profiling.

In this section, we are going to explore the functionality of the aforementioned dependency modeling strategy. To be specific, we set three model variations as follows:

(1) **independent HC**: The relatedness of the current behavior to each historical behavior is not modeled. We discard the inputs of the candidate document as well as the current query, which represents current intents, at the user-level transformer. Instead, we take the summarized outputs of the history sequence I_H as the unified historical interest. For the current intents, we concatenate the candidate document and the current query and send them to another transformer. The outputs are summarized as the current interest. Further, we append it and a "[CLS]" token to the unified historical interest and send them to a transformer. The output of the "[CLS]"

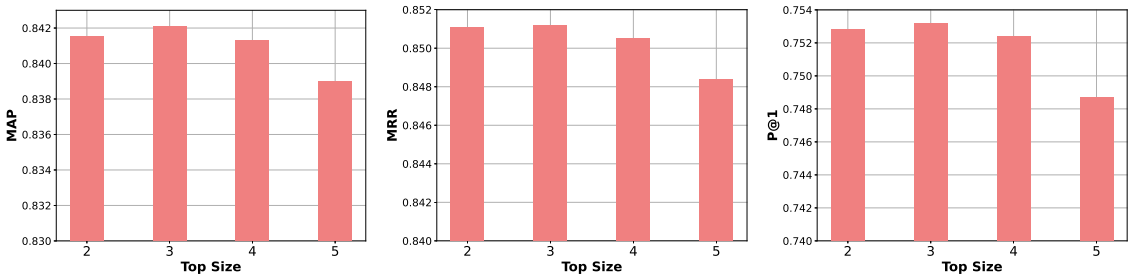


Fig. 5. Results with different numbers of selected passages.

token is taken as the predicted user profile. In the re-ranking stage, this predicted user profile is taken as the user profile **I** in Equation 8.

(2) **compared**: This experiment is designed to investigate how the impact of negative samples deteriorates the ranking quality, and the effectiveness of using the predict token. The dependency modeling strategy is the same with DIMPS, but without the notation of "predicted interest vector". We discard the input of the "[CLS]" token at the sequence-level transformer. Instead, we summarize the outputs of history sequences I_H (history-intent) as the user profile, and then compare its cosine similarity regarded the outputs of the candidate document for the personalized score.

The experimental results shown in Table 7 match our expectations. All three models damage the ranking quality of DIMPS. The "independent HC" gains the lowest accuracy, demonstrating that the model fails to take advantage of the intent-oriented information from the history without the attention to current behaviors. Whereas, with the communications among historical and current intents ensured, the "compared" model still suffers from considerable drops on all metrics. The major reason lies in the uncertainty of the candidate documents' satisfactoriness. If an unsatisfied candidate document is fed to the network, its representations will be inclined to 1) the correct current intent because of the intent-oriented document encoder, 2) and historical intents because the dynamic interest modeling poses the impact of the history to it. Therefore, as we would like to use candidate documents as complementary intent representations, unlike most user profile-based personalized methods, we set a "[CLS]" token to reflect how the predicted search sequence accords with user interests.

6.4 Experiments with Different Numbers of Selected Passages

To analyze our model's performance regarding the number of selected passages in the document pruning module, we set the top size K at 2, 3, 4, and 5 respectively. From Figure 5, we observe that, in general, more selected passages help the model to capture user intents. However, the accuracy deteriorates when setting the K larger than 3, indicating the inclusion of more noise when learning from long documents. Moreover, it also verifies the necessity of pruning documents according to corresponding intents as well as the functionality of our pruning strategy to extract such intents. As the balance of efficiency and effectiveness, we take the K for 3 in our DIMPS model.

6.5 Experiments with Different Numbers of Historical Behaviors

To investigate the impact of history length in DIMPS, we test the model with different numbers of historical search behaviors. The results plotted in Figure 6 show that when the behavior numbers are lower than 10, the accuracy grows as lengthening the history. However with a long history, when increasing the behavior number the accuracy tends to drop more significantly. Specifically, the MAP gap between behavior numbers 10 and 15 is

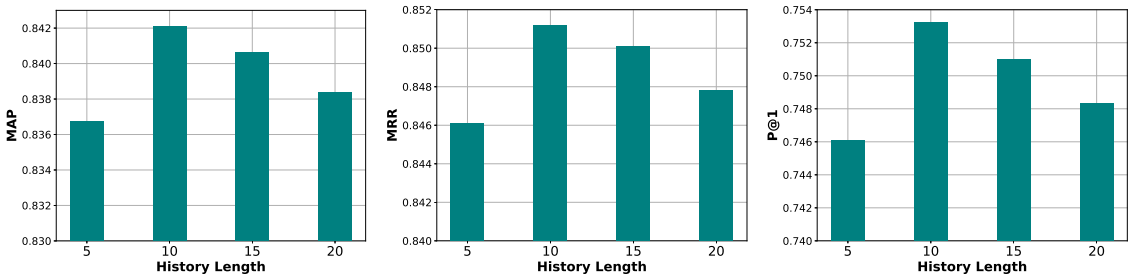


Fig. 6. Results with different numbers of historical behaviors.

