Multi-grained Document Modeling for Search Result Diversification

ZHIRUI DENG, Gaoling School of Artificial Intelligence, Renmin University of China, Haidian District, China
ZHICHENG DOU, Gaoling School of Artificial Intelligence, Renmin University of China, Haidian District, China
ZHAN SU, Gaoling School of Artificial Intelligence, Renmin University of China, Haidian District, China
JI-RONG WEN, Gaoling School of Artificial Intelligence, Renmin University of China, Haidian District, China

Search result diversification plays a crucial role in improving users' search experience by providing users with documents covering more subtopics. Previous studies have made great progress in leveraging inter-document interactions to measure the similarity among documents. However, different parts of the document may embody different subtopics and existing models ignore the subtle similarities and differences of content within each document. In this article, we propose a hierarchical attention framework to combine intra-document interactions with inter-document interactions in a complementary manner in order to conduct multi-grained document modeling. Specifically, we separate the document into passages to model the document content from multi-grained perspectives. Then, we design stacked interaction blocks to conduct inter-document and intra-document interactions. Moreover, to measure the subtopic coverage of each document more accurately, we propose a passage-aware document-subtopic interaction to perform fine-grained document-subtopic interaction. Experimental results demonstrate that our model achieves state-of-the-art performance compared with existing methods.

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

Additional Key Words and Phrases: Intra-document relations, search result diversification, multi-grained document modeling

ACM Reference Format:
Zhirui Deng, Zhicheng Dou, Zhan Su, and Ji-Rong Wen. 2024. Multi-grained Document Modeling for Search Result Diversification. *ACM Trans. Inf. Syst.* 42, 5, Article 126 (April 2024), 22 pages. https://doi.org/10.1145/3652852

This work was supported by the National Natural Science Foundation of China No. 62272467, the fund for building world-class universities (disciplines) of Renmin University of China, and Public Computing Cloud, Renmin University of China.
Authors' address: Z. Deng, Z. Dou (Corresponding author), Z. Su, and J.-R. Wen, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing, 100872, China; e-mails: zrdeng@ruc.edu.cn, dou@ruc.edu.cn, suzhan@ruc.edu.cn, jrwen@ruc.edu.cn.

ACM 1046-8188/2024/04-ART126
https://doi.org/10.1145/3652852

1 INTRODUCTION

Traditional search engines only consider the relevance of documents and neglect the diversity. It is often the case that a query is short and ambiguous [13, 22, 30, 31] and different users issuing the same query are finding different information. For example, when a user's query is "map", it is uncertain whether she wants to seek information about "google map" or "map container in programming languages". Ranking documents solely based on their relevance to the query may favor the documents about "google map", whereas the user who is looking for "map container" may fail to find relevant information from the top results. Search result diversification is proposed to solve this problem. It aims at returning diverse documents covering different user intents within the top results.

Existing search result diversification approaches either model the (implicit) relationship between documents [3, 32, 35, 36, 38, 41, 44] or the (explicit) document's coverage of manually constructed subtopic (e.g., Google suggestions) [9, 21, 23, 29] to improve the diversity of search results. These methods regard a document as a unit to measure the document-document similarity or subtopic-document coverage. However, different parts of a document may cover different subtopics. For example, a document that introduces "Starbucks", may state "Starbucks menu" in one part and "Starbucks location" in another part. Treating a document as a unit may lead to the information in different parts of the document not being fully captured and modeled which will further restrict their capacity. Therefore, *how to model fine-grained subtopic differences within documents while balancing the efficiency and effectiveness* becomes a crucial problem for search result diversification.

Currently proposed search result diversification models only leverage a single global representation which represents a document as a single vector to perform inter-document interaction while ignoring term and passage level fine-grained interaction to measure the novelty of documents. In this instance, the relationship between documents cannot be well learned. Besides, measuring the coverage of subtopics only with the document and subtopics' single global representation will neglect the fine-grained interaction between the document and subtopics, resulting in inaccurate subtopic coverage estimates. Moreover, there might be duplicate content and noisy information in a document. For instance, for a document about "Alan Turing", there may be more than one part involving "Turing's achievements". It is crucial to discard duplicate content and extract the document content that can reflect its main content in order to reduce computational consumption. However, previous models adopted the entire document content to build a document representation which reduced efficiency and misled the diversified ranking model. Thus, it is prominent to capture the local semantic similarities and differences between different parts of the document. This motivates us to leverage multi-grained information contained in the documents to represent documents instead of solely using a single global document representation to model document content relationships.

Based on the above considerations, we propose a **H**ierarchical **A**ttention framework for search result **D**iversification (HAD) which **models intra-document relations and inter-document relations from multi-grained perspectives**. Specifically, our proposed model HAD is composed of four steps. **First**, we separate the document into passages to represent the document content from multiple granularities. To conduct efficient inter-document interactions, we select pivot entities from documents and leverage them to perform interactions between different candidate documents. Through a term-level encoder, we obtain the initial representation of passages, documents, and subtopics. **Second**, we design L layer stacking interaction blocks. Each block consists of a document encoder and a sequence encoder. The former captures the relationships between passages within a document whereas the latter enhances the interaction between different candidate documents. Moreover, we generate document representation by merging signals from different

perspectives with multi-head attention to strengthen information propagation. **Besides**, we conduct a passage-aware document-subtopic interaction to enhance the fine-grained interaction between documents and subtopics and measure the subtopic coverage of each document more accurately. **Finally**, we combine the above passage-aware diversity features with the relevance features to obtain the document's final ranking score through an MLP.

We conduct experiments on the Web Track dataset from TREC 2009 to 2012. Experimental results show that our model outperforms existing approaches in terms of all evaluation metrics. Extensive experiments are carried out to study the role of different components in our model. Besides, we evaluate the efficiency and effectiveness of our model and validate that our model can achieve better experimental results on all evaluation metrics while spending less time. We also conduct quantitative analyses and a case study to analyze why our model can achieve better ranking results.

In summary, our major contributions are three-fold:

(1) We present a hierarchical attention framework for search result diversification to capture intra-document and inter-document interaction from multi-grained perspectives.
(2) We introduce passage-aware document-subtopic interaction to measure the subtopic coverage of the document's different parts.
(3) We devise a pivot entity selection component to reduce computational consumption and distill the document content.

The rest of our article is organized as follows. We introduce related work about search result diversification and attention and hierarchical mechanism in IR in Section 2. The details of the proposed hierarchical attention framework for search result diversification HAD are illustrated in Section 3. We also state the training and optimization in Section 3. Section 4 describes the datasets, evaluation metrics, baselines, and implementation details of our model. We report the overall experimental results and several extended experimental results in Section 5. Section 6 is the conclusion of our work.

2 RELATED WORK

2.1 Search Result Diversification

Search result diversification aims at tackling the problem of ambiguous queries and to satisfy users' different information needs. Existing search result diversification algorithms can be roughly classified into explicit approaches, implicit approaches and ensemble approaches in terms of whether the subtopics are explicitly modeled and whether the similarity of documents is captured. In this section, we will briefly review the three types of approaches and discuss their relationship with our model.

2.1.1 Implicit Diversification Approaches. The implicit methods [37] consider the relevance between a candidate document and the query and the divergence of the candidate document to the selected documents. MMR [3] was the foundation of most implicit approaches. It leverages the linear combination of the traditional query-document relevance and the novelty of the document to the selected documents to produce a score of document d_i and generates diversified document ranking. The score of document d_i is denoted as Equation (1).

$$S_i = \lambda S^{rel}(d_i, q) + (1 - \lambda) \max_{d_j \in D_s} S^{div}(d_i, d_j), \tag{1}$$

where S_i is the score of document d_i, $S^{rel}(d_i, q)$ indicates the relevance of document d_i and query q, $S^{div}(d_i, d_j)$ reflects the similarity between document d_i and document d_j, D_s is the set of selected documents and λ is a hyperparameter which balances the relevance score and the diversity score.

