Passage-aware Search Result Diversification

ZHAN SU, School of Information, Renmin University of China, Beijing, China
ZHICHENG DOU, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
YUTAO ZHU, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
JI-RONG WEN, Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, Beijing, China

Research on search result diversification strives to enhance the variety of subtopics within the list of search results. Existing studies usually treat a document as a whole and represent it with one fixed-length vector. However, considering that a long document could cover different aspects of a query, using a single vector to represent the document is usually insufficient. To tackle this problem, we propose to exploit multiple passages to better represent documents in search result diversification. Different passages of each document may reflect different subtopics of the query and comparison among the passages can improve result diversity. Specifically, we segment the entire document into multiple passages and train a classifier to filter out the irrelevant ones. Then the document diversity is measured based on several passages that can offer the information needs of the query. Thereafter, we devise a passage-aware search result diversification framework that takes into account the topic information contained in the selected document sequence and candidate documents. The candidate documents' novelty is evaluated based on their passages while considering the dynamically selected document sequence. We conducted experiments on a commonly utilized dataset, and the results indicate that our proposed method performs better than the most leading methods.

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

Additional Key Words and Phrases: Search result diversification, passage ranking, intra-document attention

ACM Reference Format:
Zhan Su, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2024. Passage-aware Search Result Diversification. *ACM Trans. Inf. Syst.* 42, 5, Article 136 (May 2024), 29 pages. https://doi.org/10.1145/3653672

1 INTRODUCTION

Search result diversification is an essential research topic in the information retrieval area. Considering that accurately capturing users' genuine search intentions solely through brief and vague queries poses a challenge, diversification methods can improve the search results by offering diverse documents that cover different subtopics of a query. In terms of incorporating subtopics, most search result diversification approaches fall into two categories: *explicit* methods and *implicit*

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' addresses: Z. Su, School of Information, Renmin University of China, 59 Zhongguancun Street, Haidian District, Beijing; e-mail: suzhan@ruc.edu.cn; Z. Dou (Corresponding author) and Y. Zhu, Gaoling School of Artificial Intelligence, Renmin University of China, 59 Zhongguancun Street, Haidian District, Beijing; e-mails: dou@ruc.edu.cn, yutaozhu94@gmail.com; J.-R. Wen, Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, Beijing, China; e-mail: jirong.wen@gmail.com.

ACM 1046-8188/2024/05-ART136
https://doi.org/10.1145/3653672

Query: Battles in the Civil War

Passages of the document	Subtopics
P_1: The United States Civil War featured many military actions. Among the most significant were the First Battle of Bull Run, the Battle of Shiloh, the Battle of Antietam …	The major **battles** in the US civil war
P_2: Over the course of the war, the Commanding General of the United States Army was, in order of service, Winfield Scott, George B. McClellan, Henry Halleck …	Famous **generals** in the Civil War
P_3: Greatest Movies About the American Civil War Gone with the Wind (1939) ... Horse Soldiers (1959) ... Glory (1989) ... Cold Mountain (2003)…	**Movies** about the Civil War battles

Fig. 1. Example of multiple passages and the corresponding subtopics within a document. Different passages within a single document can cover different subtopics of the query. Specifically, passages P_1, P_2, and P_3 cover the subtopics of major battles, famous generals, and movies in the Civil War, respectively.

methods. The explicit approaches [15, 23, 38] assess the documents' diversity from the perspective of the query. In this way, a query is extended to multiple query aspects (a.k.a subtopics), and the search results' diversity is evaluated based on their coverage of these subtopics. While the implicit methods [10, 40, 48, 57] model diversity from the document side, namely punishing the redundant documents and encouraging the novel documents, without explicitly involving subtopics. Both explicit and implicit methods have their own merits and can achieve excellent performance.

Most implicit methods [40, 47, 48, 57] either treat the document as several handcraft ad hoc features (e.g., TF-IDF and BM25) or a pre-trained document representation (e.g., doc2vec). From our perspective, both representation approaches view a document as a whole. As a result, signals from different parts of the document are mixed up, which is a drawback for the ranking models to sense different subtopics contained in the diverse documents. Besides, it is not suitable to compress different documents into a fixed-length vector, since the document length can vary largely. Inspired by the explicit way of expanding the query into multiple subtopics, in this work, we propose to present a document as multiple passages and model document diversity at the passage level for implicit methods.