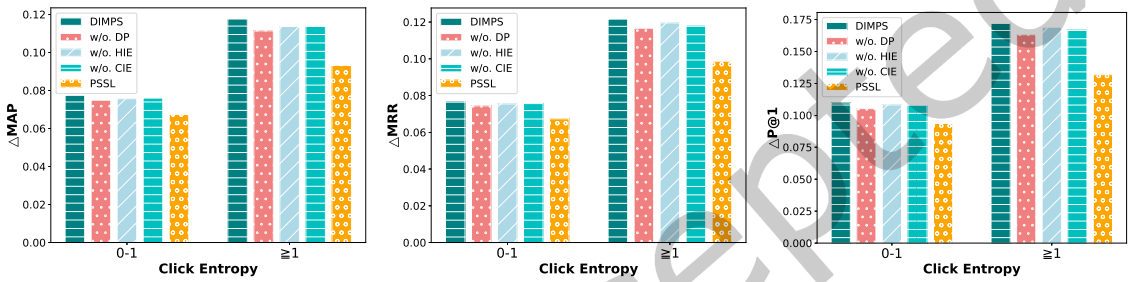


Fig. 7. Results on ambiguous and non-amibiguous queries

1.45%, while the MAP gap between behavior numbers 15 and 20 is 2.28%. We suppose this is because, in the long history, the text related to current intents in the clicked documents is more sparse, which results in much noise in the selected current intent-oriented passages. For a longer history, the intent-oriented document encoder should be further adjusted to ensure effective information extraction. In future work, we may exclusively design the intent-oriented document encoder for long-term history, with less expensive structures compared to the current short-term ones.

In addition, our lowest accuracy still outperforms the SOTA method PSSL, which requires 50 historical searches. This phenomenon also proves that our DIMPS does not require a long user history to achieve adorable personalized results.

6.6 Experiments with Ambiguous and Non-ambiguous Queries

We first divide the dataset into ambiguous and non-ambiguous queries, the intents of the former are more ambiguous with words that could be interpreted into diverse meanings like "MAC" while the latter are more clear. Click entropy [17] is an effective measurement for the queries' ambiguity. Previous studies [17, 44] have shown that a larger click entropy often indicates more potential for search results personalization because of the larger ambiguity. We categorize the queries with the cutoff of click entropy at 1.0. Figure 7 shows the performance improvement in MAP over BM25 by the state-of-the-art model PSSL, our proposed dynamic interest modeling model DIMPS, and the two ablation models.

It is noted that all personalized models boost the original ranking on both groups. Generally, the ambiguous queries (click entropy≥1) have higher improvements than non-ambiguous queries (click entropy<1). The results demonstrate the effectiveness of all methods for clarifying search intent. Besides, all the dynamic interest modeling