MMR is an unsupervised method, and based on it, a lot of supervised methods have been proposed. SVM-DIV [41] leveraged structural SVM [8] to predict diverse subsets and measure the diversity of the documents. R-LTR [44] used a process of sequential document selection and defined a ranking function and loss function to generate the diversified ranking. PAMM [35] adopted the ranking model that maximized marginal relevance at ranking and directly optimized the diversity evaluation measures. NTN [36] automatically learned a non-linear document novelty function from the data with a neural tensor network and could be combined with R-LTR and PAMM. DALETOR [38] and MO4SRD [40] both designed a direct metric optimization framework for search result diversification. To capture the intent coverage of the document, Graph4DIV [32] leveraged graphs to model the information interactions between documents.

2.1.2 Explicit Diversification Approaches. Compared with implicit approaches which solely leverage the similarity between documents to measure the novelty of documents, explicit approaches [39] additionally utilized the coverage of subtopics to model the novelty of documents. xQuAD [29] leveraged sub-queries to explicitly account for different aspects of the query and diversified a document ranking by estimating the degree of a document satisfying each uncovered aspect. PM2 [9] treated the problem as finding a proportional representation for the document ranking. HxQuAD/HPM2 [21] represented subtopics in a hierarchical structure and TxQuAD/TPM2 [10] leveraged subtopic terms to promote diversity. To avoid heuristic and handcrafted functions and parameters, researchers have proposed supervised methods. Jiang et al. [23] proposed DSSA which leveraged attention mechanism and **recurrent neural networks (RNNs)** to model subtopics in a supervised learning framework.

2.1.3 Ensemble Diversification Approaches. Implicit approaches leveraged the similarity of documents to measure the novelty of documents while explicit approaches used the coverage of subtopics to determine the diversity of documents. Inheriting the advantages of implicit and explicit approaches, ensemble approaches utilized both features to improve the model's performance. DVGAN [25] introduced a **generative adversarial network (GAN)** to combine explicit document-subtopic relevance features with implicit document-document similarity features and generate training samples effectively. DESA [26] utilized a self-attention network to model all candidate documents as a whole sequence and score all documents simultaneously. GDESA [27] leveraged a self-attention encoder-decoder structure and an RNN-based document selection component to combine the global interactions among all the documents and the interactions between the selected sequence and each unselected document.

Although these methods diversify search results based on the similarity of documents and the coverage of subtopics, they have a common problem: they take the document as the basic unit of search result diversification and measure the novelty of documents solely based on inter-document interaction. Compared with these methods, our method not only uses the inter-document interaction to capture the similarities and differences between different documents' content but also measures intra-document relations by using fine-grained interaction within documents from multi-grained perspectives.

2.2 Attention and Hierarchical Mechanism in IR

Since attention [1] mechanism was proposed, plenty of models such as Transformer [33] and BERT [11] have adopted attention networks to replace RNNs [5, 19] and **convolution neural networks (CNNs)** [17]. In particular, the transformer has shown effectiveness in not only the information retrieval [42] but also the computer vision [12]. Because the attention mechanism can learn long-range dependencies, and obtain documents' attention distributions, it is suitable for search result diversification which needs to capture the similarity between documents and the subtopic

coverage of documents. Previous models have adopted the attention mechanism into search result diversification. DESA [26] adopted an attention mechanism substitute for RNNs to make full interaction between documents and approach the global optimal diversity ranking. DALETOR [38] built a neural network with multi-head self-attention to verify the effectiveness of their loss function. Inheriting the spirit of these models, we leveraged the attention mechanism to depict the intra-document and inter-document relations and built an ensemble model to learn a better diversified document ranking.

Besides, the hierarchical structure has been proven to be effective in many fields, such as personalized search [16, 43] and session search [28]. Hierarchical structure can capture semantic information at different levels and conduct more fine-grained semantic interaction. Therefore, it is reasonable to introduce the hierarchical structure into search result diversification. Existing studies [21] argued that subtopics should be organized in a hierarchical structure instead of a flat list in order to better represent user intents. However, they ignored the relationship of content within each document where the author covered multiple topics in different document parts.

In this work, we leverage the attention mechanism to describe the content of documents more precisely. Besides, we inherit the advantage of hierarchical structure to model the candidate documents from multi-grained perspectives and obtain all documents' ranking scores simultaneously.

3 METHODOLOGY

Search result diversification models aim at returning diversified search results so that the top search results can cover users' intent as much as possible. However, most existing models only consider inter-document relations: each document is represented by a single vector, and diversification is achieved by measuring the similarity between document vectors. In this article, to better estimate similarities between documents more effectively and accurately, we propose to further model intra-document relations. We design a hierarchical attention framework to obtain multi-grained representations of a document by incorporating representations of its terms and passages. With the multi-grained document representation, we can estimate the novelty of a document and the similarity between documents more effectively. Moreover, we conduct fine-grained interaction between documents and subtopics to measure the subtopic coverage more accurately.

3.1 Overview of Our Model HAD

The notations and their descriptions in this article are shown in Table 1. Formally, given a query $q \in \mathcal{Q}$, we leverage both the subtopic t_k in the subtopic set $\mathcal{T}$ and documents in the candidate document set $\mathcal{D}$ to re-rank the documents and obtain a final diversity ranking document list $\mathcal{R}$.

As the search result diversification problem is NP-hard, the greedy selection approach has been used in most existing models [23, 29, 32] which iteratively select the next document from the candidate document set based on its relevance to the query and the novelty to the selected document set. Recently, researchers propose to model all candidate documents as a whole sequence and measure their information utilities globally [26, 38]. Following these works, in this article, we propose an ensemble model and score all candidate documents simultaneously.

The overall structure of our model HAD is shown in Figure 1. With the above concepts and notations, we briefly introduce our model HAD as follows.

(1) Text Representations (Section 3.2).

For document d_i, we first employ a sliding window to divide the document into multiple passages in order to take advantage of local semantic interaction within the document. Besides, previous studies retain all content in the document, which will hurt efficiency and introduce redundant information. To reduce the duplication of content, we select pivot entities that can reflect the main content of the document and concatenate them as a new document $\hat{d}_i$.

Table 1. Notations

Notation	Description
$\mathcal{Q}$, q	$\mathcal{Q}$ is the query set and the query $q \in \mathcal{Q}$.
$\mathcal{D}$, d_i	$\mathcal{D}$ is the candidate document set and $d_i \in \mathcal{D}$.
$\mathbf{R}_q$	the relevance features of query q
$\mathbf{R}_i$	the subtopics' relevance features for d_i
S_i	diversity score of document d_i
$\mathcal{R}$	the final diversity ranking document list
$\mathbf{D}_i^o$	the initial embedding of document d_i.
$\overline{\mathbf{D}}_i$	the context-aware representation of d_i after inter-document attention.
$\mathbf{D}_i^d$	the global representation of d_i.
$\mathbf{D}_i^p$	the passage-aware representation of d_i.
$\mathbf{D}_i^t$	the subtopic coverage feature of document d_i.
$\mathcal{T}$, t_k	the subtopic set and $t_k \in \mathcal{T}$.
$\mathbf{T}_k^o$	the initial embedding of subtopic t_k.
$\overline{\mathbf{T}}_m$	the subtopic coverage representation after multi-head attention.
p_{ij}	p_{ij} is the jth passage in document d_i.
$\mathbf{P}_{ij}^o$	the initial embedding of passage p_{ij}.
$\overline{\mathbf{P}}_{ij}$	the context-aware representation of p_{ij} after intra-document attention.
$\mathbf{P}_{ij}$	the passage-aware representation of p_{ij} after multi-head attention in document encoder.

To boost term interaction within document content and integrate more semantic knowledge into the initial embedding of documents, for document d_i, we fed passage p_{ij} which is the jth passage in document d_i and the new document $\hat{d}_i$ into a term-level encoder to obtain a better generalized initial embedding $\mathbf{P}_{ij}^o$ and $\mathbf{D}_i^o$, respectively. Similarly, we also adopt a term-level encoder to obtain a subtopic initial embedding $\mathbf{T}_k^o$ for subtopic t_k.

(2) Stacking Interaction Blocks (Section 3.3).

We hire L layer stacking interaction blocks to enhance intra-document and inter-document interactions, integrate multi-grained document representation to promote information propagation, and acquire passage and document representations. Each layer of stacking interaction blocks consists of a context-aware document encoder and a context-aware sequence encoder.