There are several advantages of presenting documents as multiple passages compared with the traditional ways of search result diversification. *First*, it is more appropriate to present subtopic information. Different passages may contain various subtopic information in a long document. For example, as shown in Figure 1, a document relevant to the query "Battles in the Civil War" contains three passages P_1, P_2, and P_3, each of which covers a distinct subtopic. If we adopt the widely used presentation approaches by encoding the whole document as a single vector, then information from different passages (or subtopics) will be hard for the model to sense. Hence, in the search result diversification tasks, passages are more suitable units for presenting subtopic information than documents (in traditional ways), as diverse documents often cover many subtopics. *Second*, it is more capable to preserve document content. The traditional ad hoc features mainly model documents at the term level (such as BM25), while passages are coherent semantic units that can better preserve the contents. Compared with the term-level features, passage representations contain context information that is relevant to the queries. *Third*, it is more flexible to present documents of different lengths. The traditional one vector representation approach

can overwhelm the popular document encoders (like BERT [17]) in the case of long documents. Although some methods [3, 5, 20] have been proposed to alleviate this problem, a more straightforward approach is to divide the long documents into passages. In this way, short documents have fewer passages while long documents have more passages. The effects of the passages are also validated in IR [7, 9, 21, 30, 37, 44].

Given that search result diversification is NP-hard [1], most explicit and implicit methods adopt a greedy selection strategy, namely continuously choosing the most novel document from the remaining candidate documents concerning the current sequence of selected documents. Therefore, measuring whether each specific sub-intent of the query has been satisfied by the selected documents is an essential part of this greedy framework. In this article, we focus on the implicit search result diversification methods and propose a **Passage-Aware Diversification model (PAD)** with the greedy framework.

More specifically, the diversification is done in the steps as follows. (1) We train a passage relevance classifier to evaluate the quality of passages and select the most representative passages for each document based on the relevance scores. This approach ensures that only a limited number of passages are provided to the diversification model, thereby maintaining efficiency to an acceptable level. (2) We design a context-aware passage interaction framework that can automatically aggregate the passage information to form document representation considering the selected documents at each step of the greedy selection process. The next selected document should cover those query intents that are less satisfied by the list of selected documents until now. In the framework, we devise three passage-aware encoders, namely the **global passage encoder (GloEnc)**, the **selected documents encoder (SelEnc)**, and the candidate **document encoder (DocEnc)**, to discover the rich information contained in the passages. The GloEnc will provide a global view of the passages from all documents, while the SelEnc will aggregate the passage information from the selected document sequence. The features from the SelEnc reflect the current satisfied state of the query, which is important for the model to select the novel documents that cover the subtopics with lower satisfaction. Besides, to comprehensively leverage the different passage representations, we do not directly use the linear combination of the passage representations as the final document representation. Considering that the document novelty is dynamically changed depending on the selected document sequence, we take the features from the selected sequence into account and design a DocEnc to automatically generate the document diversity features from multiple passages with an attention mechanism. (3) Each candidate document's novelty score is then measured by comparing its passage-aware representation outputted by DocEnc with the representations of the current selection states outputted by SelEnc.

Our experiments are conducted on the widely utilized TREC Web Track dataset. Experimental results show that our PAD model outperforms existing methods, demonstrating the effectiveness of modeling passages in search result diversification.

Our contributions can be summarized as follows:

(1) We propose to exploit several essential passages to better represent diverse documents that cover multiple subtopics in search result diversification. We reveal that modeling passages can capture the fine-grained interaction between the documents and, hence, improve diversification quality.
(2) We develop a relevance classifier to identify whether a passage is relevant to the query. With the classifier, we are able to select important passages from a document other than simply including all passages, which reduces the noise and improves the efficiency of the diversification model.
(3) We devise a passage-aware framework that can automatically derive the document diversity from the passage relationships via passage interactions between the selected documents and

Table 1. Types of Search Result Diversification Methods

	Implicit	Explicit and Ensemble
Supervised	**PAD** (Our approach), KEDIV, DALETOR, Graph4DIV, NTN, PAMM, R-LTR, SVM-DIV,	GDESA, DESA, DVGAN, DSSA
Heuristic	MMR	HPM2, TPM2, HxQuAD, TxQuAD, PM2, xQuAD, IA-Select

Our approach is in bold.

candidate documents, which is helpful to dynamically capture the documents that cover the less satisfied query intents.