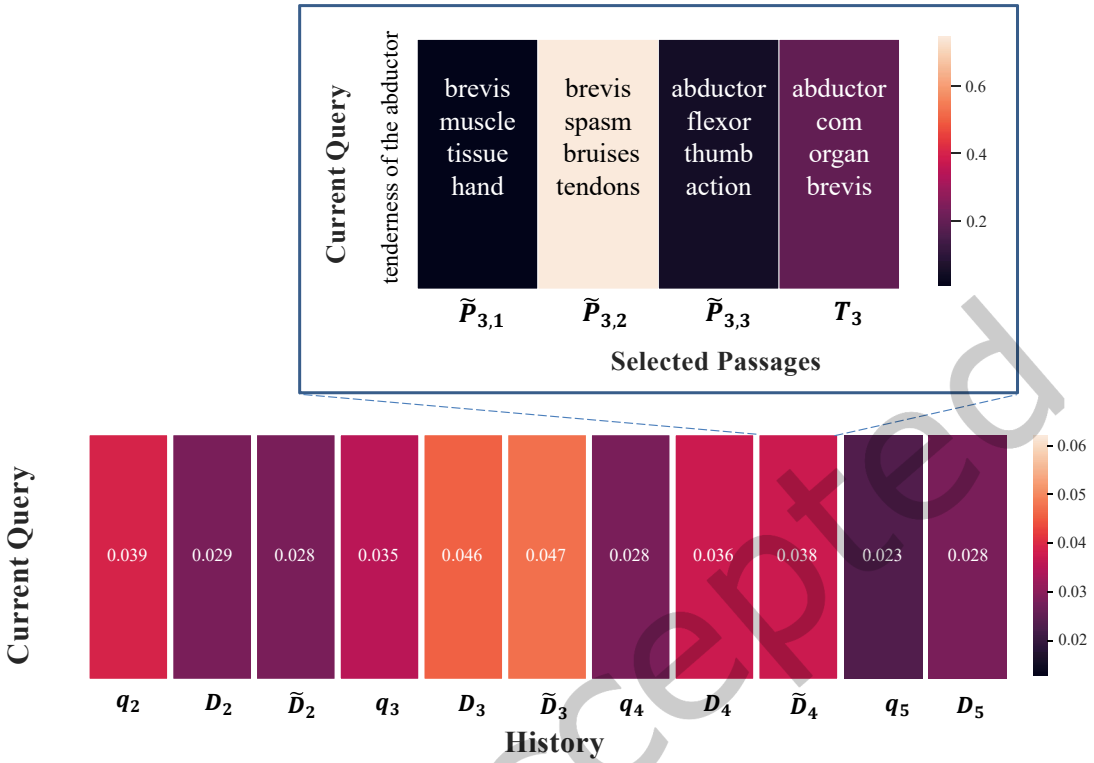


Fig. 8. The weights of a part of past search history applied by the current query "tenderness of the abductor". A lighter area indicates a larger weight. $\mathbf{q}_i$ represents the ith query vector in the user' search history. $\mathbf{D}_i$ refers to the historical intent-oriented document representation corresponding to the ith search. While $\widetilde{\mathbf{D}}_i$ is the current intent-oriented document representation of the ith search. $\widetilde{\mathbf{P}}_{i,j}$ is the jth passage-level intent-oriented representation from the current intent-oriented encoding part of the ith search. Corresponding keywords are shown in the sub-figure.

methods consistently outperform PSSL on both query categories. Specifically, our DIMPS outperforms the baseline model PSSL by 1.00% when click entropy<1 and 2.44% when click entropy≥1. Likewise, a greater drop can be seen when click entropy≥1 between DIMPS w/o. DP and PSSL. Similar performance gap is observed between the other two ablation models and the PSSL. This demonstrates the contribution of the proposed intent-oriented dynamic interest modeling strategy in disambiguating queries. Particularly, if we strip off the current intent-oriented encoding parts, the performance drops more noticeably on ambiguous queries than on non-ambiguous queries. This confirms that the clicked document contains informational clues for the deduction of users' current intent, which accords with our assumption. A similar phenomenon can be seen on DIMPS w/o. DP. We believe this shows the necessity of dynamically extracting supplemental pieces in the clicked history, especially under ambiguous queries.

Next, to verify the function of our intent-oriented document encoder and interest modeling module in detail, we will show a visualized example.

6.7 Case Study

As we stated before, static document representations are noisy and broad which hinders us from capturing accurate user intents and re-ranking documents. In this paper, we implement a dynamic intent-oriented document encoder and a dynamic interest modeling module to construct intent-aware interest sequences. Table 4 demonstrates the effectiveness of this kind of dynamic interest modeling in improving personalization. To further analyze how the representing and modeling work, we show an example by sampling one user's query log in the B dataset. We visualize the attention weights applied to all behavior representations by the current query in the final history-level transformer. Furthermore, we also represent the attention weights applied to passage-level representations by the current query in the document encoding part.

In this case, the user enters "tenderness of the abductor" to get information about a muscle problem. As shown in Figure 8, all three vectors from each behavior are assigned different weights by the issued query. We focus on the representations of search position 4 with the query word "hand muscle", which does not reveal direct relations with the current query. However, the clicked document includes much information about muscle problems, which means this is a behavior highly related to the current intent.

From Figure 8, we can observe that:

- The the postion 4's query representation $\mathbf{q}_4$ obtains a small weight, which accords with intuitions. This verifies that without dynamic document profiling, like previous methods the model cannot capture the interests underlying document documents.
- Both the postion 4' historical intent-oriented representation $\mathbf{D}_4$ and the current intent-oriented representation $\widetilde{\mathbf{D}}_i$ gain larger weights compared to its query representation. This indicates that our model successfully leverages the additional features from the intent-oriented document encoder, and correctly considers them contributing to depicting user interests.

Now we have proved that in dynamic interest modeling, the model successfully emphasizes the informational parts that benefit the user profiling. To further analyze how the relevance of the current intent is attended in the current intent-oriented encoding, in Figure 8, we also show the attention weights at position 4 among the three selected passage representations and the title representation. It is noted that:

- The second passage $\widetilde{P}_{3,2}$ that includes a detailed description of muscle problems related to abductor tenderness are applied the largest weights, which verifies our current intent-oriented encoding attends important information about current search intent from historical data.

It is obvious that either the document representation itself (consists of several passage representations), or its importance for interest modeling, is dynamically decided by the current query. From the weights of this case, we can see that such fine-grained modeling of a document depends on its contribution to user profiling. That means, our model is effective at fine-grained modeling of documents in personalization, leading to its adorable ability to improve personalized ranking quality.

In conclusion, the visualization of weights on passage-level intent-oriented representations demonstrates that our intent-oriented encoding module can extract features related to the current intent from the history data, while the weights of all behavior representations verify that our interest modeling module is capable of highlighting relatively important parts when building the search interest.