Context-aware Document Encoder. To model intra-document relations, we first design a context-aware document encoder, as shown in the green part of Figure 1. The document encoder is comprised of an intra-document attention and a multi-head attention. The former interacts all passages in each document and generates a context-aware passage representation $\overline{\mathbf{P}}_{ij}$ for passage p_{ij}. The latter combines the initial passage representation $\mathbf{P}_{ij}^o$ with the context-aware passage representation $\overline{\mathbf{P}}_{ij}$ to enhance information propagation between different perspectives' passage representation and generate a passage-aware representation $\mathbf{P}_{ij}$. For the last layer of stacking interaction blocks, we increase a pooling operation to obtain the passage-aware document representation $\mathbf{D}_i^p$.

Context-aware Sequence Encoder. As shown in the blue part of Figure 1, we first interact all candidate documents and obtain a context-aware document representation $\overline{\mathbf{D}}_i$ for document d_i. Through a multi-head attention, we combine the document's passage-aware representation $\mathbf{P}_{i1}$:$\mathbf{P}_{im}$ with $\overline{\mathbf{D}}_i$

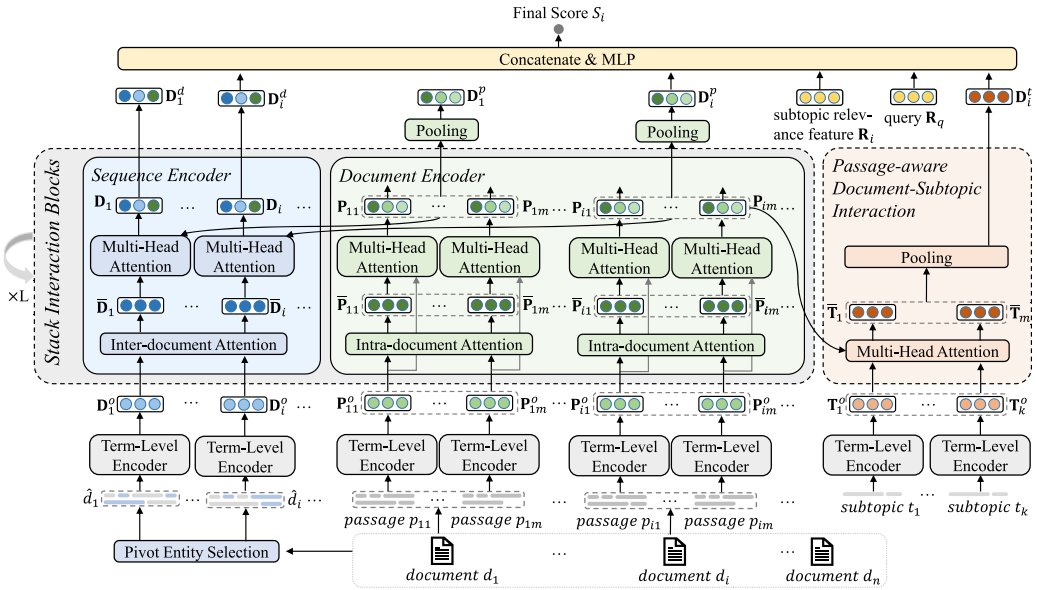


Fig. 1. The architecture of our model HAD. For a candidate document list, we encode passages and document entities through the term-level encoder. We design stack interaction blocks to perform inter-document and intra-document attention and multi-grained interaction. Then, we conduct passage-aware document subtopic interaction to perform fine-grained interaction between documents and subtopics. Finally, we concatenate the multi-grained document's diversity representation and the document's relevance features and produce the final score of the document.

to strengthen the local and global information interaction and generate a global document representation $\mathbf{D}_i^d$.

(3) Passage-aware Document-Subtopic Interaction. (Section 3.4).

In order to capture the fine-grained subtopic-coverage information of documents, we interact the initial subtopic embedding $\mathbf{T}_1^o$:$\mathbf{T}_k^o$ with the passage-aware representation $\mathbf{P}_{i1}$:$\mathbf{P}_{im}$ of document d_i and obtain each passage's subtopic coverage representation $\overline{\mathbf{T}}_1$:$\overline{\mathbf{T}}_m$. With a pooling operation, the model produces the subtopic coverage representation $\mathbf{D}_i^t$. This is shown in orange in Figure 1.

(4) The Final Ranking. (Section 3.5).

From stacking interaction blocks, we have obtained the passage-aware document representation $\mathbf{D}_i^p$ and the global document representation $\mathbf{D}_i^d$. Through passage-aware document-subtopic interaction, we have acquired the subtopic coverage representation $\mathbf{D}_i^t$ for document d_i. We concatenate these diversity features with the relevance features and generate the final document score with an MLP. We rank the documents with these scores and get the final document ranking.

In the remaining part of this section, we will introduce the details of these components and how we leverage these components to diversify document ranking.

3.2 Text Representations

In this section, we introduce the process of altering the representation of documents and subtopics to better represent their content. We carry out passage generation which partitions a document into passages, pivot entity selection which selects important entities in the document, and term-level encoder which obtains initial embedding of passages, documents, and subtopics. We will introduce the details of these components in the following parts.

3.2.1 Passage Generation. Different parts in the documents may imply different subtopics and representing a document as a single vector loses local semantic relations in each document. For example, in a document about "Starbucks menu", the first paragraph may be about "drinks on Starbucks menu" while the second paragraph may introduce "the calories of these drinks". Therefore, it is necessary to split the document into passages and model intra-document relations in order to capture subtle differences in the document content.

Following previous works [15, 20] , we first employ a sliding window of size w with an overlapping factor o, where $o < w$, to divide the document into passages, as shown in Equations (2) to (3). There are w tokens in each passage. The first passage starts from the first position of the document. For each document, we keep m passages and too short documents will be padded.

$$p_{i1} = d_{1:w}, \tag{2}$$

$$p_{ij} = d_{(w-o)*(j-1)+1:(w-o)*j+o}, \tag{3}$$

where p_{ij} is the jth passage in document d_i and $d_{1:w}$ is the token sequence from the 1th to the wth in document d_i, and we omit i for convenience.

3.2.2 Initial Passage Representation. Enhancing interaction between terms plays a crucial role in obtaining passages' semantic information and co-occurrence information at the term level. Besides, integrating more semantic knowledge into the initial embedding of passages can strengthen the representation ability of the passage representation and promote downstream diversified ranking results. Thus, we design a term-level encoder to obtain a better passage initial embedding. Endowed with the benefit of pretrained language model, we adopt BERT as the term-level encoder because it is pretrained on the large corpus and has achieved great performance on plenty of natural language processing tasks. Other advanced pretrained models are also practicable to obtain the passage initial embedding.

Specifically, given passage p_{ij}, we use BERT to encode the passage and get the representation of [CLS] token as the passage representation $\mathbf{P}^o_{ij}$, i.e.,

$$\mathbf{P}^o_{ij} = \text{BERT}_{[\text{CLS}]}(p_{ij}). \tag{4}$$

3.2.3 Initial Document Representation. In the above section, we have separated the document into several passages for the sake of modeling intra-document semantic relations. Moreover, the inter-document relations are also important to measure the similarities and differences between different documents. Under this circumstance, there are two problems to model the inter-document relations simply based on passages: in terms of effectiveness, there is plenty of useless and redundant information in the document, and utilizing this information will mislead the diversified ranking model. From the perspective of efficiency, the computational consumption will be high if we model all documents based on passages, because we need to model all passages in all documents simultaneously. Moreover, processing long documents through a pretrained model like BERT is impossible, because the input of BERT has a maximum length.

To cope with these problems, we select pivot entities in each document and leverage them to represent the document. The pivot entities are informative entities that play a crucial role in document semantics. We adopt an open-source entity linking tool TagMe [14] to annotate entities in each document because it can achieve high accuracy and reliability in identifying and linking entities in various types of texts. As shown in the following example, the tool can map conceptual entities within text to corresponding Wikipedia concepts.