2 RELATED WORK

2.1 Search Result Diversification

Previously proposed search result diversification approaches mainly cover three categories: implicit methods, explicit methods, and ensemble methods. As discussed in Section 1, implicit approaches focus on document novelty, while explicit methods leverage subtopics to evaluate document diversity. The ensemble approaches exploit these two types of features for diversified ranking. The rough categorization of these methods is shown in Table 1. Heuristic methods leverage some handcrafted signals to measure the novelty of different documents, while supervised methods mainly exploit the document representations to automatically extract diversity features for diversified learning to rank with human-annotated diversity labels.

(1) **Implicit Methods.** The groundbreaking work of the search result diversification is the **maximal marginal relevance (MMR)** [10], which took the diversity and relevance of documents into account and adjusted their weights via a tunable parameter λ. The calculation of document d_i's MMR score is shown as follows:

$$\text{MMR}(d_i, q) = \lambda \text{Sim}(d_i, q) - (1 - \lambda) \max_{d_j \in \mathcal{S}} \text{Sim}(d_i, d_j), \tag{1}$$

where $\mathcal{S}$ is the selected document sequence and Sim(.) is the function measuring the similarity of documents and queries. The combining way of evaluating a document's diversity and relevance is adopted by most following researches [23, 34, 38, 40, 57]. Early research [39] found that novelty can break the tie between similarly diverse results. Recent diversification researches focus more on supervised approaches. For example, Yue and Joachims [55] leveraged structural-SVM to predict the document sets that cover different subtopics based on word-level features. Zhu et al. [57] took search result diversification as **relational learning to rank (R-LTR)** based on several classical human-designed document features. PAMM [47] learned to enlarge the predicted score margin between the positive document rankings and negative rankings. Xia et al. [48] leveraged a **neural tensor network (NTN)** to automatically generate diversity features from document representations to reduce the workload of manually designing novelty features. Yan et al. [51] proposed a differentiable loss to optimize the diversified ranking model. Yu [54] proposed a probabilistic scoring function to determine document's rank position. Graph4DIV [40] is an implicit graph-based method that presents the document relationship on the graph. Other methods introduce external sources (e.g., Knowledge Graph) to model document diversity. For example, KEDIV [41] exploits the knowledge base to measure the relations of the queries and documents in the search result diversification. Graph4DIV and KEDIV focus on the document relationship modeling and achieve good performance. Different from these two approaches, we exploit multiple passages and their relations to measure document diversity in this article.

Compared with the methods above, our PAD is also an implicit method. However, PAD can leverage the passage-level information to consider the inner-document passage structure and inter-document relationship during the diversification process, which is helpful in capturing a more fine-grain selection status in the selection procedure.

(2) **Explicit Methods.** Further from MMR, Santos et al. [38] explicitly calculated the document's probability on different query aspects (a.k.a subtopics) in the diversification task. Besides, Dang and Croft [15] proposed that the proportions of different aspects in the final result list should be consistent with the popularity of subtopics. Several methods (e.g., TxQuAD [16] and HPM2 [22]) are proposed based on the framework of xQuAD [38] and PM2 [15], which leveraged document terms or hierarchical structure to improve the diversity. Although some methods (e.g.TPM2 and TxQuAD [16]) model document diversity at the word level, the sequential information of the passages contained in the documents is ignored. Besides, some explicit methods consider document coverage to different query aspects. For example, Ozdemiray and Altingovde [33] adopted score and rank aggregation techniques in diversification. Our method PAD takes the passage, the continuous semantic units, into account and leverages the passage difference to measure the document diversity.

Apart from the heuristic methods, explicit supervised diversification approaches also acquire excellent performance. For example, Yigit-Sert et al. [53] exploited several query performance predictors to evaluate the document's coverage on different subtopics. Besides, DSSA [23] is a representative supervised explicit method that uses RNN and attention to automatically measure the coverage of the document's subtopic during the greedy selection. Different from DSSA, our implicit approach models the diversity of the documents based on passages of the documents, which enables our model to detect different passages that answer different query intents.