6.8 Efficiency-Effectiveness Analysis

To extract features related to the current intents, we profile each historical clicked document set according to the current query. This means we need to run the pruning part and the encoding part on each historical interaction. None of the computations about the current intent are reusable because each new query requires full recomputation. Hence, we pay great attention to reducing the computational cost. For example, the whole

Table 8. A Computational Complexity Summary of the Baseline Personalized Models Discussed in the Article

Model	Computational Complexity	Description
HRNN	$O(ND^2 + ND)$	$O(ND^2)$ is the complexity from its recurrent networks, and $O(ND)$ is from its query-aware dynamic attention.
PEPS	$O(L^2D + LD^2)$	The multi-head attention contextual representation has the complexity of $O(L^2D)$. $O(LD^2)$ is from the GRU-based query reformulation module.
HTPS	$O(NL^2D + N^2D)$	$O(NL^2D)$ is from its word-level transformer. $O(N^2D)$ is from its transformer encoder on history.
PSSL	$O(N^2D)$	It is from its transformer encoders on user history.
DIMPS	$O(NKD + N^2D)$	$O(NKD)$ is the complexity of the query-aware document encoding. $O(N^2D)$ is the complexity of the interest modeling.
DIMPS self-attention	$O(NK^2D + N^2D)$	$O(NK^2D)$ is the complexity of the self-attention document encoding. $O(N^2D)$ is the complexity of the interest modeling.

procedure of document pruning and passage encoding is performed offline, without the need to learn with current queries. The fast TF-IDF is used to extract query-related passages. The selected passages are also previously prepared by pre-trained deep learning models. During the learning procedure, we discard the modeling of inter-passage interactions but utilize simple query-aware attention to build document profiles. What's more, the model also shows its superiority over baselines even fed with a small number of historical behaviors.

All the above efforts lead to satisfactory results over efficiency as well as effectiveness. In this section, we first analyze the computation complexity in theory for all personalized baselines and the proposed model. Second, we compare the query latency of the DIMPS "self-attention" model, DIMPS without passage encoding, and the DIMPS original model, to represent the effects on computational cost in the document profiling stage. Third, we compare the query latency between the state-of-the-art model PSSL and our proposed model under different history lengths.

Overall Complexity Analysis.

We list the computational complexity of all the deep-learning personalized baselines as follows in Table 8. Note that in real applications is the query latency that affects the user, so we do not analyze the complexity of the pre-training or offline preprocessing stage of the models. N is the number of search behaviors in the user history. D is the representation dimension. K is the number of passages in each intent-oriented pruned document. L is the sequence length of queries or documents.

Table 9. Efficiency and effectiveness of DIMPS with different document encoding strategies.

Model	MAP	MRR	P@1	Queries per second
DIMPS self-attention	**.8425**	**.8516**	**.7544**	92,783
DIMPS Emb	.8407	.8502	.7520	**154,465**
DIMPS	.8421	.8512	.7532	101,968

Our DIMPS model has comparable computational complexity with all the baselines. Thanks to the offline document pruning stage, using document passages instead of titles does not bring unacceptable expenses.

Efficiency-Effectiveness Analysis for DIMPS

In this section, we exploit the efficiency and effectiveness of our DIMPS model with different document encoding strategies. Specifically, we test the following two model variants:

- **self-attention**. It is the model mentioned in Section 6.2. The query-aware attention in the document encoder is replaced by self-attention models. This model has shown a better ability to capture relevance among passages and gains better ranking results in Table 6. In this section, we will further analyze its efficiency.
- **Emb**. The transformer-based encoding part described at the beginning of Section 3.1.2 is abandoned. While, we generate passages and query vectors by mapping each word through a pre-trained embedding table, which is fine-tuned during the training, and summarize all the word representations. The embedding matrix is obtained by training a word2vec [33] model following [19, 31, 65]. The embedding size is 100. We set this experiment to explore our model's performance when the offline transformer encoding part is too expensive to employ in real applications.

We provide the costs of two models in Table 9. All the experiments are conducted on a 256 GB memory server with a single Titan V GPU.

As for the self-attention model, it is shown that the accuracy improvement is not significant but the cost grows more obviously. Besides, we also give theoretical analysis on efficiency in Table 8. It is shown that the computational cost of self-attention models, with the complexity of $O(NK^2D)$, gains considerable increase with more input passages. On the contrary, the DIMPS model with query-aware attention, with the complexity of $O(NKD)$, would be far less expensive. We did not test the model with a history longer than 20 or passage numbers larger than 3 because the document pruning part would cost much as these numbers grow. But as the document pruning can be performed offline, it is possible that in real-life applications these numbers would be much larger. At last, we adopt the efficient dynamic query-aware mechanism into DIMPS to ensure our model structure is not too complex to apply in real-life applications, where longer inputs could be preferred.

As for the Emb model, it yields better results on effectiveness and efficiency over PSSL in the next section. This proves that our model's superiority is not heavily dependent on the pre-trained passage encoding part. Even if generate passage vectors during training, the computational cost and accuracy are still adorable.