Example text: Find more *Starbucks products* in our *Tmall store. Learn* more about *Starbucks coffee culture.*

Extracted entities: ("Starbucks": 0.7795), ("products": 0.0305), ("Tmall": 1.0), ("store": 0.0019), ("Learn": 0.0012), ("Starbucks coffee": 0.3593), ("coffee culture": 0.4030)

These entities could be specific people, locations, or any other concept covered in the Wikipedia knowledge base. The probability score provided by TagMe reflects the confidence level in the association between a detected entity in the text and its corresponding Wikipedia concept. It should be noted that in Section 3.2.2, we do not extract entities from passages, because the passage is short enough to be modeled by pretrained model and extracting entities will destroy sentence structure in passages which will adversely affect the final result. After extracting entities, we keep some of the most important entities for document d_i, and concatenate these entities as a new document $\hat{d}_i$. Similar to the passage representation, we use BERT to encode the document, i.e.,

$$\mathbf{D}_i^o = \text{BERT}_{[\text{CLS}]}(\hat{d}_i). \tag{5}$$

3.2.4 Initial Subtopic Representation. We use the same method to encode subtopics. For the subtopic t_k, we have:

$$\mathbf{T}_k^o = \text{BERT}_{[\text{CLS}]}(t_k).$$

3.3 Stacking Interaction Blocks

Previous works solely consider the inter-document interaction to measure the similarity of documents but neglect the intra-document passage dependencies where different parts of the document may contain different subtopics. To capture intra-document relations and obtain a document representation reflecting document fine-grained semantic distinction, we propose stacking interaction blocks to generate multi-grained document representation. Its architecture is shown in Figure 1. The stacking interaction blocks are comprised of a stack of L identical layers and the basic components of each layer are a document encoder and a sequence encoder. In the following of this section, we will introduce the details of each component.

3.3.1 Context-Aware Document Encoder. From Section 3.2, we have acquired the passage initial embedding by a term-level encoder. In this instance, only the document content inside a passage can interact with each other. However, the interaction between passages in a document is also important for capturing intra-document relations and enhancing fully fine-grained interaction inside the document. Besides, separating a document into passages is conducive to improving the efficiency of our model. Therefore, we propose a context-aware document encoder.

More specifically, we use intra-document attention and multi-head attention to retain local semantic information in the documents and carry on information propagation between a passage's representation from different perspectives. Following [33], we adopt a transformer's multi-layer encoder block Trm(·) to build the intra-document attention in document encoder to capture the novelty of the passage and the relations between passages in a document, shown in Equation (6).

$$\overline{\mathbf{P}}_{i1} : \overline{\mathbf{P}}_{im} = \text{Trm}(\mathbf{P}_{i1}^o : \mathbf{P}_{im}^o), \tag{6}$$

where $\mathbf{P}_{i1}^o : \mathbf{P}_{im}^o$ and $\overline{\mathbf{P}}_{i1} : \overline{\mathbf{P}}_{im}$ are the initial embedding and the context-aware representation of passage 1 to m of in document d_i, respectively.

Besides, we want to enhance the influence of information from the previous layer. Inspired by [18], we leverage a multi-head attention, denoted as MHA(·) in Equation (7) to aggregate the initial passage representation $\mathbf{P}_{ij}^o$ and the context-aware passage representation $\overline{\mathbf{P}}_{ij}$ obtained from the intra-document attention.

$$\mathbf{P}_{ij} = \text{MHA}(\overline{\mathbf{P}}_{ij}, \mathbf{P}_{ij}^o, \mathbf{P}_{ij}^o). \tag{7}$$

For the last layer of stack interaction blocks, we employ a mean pooling operation to aggregate all passage representations to generate a passage-aware document representation $\mathbf{D}_i^p$, calculated by Equation (8).

$$\mathbf{D}_i^p = \frac{1}{m}\sum_{j=1}^{m}\mathbf{P}_{ij}. \tag{8}$$

3.3.2 Context-Aware Sequence Encoder. The interaction between candidate documents plays a prominent role in comparing semantic information between different documents and gaining the global semantic perspective document representation. Besides, aggregating multi-grained document representation is important to perform information propagation, take advantage of different-grained document representation and capture important information in each document. Therefore, we design a document sequence encoder to interact different candidate documents and different granularity representations of documents, respectively.

To model the relations between different candidate documents and obtain a context-aware document representation that can better reflect the document's main content, we use a transformer encoder $\mathrm{Trm}(\cdot)$ as the inter-document attention to encode all candidate documents. It takes the initial embedding of all candidate documents $\mathbf{D}_1^o : \mathbf{D}_n^o$ obtained from Section 3.2.3 and produces the context-aware representation of all document $\overline{\mathbf{D}}_1 : \overline{\mathbf{D}}_n$.

$$\overline{\mathbf{D}}_1 : \overline{\mathbf{D}}_n = \mathrm{Trm}(\mathbf{D}_1^o : \mathbf{D}_n^o). \tag{9}$$

To enhance information propagation between different granularity representations of documents and get a better document representation, we adopt a multi-head attention to interact the passage representations $\mathbf{P}_{i1} : \mathbf{P}_{im}$ with the context-aware document representation $\overline{\mathbf{D}}_i$. We use the multi-head attention $\mathrm{MHA}(\cdot)$ because it can incorporate different aspects' information with fine-grained interaction. The document representation is regarded as Q and the passage representation is K which is equal to V. With multi-head attention, we obtain the global document representation $\mathbf{D}_i^d$ which involves multi-grained document representation.

$$\mathbf{D}_i = \mathrm{MHA}(\overline{\mathbf{D}}_i, \mathbf{P}_{i1} : \mathbf{P}_{im}, \mathbf{P}_{i1} : \mathbf{P}_{im}). \tag{10}$$

We treat the output of the Lth layer $\mathbf{D}_i$ as the final global perspective document representation $\mathbf{D}_i^d$.

3.3.3 Stacking Blocks. The stacking interaction blocks have L layers, in this section, we will introduce how the blocks are stacked.

From section 3.3.1, we have obtained passage representation $\mathbf{P}_{ij}$ for passage p_{ij} through the document encoder. We take $\mathbf{P}_{ij}$ as the next layer passage input replacing $\mathbf{P}_{ij}^o$ and feed it to the next layer intra-document attention. Similarly, through the sequence encoder illustrated in section 3.3.2, we have acquired document representation $\mathbf{D}_i^d$ for document d_i. The document representation $\mathbf{D}_1^d : \mathbf{D}_n^d$ is regarded as the next layer input $\mathbf{D}_1^o : \mathbf{D}_n^o$ for sequence encoder.

3.4 Passage-Aware Document-Subtopic Interaction

Different parts of the document may cover different subtopics and measuring the subtopic coverage differences of the document's different parts is important to provide a better diversified ranking. Existing works incorporate candidate documents and subtopics with a single document vector but ignore the passage-aware interaction between them which can estimate the coverage of subtopics more accurately. Besides, existing methods leverage google suggestions [21] as subtopics.[1] There

[1]Google query suggestions: http://playbigdata.ruc.edu.cn/dou/hdiv/

Table 2. Relevance Features Utilized by Existing Models

Features	Description	The length of features
TF-IDF	TF-IDF model	5
BM25	BM25 model with default parameters	5
LMIR	LMIR model with Dirichlet smoothing	5
PageRank	The score of PageRank	1
#Inlinks	inlinks' number	1
#Outlinks	outlinks' number	1

are some redundancy subtopics compared with the real user intents. With these considerations, we propose passage-aware document-subtopic interaction to relieve the subtopic's potential redundancy and address the document's fine-grained coverage of subtopics.

The passage-aware document-subtopic interaction consists of a multi-head attention MHA(·), shown in Equation (11). The subtopic representation $\mathbf{T}_1^o : \mathbf{T}_k^o$ of subtopic $t_1 : t_k$ and the passage representation $\mathbf{P}_{i1} : \mathbf{P}_{im}$ of passage $p_{i1} : p_{im}$ are fed into the multi-head attention as Q and $K = V$, respectively.

$$\overline{\mathbf{T}}_1 : \overline{\mathbf{T}}_m = \text{MHA}(\mathbf{P}_{i1} : \mathbf{P}_{im}, \mathbf{T}_1^o : \mathbf{T}_k^o, \mathbf{T}_1^o : \mathbf{T}_k^o), \tag{11}$$

where $\overline{\mathbf{T}}_1 : \overline{\mathbf{T}}_m$ is subtopic coverage of passage 1 to m.