Moreover, the tf-idf document pruning part can be deleted to save cost since the DIMPS w/o. DP in Section 6.1 still achieves competitive performance over PSSL.

The Comparison between DIMPS and PSSL.

The comparison of query latency between our method and the state-of-the-art model PSSL is shown in Figure 9, where we test the two models under different history lengths. Historical behavior numbers are denoted in Figure 9. All the experiments are conducted on a 256 GB memory server with a single Titan V GPU. To ensure a fair

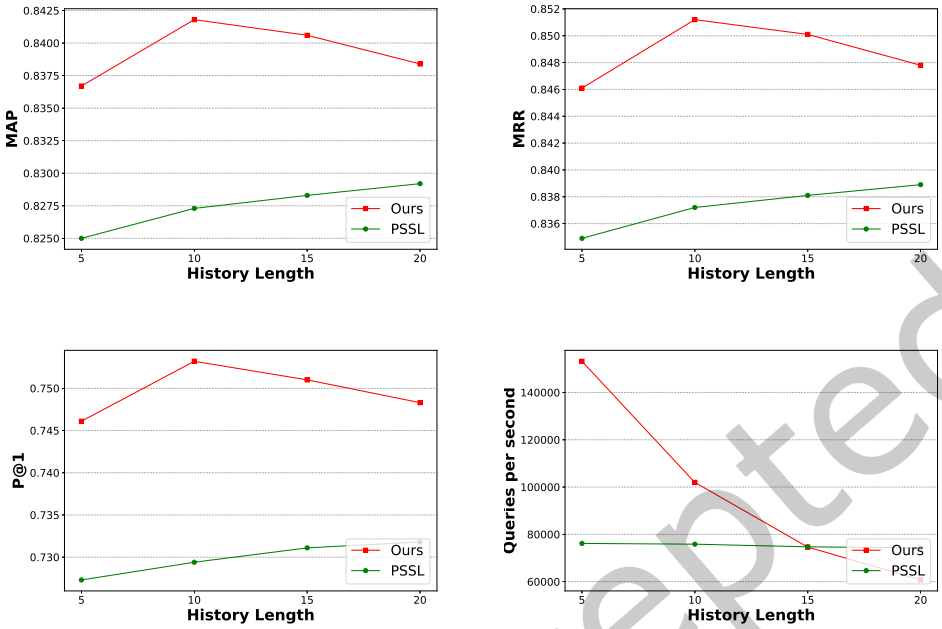


Fig. 9. Effectiveness and efficiency of our proposed model and PSSL with different history lengths. Compared to the SOTA model PSSL, our DIMPS model gains better accuracy and costs comparable computational resource.

comparison, the PSSL model is implemented by its authors' published code. The dataset used in this section is also the same as the one used in the PSSL paper.

We can see that our method has comparable query latency with PSSL, while significantly outperforming it in terms of ranking quality. As expected, modeling interest from more historical behaviors increases the computational cost. The increases of DIMPS are larger than PSSL, it is because the DIMPS needs to perform document profiling for each behavior, while PSSL only needs to add the behavior to the history-level transformer encoder. Whereas, all of the DIMPS models show obvious advantages in performance over the PSSL models. In this figure, the best efficiency-effectiveness balance is achieved with a history of length 10, which indicates our intent-oriented extraction strategy is more suitable for short-term history. We believe this is because a longer history tends to reflect more long-standing user characteristics, while its fine-grained evidence is not so informational for user profiling.

In summary, compared to PSSL, our model shows better effectiveness and comparable efficiency.

6.9 Applications of the Document Pruning

In this paper, we design a document pruning approach to prune document contents according to the queries. In this way, we can leverage passage-level information to capture more acceptable user interests. As the results shown in Table 6, the model without query-aware attention still outperforms previous methods. This implies the document pruning module effectively extracts useful features, while simply using the titles or the entire text of the documents indeed brings much noise. Naturally, we wonder if the performance of other personalized methods will be improved with the utilization of the query-aware document pruning module. Hence, in this section, we

Table 10. Results of Baselines with Document Pruning

Model	MAP	MRR	P@1	P-improve
HTPS w/o. DP	.8207	.8304	.7219	.2773
HTPS DP	.8226	.8331	.7268	.2759
PSSL w/o. DP	.8299	.8295	.7329	.2709
PSSL DP	.8301	.8398	.7338	.2740

Table 11. Results of Baselines with Candidate Document-aware Interest Modeling

Model	MAP	MRR	P@1	P-improve
HTPS	.8224	.8324	.7286	.2552
HTPS ca	.8236	.8340	.7288	.2758
PSSL	.8301	.8398	.7338	.2688
PSSL ca	.8280	.8379	.7315	.2715

apply the document pruning for two baselines, HTPS and PEPS. We choose the two models for experiments because they characteristically follow the popular user profiling paradigm shown in Figure 1(a): building a user profile from the sequence of past queries and clicked documents.

To be specific, we prune the clicked documents according to their corresponding historical queries. The experiments are noted as "DP" in the table. We do not prune the document according to the current interest, because the current intent-oriented features need to be specially modeled, which is not included in the baseline models. The pruned passages are encoded to vectors as in Section 3.1.1. Then, the vectors are appended to the original title word embeddings of the documents. That is, they are treated as additional words from the documents in the baseline models.

For comparison, we also experiment with the baselines with the first passages of the documents. The experiments are noted as "w/o. DP" in the table.

The number of passages is set as 3. The length of history is 20. Note that the history lengths is not the same as the ones in the original papers. Because PSSL does not measure the length by the number of clicked behaviors, it just sends a certain number of queries and all the clicked documents. While, the HTPS uses 50 past behaviors as user history, which is quite time-consuming for our document-pruning strategy. The models in this section are implemented according to the code published by their authors.

The experiments shown in 10 illustrate that both the "DP" models outperforms "w/o. DP"models. Specifically, for HTPS the document pruning improves the accuracy by 0.19% on MAP, while for PSSL, it improves 0.02% on MAP. This verifies that our model-free document pruning algorithm could help the models to build a more accurate user profile by the query-aware pruned passages.

6.10 Applications of the Candidate Document-aware Interest Modeling

In our DIMPS, we send the candidate document into the interest modeling procedure to capture the interactions between the candidate document and user behaviors. The inferiority of "independent HC" in Table 7 and "DIMPS w/o. c" in Table 4 compared to DIMPS indicates the effectiveness of such candidate document-aware interest modeling strategy. Similarly, we would like to investigate the functionality of this strategy on other personalized methods. Like in Section 6.9, we choose the two typical user profiling baselines HTPS and PSSL for experiments.

In implementation, we append the candidate document representation, used in their original models, to the last of the history sequence. The rest of the interest modeling procedure remains unchanged. The candidate document-aware models are denoted as "ca". Comparison of the tested models and original models is shown in Table 11. The models in this section are implemented according to the code published by their authors.

It is observed that the candidate document-aware strategy improves the performance of HTPS, but deteriorates the performance of PSSL. The deterioration may come from the neglect of the relationship between history and documentation in PSSL's pre-training stage. Besides, from the "compared " model in Section 6.3 we can see the ways of getting ranking scores influence the performance of candidate document-aware modeling. To conclude, we believe that incorporating candidate documents by appending them to the history sequence could lead to improvements for some methods, but it needs more special consideration for some sophisticated scenarios. The dependency between history and candidate documents should be further studied to design a more functional candidate document-aware interest modeling strategy. We would leave this to our future work.

7 CONCLUSION

In this paper, we address the personalized search problem by building intent-aware dynamic document representation to construct more accurate user profiles. To implement this idea, we propose a dynamic interest modeling strategy from passage-level evidence. First, we explicitly model features related to corresponding and current intents from documents. Furthermore, we organize documents and their corresponding queries into a sequence and build a dynamic user profile. Finally, we evaluate the learned profile to re-rank the candidate documents.

Experiments on the large-scale dataset illustrate our model's superiority over the state-of-the-art methods. Besides, we perform ablation experiments on the two main components in our dynamic intent-oriented document encoder. Results demonstrate our model's ability to capture search intents from history. The analysis of our passage encoding and sequence modeling is also presented with experimental results. It is confirmed that using queries to guide the extraction of corresponding intents from passage-level evidence is an effective way for document representation, while capturing the full interactions among history, current query, and candidate documents is necessary for better re-ranking. Then we test our model's performance on ambiguous and non-ambiguous queries respectively, to test its capacity to build intents even with inaccurate search words. Furthermore, by presenting the visualization of weights, we also verify our model's effectiveness in modeling dynamic intent-aware documents as well as the whole search sequences to depict a better user profile. We also conduct efficiency-effectiveness analysis between the state-of-the-art model and our model with varying history lengths as well as one another passage encoding strategy. Results have proved the possibility of deploying DIMPS into real-world applications. Besides, we test the functionality of the document pruning module on several baselines. The improvements verify that our document pruning can offer efficient document vectors for other personalized approaches, alleviating the noise problem caused by using the entire text of the documents.

In future work, we are interested in designing an exclusive intent-orient interest modeling strategy for long-term history to improve the performance of our model. Second, we want to explore how the ranking quality changes when substituting some of the components in the offline document pruning and passage encoding stage, which would be beneficial in finding the tradeoff configurations under different search requirements.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author. This work was supported by the National Natural Science Foundation of China No. 62272467 and No. 61872370, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China NO. 22XNKJ34, Public Computing Cloud, Renmin University of China, and the fund for building world-class universities (disciplines) of Renmin University of China. The work was partially done at 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 International Conference on Web Search and Data Mining (WSDM) (international conference on web search and data mining (wsdm) ed.). Association for Computing Machinery, Inc. https://www.microsoft.com/en-us/research/publication/diversifying-search-results/

[2] 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. 135–144.