To incorporate different parts' subtopic coverage of each document, we hire a mean pooling to obtain the $\mathbf{D}_i^t$ which represents the subtopic coverage of each document, shown in Equation (12).

$$\mathbf{D}_i^t = \frac{1}{m} \sum_{j=1}^{m} \overline{\mathbf{T}}_j. \tag{12}$$

3.5 The Final Ranking

In this section, we will introduce the process of combining the relevance feature with the diversity features with an MLP layer and obtaining the final diversity score of each document, denoted as Equation (13).

$$S_i = \text{MLP}(\mathbf{D}_i^p, \mathbf{D}_i^d, \mathbf{D}_i^t, \mathbf{R}_q, \mathbf{R}_i), \tag{13}$$

where $\mathbf{D}_i^p$ is the passage-grained document representation, shown in Section 3.3.1; $\mathbf{D}_i^d$ is the global perspective document representation, denoted in Section 3.3.2; $\mathbf{D}_i^t$ is the subtopic coverage representation, calculated in Section 3.4; $\mathbf{R}_q$ is the relevance feature of query q, and $\mathbf{R}_i$ is the relevance feature of the subtopic of document d_i.

Before obtaining the final score, we hire a linear layer to project the passage-aware document representation $\mathbf{D}_i^p$, the global document representation $\mathbf{D}_i^d$ and the subtopic coverage representation $\mathbf{D}_i^t$ into the same semantic space.

Same as previous work [23, 26], the relevance feature of query q consists of the 18 dimensional traditional relevance features, including BM25, TF-IDF, PageRank, and so on, shown in Table 2. More specifically, we regard the TF-IDF feature produced by the TF-IDF model, the BM25 feature which is BM25 with default parameters and LMIR with Dirichlet smoothing as the traditional relevance features. We also add the PageRank score, the number of inlinks and the number of outlinks as the relevance features. For a query, that has several subtopics, we calculate the relevance representation of the subtopic of document d_i with an MLP layer.

$$\mathbf{R}_i = \text{MLP}(x_{t_1}, x_{t_2}, ..., x_{t_k}), \tag{14}$$

where x_{t_k} is the traditional 18 dimensional relevance features of subtopic t_k.

3.6 Training and Optimization

Following previous works [23, 26, 32], we use the list-pairwise loss function for optimization. Because of the limited dataset for search result diversification, we adopt the list-pairwise sampling approach proposed by Jiang et al. [23] to generate the ground-truth ranking. Moreover, we inherit loss function based on the same sampling strategy as [23], shown in Equation (15).

$$\mathcal{L} = \sum_{q \in Q} \sum_{o=1}^{|O_q|} |\Delta M| [y^o \log(P(r^o_{1,2})) + (1 - y^o) \log(1 - P(r^o_{1,2})], \tag{15}$$

where O_q is the pair samples set of query q. ΔM is the weight of pair sample (r_1, r_2). $y^o = 1$ is the positive label and 0 is the negative. $P(r^o_{1,2})$ is the probability of pair (r_1,r_2) to be positive.

$$P(r^o_{1,2}) = \frac{1}{1 + \exp(s_{r_1} - s_{r_2})}, \tag{16}$$

$$\Delta M = |M(r_1) - M(r_2)|, \tag{17}$$

where $M(\cdot)$ evaluates the equality of the model's diversity ranking. $M(r_1) > M(r_2)$ implies that (r_1, r_2) is positive pair and $M(r_1) < M(r_2)$ shows (r_1, r_2) to be negative.

4 EXPERIMENTAL SETTINGS

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets. Same as previous works [23, 25, 26, 32], we conduct experiments on the ClueWeb09 dataset [2]. The ClueWeb09 dataset contains 200 queries of WebTrack dataset from TREC 2009 to 2012. Since the query #95 and #100 do not have diversity judgments, we only use 198 queries in our experiment. For each query, there are 3 to 8 annotated subtopics that come from Google query suggestions which are provided by Hu et al. [21]. In this work, we only use the first level of the subtopics with no hierarchical subtopics and the max subtopic number of the queries is 10. The weights of all subtopics are uniform.

4.1.2 Evaluation Metrics. Following previous work [25], we leverage the official diversity evaluation metrics of Web Track which contain ERR-IA [4], α−nDCG [6], and NRBP [7]. The definition of these evaluation metrics is introduced as follows.

ERR-IA. ERR-IA is the improvement of ERR. ERR considers the diversity between documents and can be computed as Equations (18) and (19).

$$\text{ERR} = \sum_{i=1}^{N} \delta(i) \Pr(i), \tag{18}$$

$$\Pr(i) = \prod_{k=1}^{i-1} (1 - r(k)) r(i), \tag{19}$$

where N stands for the number of documents, $\delta(i)$ is the position function, $r(i)$ is the probability that the user can be satisfied by document i.

ERR-IA takes into account the **intent-aware** (**IA**) and is given by

$$\text{ERR-IA} = \sum_{t=1}^{T} \sum_{i=1}^{N} \delta(i) \text{Pr-IA}(i, t), \tag{20}$$

$$\text{Pr-IA}(i, t) = \prod_{k=1}^{i-1} (1 - r(k, t)) r(i, t), \tag{21}$$

where T is the number of subtopics, $r(k,t)$ stands for the relevance of the kth document to the tth subtopic. In this article, we leverage the cosine similarity between the document and the subtopic to generate $r(k,t)$.

α-nDCG. α-nDCG aims at balancing the relevance and the diversity of the document ranking lists. The definition of α-nDCG is based on nDCG which is widely used in information retrieval. A document list should be given a higher nDCG when it puts a more relevant document in a higher position. nDCG is the normalization of DCG and calculated by Equation (22).

$$\text{nDCG(R)} = \frac{\text{DCG(R)}}{\text{DCG(R}^*\text{)}}, \tag{22}$$

where R is a document ranking list and R^* is the optimal document ranking. The DCG score of the document ranking list R is denoted in Equation (23).

$$\text{DCG(R)} = \sum_{i=1}^{N} \frac{2^{r(i)} - 1}{\log(i+1)}, \tag{23}$$

where N is the number of documents and $r(i)$ is the relevance score of the ith document in the document ranking list R .

To balance the relevance and the diversity of the document ranking lists, α-nDCG changes the definition of DCG in nDCG as Equation (24).

$$\widehat{\text{DCG(R)}} = \sum_{i=1}^{N} \frac{\text{CG}(i)}{\log(i+1)}, \tag{24}$$

$$\text{CG}(i) = \sum_{t=1}^{T} J_t(i)(1-\alpha)^{C_t(i-1)}, \tag{25}$$

where T is the number of intent, $J_t(i) = 1$ if the document ranked in the ith place is relevant to intent t and 0 otherwise, $C_t(i) = \sum_{k=1}^{i} J_t(k)$ is the number of documents relevant to intent t within top i and α is a parameter.

Based on the above equation of DCG, α-nDCG is defined as

$$\alpha - \text{nDCG} = \frac{\widehat{\text{DCG(R)}}}{\widehat{\text{DCG(R}^*\text{)}}} \tag{26}$$

NRBP. Clarke et al. [7] combined α-nDCG and the rank-biased precision and designed a new measure called **novelty- and rank-biased precision** (**NRBP**). NRBP is denoted as Equation (27).

$$\text{NRBP} = \frac{1-(1-\alpha)\beta}{T} \sum_{i=1}^{N} \beta^{i-1} \sum_{t=1}^{T} J_t(i)(1-\alpha)^{C_t(i-1)}, \tag{27}$$

where α and β are constants reflecting the user's declining interest when reading down the document ranking list, either because of finding the information or losing patience.

These matrices measure the diversity of document ranking by encouraging novelty and penalizing redundancy. All evaluation metrics are computed on the top 20 results of a ranking list to be consistent with previous search result diversification works and TREC Web Track. We conduct significance testing with p-value<0.05 two-tailed paired t-test.