[3] Paul N. Bennett, Krysta Svore, and Susan T. Dumais. 2010. Classification-Enhanced Ranking. 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, 111–120. https://doi.org/10.1145/1772690.1772703

[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. 185–194.

[5] Keping Bi, Qingyao Ai, and W. Bruce Croft. 2020. Learning a Fine-Grained Review-based Transformer Model for Personalized Product Search. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2020). https://api.semanticscholar.org/CorpusID:234351913

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. the Journal of machine Learning research 3 (2003), 993–1022.

[7] James P Callan. 1994. Passage-level evidence in document retrieval. In SIGIR'94. Springer, 302–310.

[8] 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. 1849–1852.

[9] Wanyu Chen, Pengjie Ren, Fei Cai, Fei Sun, and M. de Rijke. 2019. Improving End-to-End Sequential Recommendations with Intent-aware Diversification. Proceedings of the 29th ACM International Conference on Information & Knowledge Management (2019).

[10] Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. 2022. Intent Contrastive Learning for Sequential Recommendation. In Proceedings of the ACM Web Conference 2022 (Virtual Event, Lyon, France) (WWW '22). Association for Computing Machinery, New York, NY, USA, 2172–2182. https://doi.org/10.1145/3485447.3512090

[11] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. arXiv preprint arXiv:2003.07820 (2020).

[12] Steve Cronen-Townsend, W Bruce Croft, et al. 2002. Quantifying query ambiguity. In Proceedings of HLT, Vol. 2. Citeseer, 94–98.

[13] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (2019).

[14] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In Proceedings of the eleventh ACM international conference on web search and data mining. 126–134.

[15] Chenlong Deng, Yujia Zhou, and Zhicheng Dou. 2022. Improving Personalized Search with Dual-Feedback Network. In WSDM. ACM, 210–218.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[17] 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. 581–590.

[18] Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. 2019. Explainable recommendation through attentive multi-view learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 3622–3629.

[19] 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. 347–356.

[20] Morgan Harvey, Fabio Crestani, and Mark J Carman. 2013. Building user profiles from topic models for personalised search. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2309–2314.

[21] Sebastian Hofstätter, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. 2021. Intra-Document Cascading: Learning to Select Passages for Neural Document Ranking. arXiv preprint arXiv:2105.09816 (2021).

[22] Yiqiao Jin, Yeon-Chang Lee, Kartik Sharma, Meng Ye, Karan Sikka, Ajay Divakaran, and Srijan Kumar. 2023. Predicting Information Pathways Across Online Communities. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Long Beach, CA, USA) (KDD '23). Association for Computing Machinery, New York, NY, USA, 1044–1056. https://doi.org/10.1145/3580305.3599470

[23] 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
[24] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In 2018 IEEE International Conference on Data Mining (ICDM). 197–206. https://doi.org/10.1109/ICDM.2018.00035
[25] Marcin Kaszkiel and Justin Zobel. 1997. Passage retrieval revisited. In ACM SIGIR Forum, Vol. 31. ACM New York, NY, USA, 178–185.
[26] Mesut Kaya and Derek G. Bridge. 2019. Subprofile-aware diversification of recommendations. User Modeling and User-Adapted Interaction (2019), 1–40.
[27] Furkan Kocayusufoglu, Tao Wu, Anima Singh, Georgios Roumpos, Heng-Tze Cheng, Sagar Jain, Ed Chi, and Ambuj Singh. 2022. Multi-Resolution Attention for Personalized Item Search. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (Virtual Event, AZ, USA) (WSDM '22). Association for Computing Machinery, New York, NY, USA, 508–516. https://doi.org/10.1145/3488560.3498426
[28] Robert Krovetz and W Bruce Croft. 1992. Lexical ambiguity and information retrieval. ACM Transactions on Information Systems (TOIS) 10, 2 (1992), 115–141.
[29] Xiujun Li, Chenlei Guo, Wei Chu, Ye-Yi Wang, and Jude Shavlik. 2014. Deep learning powered in-session contextual ranking using clickthrough data. In In Proc. of NIPS.
[30] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. Synthesis Lectures on Human Language Technologies 14, 4 (2021), 1–325.
[31] 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. 555–564.
[32] Zhengyi Ma, Zhicheng Dou, Guanyue Bian, and Ji-Rong Wen. 2020. PSTIE: Time Information Enhanced Personalized Search. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (Virtual Event, Ireland) (CIKM '20). Association for Computing Machinery, New York, NY, USA, 1075–1084. https://doi.org/10.1145/3340531.3411877
[33] Tomas Mikolov, Quoc Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. arXiv preprint arXiv:1309.4168 (09 2013).
[34] Zhiqiang Pan, Fei Cai, Yanxiang Ling, and M. de Rijke. 2020. An Intent-guided Collaborative Machine for Session-based Recommendation. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020).
[35] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. ACM. https://doi.org/10.1145/1146847.1146848
[36] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. arXiv preprint arXiv:1904.07531 (2019).
[37] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084
[38] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael M. Bronstein. 2020. Temporal Graph Networks for Deep Learning on Dynamic Graphs. CoRR abs/2006.10637 (2020). arXiv:2006.10637 https://arxiv.org/abs/2006.10637
[39] Rodrygo Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. Proceedings of the 19th International Conference on World Wide Web, WWW '10, 881–890. https://doi.org/10.1145/1772690.1772780
[40] Rodrygo Santos, Craig Macdonald, and Iadh Ounis. 2011. On the role of novelty for search result diversification. Information Retrieval 15 (10 2011). https://doi.org/10.1007/s10791-011-9180-x
[41] Eilon Sheetrit, Anna Shtok, and Oren Kurland. 2020. A passage-based approach to learning to rank documents. Information Retrieval Journal 23, 2 (2020), 159–186.
[42] Qijie Shen, Hong Wen, Jing Zhang, and Qi Rao. 2022. Hierarchically Fusing Long and Short-Term User Interests for Click-Through Rate Prediction in Product Search. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (Atlanta, GA, USA) (CIKM '22). Association for Computing Machinery, New York, NY, USA, 1767–1776. https://doi.org/10.1145/3511808.3557351
[43] 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. 525–534.
[44] Jaime Teevan, Susan Dumais, and Dan Liebling. 2008. To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent. In Proceedings of SIGIR 2008 (proceedings of sigir 2008 ed.). Association for Computing Machinery, Inc. https://www.microsoft.com/en-us/research/publication/to-personalize-or-not-to-personalize-modeling-queries-with-variation-in-user-intent/
[45] Saul Vargas, Pablo Castells, and David Vallet. 2011. Intent-Oriented Diversity in Recommender Systems. 1211–1212. https://doi.org/10.1145/2009916.2010124
[46] Saul Vargas, Pablo Castells, and David Vallet. 2011. Intent-oriented diversity in recommender systems. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 1211–1212.
[47] 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 Advances in neural information processing systems. 5998–6008.