4.2 Baselines

We evaluate the performance of our approach HAD by comparing it with four groups of methods.

4.2.1 Non-Diversified Methods. **Lemur** services[2] retrieved adhoc results produced by language model which is used as non-diversified results. **ListMLE** is a learning-to-rank and non-diversified method.

4.2.2 Explicit Methods. **xQuAD** [29], **PM2** [9], **TxQuAD/TPM2** [10], **HxQuAD/HPM2** [21] are representative unsupervised explicit methods which combine the relevance feature and the novelty feature of documents with a parameter λ. TxQuAD/TPM2 utilized subtopic terms to diversify document ranking. Besides, HxQuAD/HPM2 leveraged an additional parameter α with a hierarchical structure to model the subtopics. **DSSA** [23] is a supervised explicit method which adopted RNNs and attention mechanism to model the diversity of the documents.

4.2.3 Implicit Methods. Typical implicit models include **MO4SRD** [40], **R-LTR** [44], **PAMM** [35], **NTN** [36], **Graph4DIV** [32], and **DALETOR** [38] and all of them are supervised methods. We use the released code of MO4SRD and train it with Lemur results.[3] We utilize α−nDCG@20 as the optimization metrics and tune the number of positive rankings l^+ and negative rankings l^- per query, for PAMM. **R-LTR-NTN** and **PAMM-NTN** is the **neural tensor network** (**NTN**) used on R-LTR and PAMM. We re-implement Graph4DIV based on their open source code.[4] For DALETOR [38], we also reproduce it with Lemur results.

4.2.4 Ensemble Methods. **DESA** [26], **DVGAN** [25], and **GDESA** [27] are the representative ensemble diversity methods. With the attention mechanism, DESA combines the coverage of subtopics and the similarity of documents to produce the diversity score of documents simultaneously. DVGAN uses a generative adversarial network to generate training data that combines explicit and implicit features. GDESA combines the greedy selection with the global interaction to diversify document rankings.

4.3 Implementation Details

For the implementation of HAD, we first separate the content in a document into 7 passages, each passage's length w is 256, and the overlapping factor o is 64. We remove entities with probability scores less than or equal to 0.25 from all entities extracted by TagMe. We use the pretrained BERT-base model provided by Hugging Face [34] to build a term-level encoder without fine-tuning the parameter of BERT. For intra-document attention and inter-document attention, the number of attention heads are both 8, the number of layers are both 1 and the hidden dimension for the feed-forward network is 1536. The attention head for multi-head attention in document encoder, sequence encoder and passage-aware document-subtopic interaction is all 8. The layer number for stacking interaction blocks is 4. The out dimension for the linear projection in our model is 128. Besides, we train the model for 15 epochs with batch size 4. We use Adam as the optimizer with a learning rate of 9e-4. Same as previous works[26, 32], we select all hyper-parameters by 5-fold cross-validation based on the result of α−nDCG@20.

5 EXPERIMENTAL RESULTS

5.1 Overall Results

The overall experimental results are shown in Table 3. In general, our model HAD outperforms all existing models. This demonstrates the superiority of our model. We also have the following observations.

[2]Lemur service: http://boston.lti.cs.cmu.edu/Services/cluweb09_batch/
[3]MO4SRD: https://github.com/wildltr/ptranking
[4]Graph4DIV: https://github.com/su-zhan/Graph4DIV

Table 3. Performance Comparison of all Methods

Category	Method	ERR-IA		α−nDCG		NRBP	
Non-diversified	Lemur	.271†		.369†		.232†	
	ListMLE	.287†		.387†		.249†	
Explicit	xQuAD	.317†	+7.0%	.413†	+6.7%	.284†	+7.7%
	TxQuAD	.308†	+7.9%	.410†	+7.0%	.272†	+9.9%
	HxQuAD	.326†	+6.1%	.421†	+5.9%	.294†	+6.7%
	PM2	.306†	+8.1%	.411†	+6.9%	.267†	+9.4%
	TPM2	.291†	+9.6%	.399†	+8.1%	.250†	+11.1%
	HPM2	.317†	+7.0%	.420†	+6.0%	.279†	+8.2%
	DSSA	.356†	+3.1%	.456†	+2.4%	.326†	+3.5%
Implicit	MO4SRD	.283	+10.4%	.367	+11.3%	.252	+10.9%
	R-LTR	.303†	+8.4%	.403†	+7.7%	.267†	+9.4%
	PAMM	.309†	+7.8%	.411†	+6.9%	.271†	+9.0%
	R-LTR-NTN	.312†	+7.5%	.415†	+6.5%	.275†	+8.6%
	PAMM-NTN	.311†	+7.6%	.417†	+6.3%	.272†	+8.9%
	DALETOR	.364†	+2.3%	.461†	+1.9%	.333†	+1.8%
	Graph4DIV	.370	+1.7%	.468	+1.2%	.338	+2.3%
Ensemble	DESA	.363†	+2.4%	.464†	+1.6%	.332†	+2.9%
	DVGAN	.367†	+2.0%	.465†	+1.5%	.334†	+2.7%
	GDESA	.369	+1.8%	.469	+1.1%	.337	+2.4%
	HAD	**.387**	–	**.480**	–	**.361**	–

The best result is in bold. †indicates significant improvements obtained by our model in two tailed paired t-test with p−value<0.05.

(1) HAD significantly outperforms all non-diversified models and explicit models in all evaluation metrics in two tailed paired t-test with $p < 0.5$. In terms of α-nDCG, the improvement over the best explicit baseline model DSSA is 2.4%. DSSA leveraged attention mechanism to select the next document and model the subtopic coverage of the whole document at each step. In contrast, HAD compares the passage-level subtopic coverage differences of documents which can capture the subtopic coverage of the document more accurately.

(2) Our model HAD also outperforms all implicit models by a large margin, including Graph4DIV, which is the state-of-the-art method for search result diversification. Compared with Graph4DIV, HAD improves ERR-IA from 0.370 to 0.387. Graph4DIV compares the intent coverage of documents and builds an intent graph to select documents iteratively. However, HAD leverages different granularities document content and scores documents simultaneously which can achieve better performance and has a faster ranking speed. We will further discuss the balance of efficiency and effectiveness in Section 5.4.

(3) Compared with ensemble models, HAD achieves significantly better performance. In terms of DESA, NRBP is improved by 2.9%. DESA encodes each candidate document to a single vector and conducts interaction between different documents. Conversely, HAD adopts a hierarchical attention structure to model inter-document and intra-document interaction from multi-grained perspectives. The results confirm the benefit of hierarchical structure for improving the performance of search result diversification.

In summary, the results indicate that **hierarchical attention framework for search result diversification is conducive to refining document multi-grained representation, carrying out fine-grained interaction, capturing intra-document relations, and promoting diversity search**. To test the model in more detail, we conduct several supplementary experiments.

Table 4. Performance of HAD with Different Document Representations

Method	ERR-IA	α−nDCG	NRBP
HAD	**.387**	**.480**	**.361**
w/o $\mathbf{D}_i^p$	.379	.476	.350
w/o $\mathbf{D}_i^d$	.379	.476	.349

Table 5. Performance of HAD with Different Components

Method	ERR-IA	α−nDCG	NRBP
HAD	**.387**	**.480**	**.361**
w/o $\mathbf{D}_i^t$	.380	.476	.352
w/o MHA-document	.377	.473	.348
w/o MHA-sequence	.376	.471	.346
w/o intra-document attn	.381	.477	.353
w/o inter-document attn	.381	.477	.353

5.2 Ablation Studies

To verify the necessity of each component, we conduct ablation experiments on our model HAD.

5.2.1 Effects of different types' document representations. Table 4 shows experiments on the effectiveness of different types' document representations:

(1) *w/o* $\mathbf{D}_i^p$: without passage-aware document representation $\mathbf{D}_i^p$.
(2) *w/o* $\mathbf{D}_i^d$: without global document representation $\mathbf{D}_i^d$.

As expected, each individual feature contributes to the whole. Without any of them, our model fails to fully utilize rich contextual information and drops seriously in terms of all metrics. Specifically, without passage-aware document representation $\mathbf{D}_i^p$, ERR-IA drops from 0.387 to 0.379 by 0.8% and NRBP declines from 0.361 to 0.350 by 1.1%. These results prove that modeling document content from multi-grained perspectives is beneficial and can promote the ability of our model to understand the document content.

5.2.2 Influence of Different Components. We explore the role of each component in our model, and the result is shown in Table 5, including:

(1) *w/o* $\mathbf{D}_i^t$: without passage-aware document-subtopic interaction and becoming an implicit model.
(2) *w/o* MHA-document: removing multi-head attention in document encoder.
(3) *w/o* MHA-sequence: without multi-head attention in sequence encoder.
(4) *w/o* intra-document attn: removing intra-document interaction within each document.
(5) *w/o* inter-document attn: removing inter-document interaction between candidate documents.

As shown in Table 5, the removal of each component will damage the results on all evaluation metrics. Concretely, without multi-head attention in sequence encoder causes the most obvious impact on performance with α-nDCG dropping from 0.480 to 0.471 and NRBP decreasing to 0.346 from 0.361. This indicates the importance of integrating multi-grained document representation and enhancing information propagation within these multi-grained perspectives' document representation. Additionally, we find that removing the intra-document attention causes an obvious drop in all evaluation metrics. This demonstrates that interacting passages within a document is

Table 6. Performance of HAD with Different Pooling Schemes

Method	ERR-IA		α−nDCG		NRBP	
#mean pooling	**.387**	–	**.480**	–	**.361**	–
#max pooling	.386	−0.1%	.479	−0.1%	.359	−0.2%

Table 7. Performance of HAD with Different Initial Embeddings

Method	ERR-IA		α−nDCG		NRBP	
HAD-BERT	**.387**	–	**.480**	–	**.361**	–
HAD-d2v	.383	−0.4%	.475	−0.5%	.358	−0.3%
DESA-BERT	.373	−1.4%	.466	−1.4%	.346	−1.5%
DESA-d2v	.363	−2.4%	.464	−1.6%	.332	−2.9%

useful for comparing content between different passages and capturing subtle differences between them. Moreover, without passage-aware document-subtopic interaction causes a significant decline in all metrics. ERR-IA and NRBP have decreased by 0.6% percent and 0.8%, respectively. This demonstrates the effectiveness of introducing fine-grained interaction between documents and subtopics.

5.2.3 Different Pooling Schemes. We verify the influence of different pooling schemes and the results are denoted in Table 6. Compared with the mean pooling operation used in our model, we adopt max pooling to test the influence of the pooling method on our model. As shown in Table 6, the two pooling operations have similar experimental results which show the stability of our model.

5.2.4 Different Initial Embedding. We study the different initial embedding generation methods, including doc2vec [24] and BERT [11]. Similar to previous work, the dimensions of doc2vec and BERT vector are 100 and 768, respectively. Experimental results are reported in Table 7.

For comparison, we implement DESA [26] with doc2vec and BERT vector as initial embedding and the result is shown as DESA-d2v and DESA-BERT, respectively. Both the doc2vec and the BERT vision of HAD shown as HAD-d2v and HAD-BERT outperform DESA. More specifically, α-nDCG is improved from 0.466 to 0.480 by changing the base model from DESA to our model HAD. These results show the superiority of our model and verify the effectiveness of the hierarchical attention structure. Besides, these experiment results also testify that our model is not affected by the initial embedding and its robustness.

5.3 Performance of Different Hyperparameters

In our model, we design stacking interaction blocks and divide a document into several passages. To investigate the impact of the stacking interaction blocks' layer number and the passage number on our model, we train our model with different hyperparameter settings and test their performance. The result is shown in Table 8 and Table 9.

5.3.1 Layer Number. Considering the layer number, according to the structure of our model, a larger layer number will cause deeper stacking interaction blocks while a lower number will lead to inadequate learning. As shown in Table 8, layer number 3 and layer number 6 both cause the results to decline. Indeed, 4 is the best choice for layer number in HAD without insufficient training and overfitting. According to the results, our model is not sensitive to the hyperparameter layer number. Changes in the layer number will not cause obvious changes in the results.

Table 8. Performance of HAD with Different Layer Numbers

Parameter	Value	ERR-IA	α−nDCG	NRBP
Layer Number	3	.385	.479	.358
	4	**.387**	**.480**	**.361**
	5	.387	.480	.361
	6	.386	.479	.359

Table 9. Performance of HAD with Different Passage Numbers

Parameter	Value	ERR-IA	α−nDCG	NRBP
Passage Number	4	.378	.474	.350
	5	.382	.476	.364
	6	.386	.479	.359
	7	**.387**	**.480**	**.361**
	8	.387	.480	.361

Table 10. Test Time Per Query of DESA, Graph4DIV, and HAD

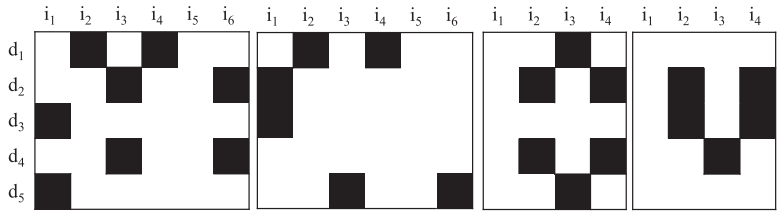
Method	test time / query (ms)		α−nDCG	
DESA	0.7	+1.7	.464	+1.6%
Graph4DIV	55.5	−53.1	.468	+1.2%
HAD	2.4	−	.480	−

5.3.2 Passage Number. As for the passage number, we experiment from 4 to 8. From Table 9, we can observe that search result diversification benefits from more document content. As the passage number grows, HAD shows a more powerful ability to diversify document ranking. With more passages, we can learn a document representation that can reflect the content of the document more comprehensively. According to Table 9, when the passage number is 7, our model gets the best performance. When the passage number exceeds 7, the results do not continue to increase because (1) a lot of documents contain less than 8 passages. Even if we increase the passage number, they will be padding. (2) the content at the end of the document does not contain new semantic information and the previous part of the document has covered the main content of the document.

5.4 Efficiency-Effectiveness Analysis

In this section, we evaluate the efficiency and effectiveness of our model HAD compared with DESA and Graph4DIV. All experiments are conducted on a single TITAN V GPU.

As shown in Table 10, **HAD has achieved tremendous acceleration compared with greedy selection model Graph4DIV and a slight increase in time compared with simultaneously scoring model DESA with the significant outperform in effectiveness.** The test time per query of HAD is 2.4 ms, and compared with the best greedy selection model Graph4DIV, it has reduced 53.1 ms with the increase of α-nDCG by 1.2%. Compared with the best simultaneously scoring model DESA, HAD improves the α-nDCG from 0.464 to 0.480 by test time per query increasing 1.7 ms. The results show that our model can improve diversified search results with better efficiency. Because most operations in our model can be calculated in parallel and the candidate document scores in our model can be calculated simultaneously without iteratively selecting the document, the actual processing time of our model is less than the theoretical analysis.



(a) Query #60 (b) Query #152

Fig. 2. Case study for the ranking of HAD and Graph4DIV. Black means relevant and white means irrelevant.

Table 11. The Annotated user Intents of Query #60 and #152

Query #60: bellevue
i_1: Find information about Bellevue, Washington.
i_2: Find information about Bellevue, Nebraska.
i_3: Find information about Bellevue Hospital Center in New York, NY.
i_4: Find the homepage of Bellevue University.
i_5: Find the homepage of Bellevue College, Washington.
i_6: Find the homepage of Bellevue Hospital Center in New York, NY.
Query # 152: angular cheilitis
i_1: What home remedies are there for angular cheilitis?
i_2: What causes angular cheilitis?
i_3: What is the common name for angular cheilitis?
i_4: How do you treat severe angular cheilitis?

5.5 Case Study

In this section, we conduct quantitative analyses to verify the effectiveness of our model. We also carry on a case study to visualize the ranking results of HAD on specific queries.