[48] Maksims Volkovs. 2015. Context models for web search personalization. arXiv preprint arXiv:1502.00527 (2015).
[49] Thanh Vu, Alistair Willis, Son N Tran, and Dawei Song. 2015. Temporal latent topic user profiles for search personalisation. In European Conference on Information Retrieval. Springer, 605–616.
[50] Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020. Fine-grained interest matching for neural news recommendation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 836–845.
[51] Jun Wang and Jianhan Zhu. 2009. Portfolio Theory of Information Retrieval. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (Boston, MA, USA) (SIGIR '09). Association for Computing Machinery, New York, NY, USA, 115–122. https://doi.org/10.1145/1571941.1571963
[52] Shuting Wang, Zhicheng Dou, Jing Yao, Yujia Zhou, and Ji-Rong Wen. 2023. Incorporating Explicit Subtopics in Personalized Search. In WWW. ACM, 3364–3374.
[53] Shoujin Wang, Liang Hu, Yan Wang, Quan Z. Sheng, Mehmet Orgun, and Longbing Cao. 2019. Modeling Multi-Purpose Sessions for Next-Item Recommendations via Mixture-Channel Purpose Routing Networks. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, 3771–3777. https://doi.org/10.24963/ijcai.2019/523
[54] Jacek Wasilewski and Neil Hurley. 2018. Intent-aware item-based collaborative filtering for personalised diversification. In Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization. 81–89.
[55] 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. 1411–1420.
[56] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Npa: Neural news recommendation with personalized attention. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2576–2584.
[57] Qiang Wu, Chris JC Burges, Krysta M Svore, and Jianfeng Gao. 2008. Ranking, boosting, and model adaptation. Technical Report. Technical report, Microsoft Research.
[58] 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. 55–64.
[59] 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. 1359–1368.
[60] Jing Yao, Zhicheng Dou, Jun Xu, and Ji-Rong Wen. 2020. RLPer: A Reinforcement Learning Model for Personalized Search. In Proceedings of The Web Conference 2020. 2298–2308.
[61] Cheng Zhai, William Cohen, and John Lafferty. 2003. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. SIGIR Forum (ACM Special Interest Group on Information Retrieval) (04 2003). https://doi.org/10.1145/860435.860440
[62] Yujia Zhou, Zhicheng Dou, Bingzheng Wei, Ruobing Xie, and Ji-Rong Wen. 2021. Group Based Personalized Search by Integrating Search Behaviour and Friend Network. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (Virtual Event, Canada) (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 92–101. https://doi.org/10.1145/3404835.3462918
[63] Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2020. Enhancing Re-finding Behavior with External Memories for Personalized Search. In WSDM. ACM, 789–797.
[64] Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2023. Enhancing Potential Re-Finding in Personalized Search With Hierarchical Memory Networks. IEEE Trans. Knowl. Data Eng. 35, 4 (2023), 3846–3857.
[65] 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. 1111–1120.
[66] Yujia Zhou, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2021. PSSL: Self-supervised Learning for Personalized Search with Contrastive Sampling. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2749–2758.