First, we compute the average ratio of the subtopics not covered by previous documents. According to Equation (28), an intent can achieve a higher score if it remains uncovered by previous documents and has a higher ranking position. The score of HAD and Graph4DIV is 2.98 and 2.83, respectively. These results confirm that our model prioritizes documents that cover unexplored subtopics over those that do not.

$$S_{\text{avg}} = \frac{1}{|Q|} \sum_{q \in Q} \sum_{d_i \in q, d_i \in \mathcal{D}} \sum_{t \in d_i} P_t(i) \frac{1}{\text{pos}_{d_i}}, \tag{28}$$

where $d_i \in q$ means d_i is the document ranked in the ith place under query q, t is the user intent contained by d_i, $P_t(i) = 1$ if intent t not covered by previous documents and 0 otherwise, $\frac{1}{\text{pos}_{d_i}}$ is the ranked position of d_i.

Second, we randomly select two queries and visualize their ranking results. Figure 2 shows the top 5 ranking results of query #60 and #152, where black means relevant and white means irrelevant. We also show the annotated user intents of query #60 and #152 in Table 11. We choose Graph4DIV for comparison, which is the best existing model. For query #60, HAD ranks a document relevant to intent i_3 and i_6 at the second position, and a document covering i_1 at the third place, while Graph4DIV ranks two documents relevant to subtopic i_1 in these two positions. This

indicates that Graph4DIV cannot identify the subtopic difference between documents as accurately as HAD which will lead to documents covering the same subtopic being ranked together. For query #152, the document ranking first in HAD covers subtopic i_3 and the second document covers subtopic i_2, and i_4. For the ranking result of Graph4DIV, the first document does not cover any subtopic and the second document covers subtopic i_2 and i_4. These results further verify that modeling intra-document and inter-document relations from multi-grained perspectives can help our model depict the subtopic divergence of documents more accurately.

6 CONCLUSION

In this article, we proposed a hierarchical attention framework that models the multi-grained document content and conducts intra-document and inter-document interaction to diversify search results. First, we separate the document into passages and extract pivot entities from documents. Moreover, we encode the passages, documents, and subtopics and integrate additional knowledge into their initial embedding. Next, we design stacking interaction blocks to capture the document's intra-document and inter-document relations. Besides, we propose a passage-aware document-subtopic interaction to perform fine-grained interaction between the document and subtopics. With a linear projection, we combine the diversity features with the relevance features and produce the final ranking score. Experimental results confirm our model can effectively model intra-document and inter-document relations from multi-grained perspectives.

In the future, we plan to mine more fine-grained intent from documents by leveraging more passage level information and selecting more useful passages in the document content. Besides, we also intend to design a model that can highlight the main content of the document in order to reduce the complexity of the model and measure the novelty of documents more accurately.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Retrieved from http://arxiv.org/abs/1409.0473

[2] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 web track. In *Proceedings of The Eighteenth Text REtrieval Conference TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009* (NIST Special Publication), National Institute of Standards. Retrieved from http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf

[3] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–336.

[4] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. 621–630.

[5] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555 (2014). Retrieved from http://arxiv.org/abs/1412.3555

[6] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 659–666.

[7] Charles L. A. Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of the Conference on the Theory of Information Retrieval*. Springer, 188–199.

[8] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.

[9] Van Dang and W. Bruce Croft. 2012. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 65–74.

[10] Van Dang and W. Bruce Croft. 2013. Term level search result diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland - July 28 - August 01, 2013*. Gareth J. F. Jones, Paraic Sheridan, Diane Kelly, Maarten de Rijke, and Tetsuya Sakai (Eds.), ACM, 603–612. DOI: https://doi.org/10.1145/2484028.2484095

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics. 4171–4186. DOI: https://doi.org/10.18653/V1/N19-1423

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is Worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net. Retrieved from https://openreview.net/forum?id=YicbFdNTTy

[13] 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.

[14] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. 1625–1628.

[15] Chengzhen Fu, Enrui Hu, Letian Feng, Zhicheng Dou, Yantao Jia, Lei Chen, Fan Yu, and Zhao Cao. 2022. Leveraging multi-view inter-passage interactions for neural document ranking. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 298–306.

[16] Songwei Ge, Zhicheng Dou, Zhengbao Jiang, Jian-Yun Nie, and Ji-Rong Wen. 2018. Personalizing search results using hierarchical RNN with query-aware attention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22–26, 2018*. Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.), ACM, 347–356. DOI: https://doi.org/10.1145/3269206.3271728

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[20] Sebastian Hofstätter, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. 2021. Intra-document cascading: Learning to select passages for neural document ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1349–1358.

[21] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 63–72.

[22] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management* 36, 2 (2000), 207–227. DOI: https://doi.org/10.1016/S0306-4573(99)00056-4

[23] Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou, Wayne Xin Zhao, Jian-Yun Nie, and Ming Yue. 2017. Learning to diversify search results via subtopic attention. In *Proceedings of the 40th international ACM SIGIR Conference on Research and Development in Information Retrieval*. 545–554.

[24] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1188–1196.

[25] Jiongnan Liu, Zhicheng Dou, Xiaojie Wang, Shuqi Lu, and Ji-Rong Wen. 2020. DVGAN: A minimax game for search result diversification combining explicit and implicit features. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 479–488.

[26] Xubo Qin, Zhicheng Dou, and Ji-Rong Wen. 2020. Diversifying search results using self-attention network. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19–23, 2020*. Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.), ACM, 1265–1274. DOI: https://doi.org/10.1145/3340531.3411914

[27] Xubo Qin, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2022. GDESA: Greedy diversity encoder with self-attention for search results diversification. *ACM Transactions on Information Systems* 41, 2 (2022), 34:1–34:36.

[28] Chen Qu, Chenyan Xiong, Yizhe Zhang, Corby Rosset, W. Bruce Croft, and Paul Bennett. 2020. *Contextual Re-Ranking with Behavior Aware Transformers*. Association for Computing Machinery, New York, NY, USA, 1589–1592. DOI: https://doi.org/10.1145/3397271.3401276

[29] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World wide web*. 881–890.

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

[31] Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2007. Identifying ambiguous queries in web search. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8–12, 2007*. Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy (Eds.), ACM, 1169–1170. DOI: https://doi.org/10.1145/1242572.1242749

[32] Zhan Su, Zhicheng Dou, Yutao Zhu, Xubo Qin, and Ji-Rong Wen. 2021. Modeling intent graph for search result diversification. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021*. Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.), ACM, 736–746. DOI: https://doi.org/10.1145/3404835.3462872

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

[35] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 113–122.

[36] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling document novelty with neural tensor network for search result diversification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 395–404.

[37] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, Wei Zeng, and Xueqi Cheng. 2017. Adapting markov decision process for search result diversification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 535–544.

[38] Le Yan, Zhen Qin, Rama Kumar Pasumarthi, Xuanhui Wang, and Michael Bendersky. 2021. Diversification-aware learning to rank using distributed representation. In *Proceedings of the Web Conference 2021*. 127–136.

[39] Sevgi Yigit-Sert, Ismail Sengor Altingovde, Craig Macdonald, Iadh Ounis, and Özgür Ulusoy. 2020. Supervised approaches for explicit search result diversification. *Information Processing and Management* 57, 6 (2020), 102356.

[40] Hai-Tao Yu. 2022. Optimize what you evaluate with: Search result diversification based on metric optimization. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, 45th Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The 12th Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 – March 1, 2022*. AAAI Press, 10399–10407. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/21282

[41] Yisong Yue and Thorsten Joachims. 2008. Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th International Conference on Machine Learning*. 1224–1231.

[42] 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, SIGIR 2020, Virtual Event, China, July 25–30, 2020*. Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.), ACM, 1111–1120. DOI: https://doi.org/10.1145/3397271.3401175

[43] Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2021. Enhancing potential re-finding in personalized search with hierarchical memory networks. *IEEE Transactions on Knowledge and Data Engineering* 35, 4 (2021), 3846–3857.

[44] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu. 2014. Learning for search result diversification. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 293–302.

Received 8 November 2022; revised 3 January 2024; accepted 6 March 2024