(3) **Ensemble Methods.** DVGAN [27] and DESA [34] are ensemble approaches that leverage both implicit (document dissimilarity) features and explicit (subtopic) features. DVGAN measured the document diversity under a Generative Adversarial Network, while DESA exploited the subtopic features to evaluate the diversity of the entire document ranking at once. The implicit part of DESA and DVGAN is similar to most implicit methods. Based on the DESA, Qin et al. [35] introduced a greedy selection strategy to further improve the diversity ranking capability of DESA. Ouyang et al. [32] leveraged graph neural network to automatically model the relations of the documents. Compared with our model PAD, they all model the document as a whole. We are dedicated to improving the document representations for implicit methods in this article. The passage-level features proposed by our approach can also be further used to enhance these document-level methods.

(4) **Other Methods.** There are also methods [18, 49, 50] exploring other ranking strategies apart from the greedy strategy. For example, inspired by the browsing behaviors of users, MDP-DIV [49] exploited the Markov decision process in the dynamic document ranking procedure. Following MDP-DIV, M^2Div [18] explored the possible rankings utilizing the Monte Carlo tree search within the framework of the MDP process. Moreover, Xu et al. [50] presented a pairwise policy gradient strategy for evaluating two document rankings, aiming to improve overall ranking quality. Liang et al. [26] considered diversification and personalization jointly. Compared with these document-level methods, we propose leveraging passage to measure document diversity and devise a passage-aware diversification framework to model passage-level features.

2.2 Passage-based Ranking Methods

Considering that some popular pre-trained language models (e.g., BERT) face the difficulty of representing long documents, some content will be neglected if we only encode the document within one vector during the representation procedure. Although many efforts have been paid to develop

Table 2. Symbols and Descriptions in PAD

Symbols	Descriptions
q, Q	a query, and the entire query set, $q \in Q$
$\mathcal{D}$	the query q's documents set, supposing $\|\mathcal{D}\| = n$
$\mathcal{S}$	the selected document set for the query q, $\mathcal{S} = \emptyset$ at initial state
C	the remnant candidate document set, $C = \mathcal{D} \setminus \mathcal{S}$, $C = \mathcal{D}$ at initial state
$\mathcal{P}_s$	the passage set of the selected documents
$\mathcal{P}_i$	the passage set of d_i used for diversity ranking
$\mathbf{R}_i$	the ith document d_i's relevance features
$\mathbf{H}_i$	the ith document d_i's diversity features
$\circ$	the element-wise product of two vectors
$[;]$	the concatenation of features

long-document encoders [3, 5, 20], splitting the documents into several passages and modeling them in the ranking process is also a promising direction [2, 21].

Previous studies [6, 7, 9, 24, 28, 30, 37, 44] have shown the effects of passage-based retrieval methods. For example, Krikon et al. [25] incorporated the similarity between passages from different documents into the ranking process. Wu et al. [45] focused on modeling context information for the document ranking task at the passage level. IDCM [21] is an intra-document cascading model that selects the top-k passage before the passage ranking. Dai and Callan [14] leveraged BERT [17] to encode passages from the documents and used several scores (first, best, and sum of all passages) for ranking. Similarly to these studies, we also utilize passages to comprehensively model documents. However, our emphasis lies in the diversification task, specifically targeting the modeling of document diversity at the passage level.

3 OUR PROPOSED METHOD: PAD

Search result diversification methods aim to identify diverse documents that address various query intents. However, most existing methods [23, 34, 40, 48, 57] treat a document as a whole. In this way, contents that cover different subtopics within a document will disturb each other. Therefore, it is hard for the subsequent ranking models to detect the subtopic information contained in the document representations.

In this article, we propose a passage-aware search result diversification approach PAD, which provides a passage view of the documents for the diversified ranking models and generates the context-aware passage representations according to the dynamic selected document list during the diversification process.

Specifically, we split the documents into lists of passages and train a classifier to infer the relevance of each passage to the query. With the help of the classifier, we are able to filter out the irrelevant passages and obtain top-k important passages to represent each candidate document. Furthermore, we model the interactions of intra-document passages and inter-document passages based on the attention mechanism during the greedy selection procedure.

3.1 Problem Formulation

The symbols and their explanations in this article are provided in Table 2. Supposing $\mathcal{D}$ ($|\mathcal{D}| = n$) is the initial document set retrieved by a query q ($q \in Q$), since search result diversification is NP-hard [1], most diversification approaches can be described as a sequential selection. That is, choosing the most diverse document d_i from the document set $\mathcal{D}$ according to the ranking score $f(d_i)$ continuously. To avoid redundancy and improve the diversity of the final document ranking $\mathcal{R}$, diversified ranking models should not only consider the document set $\mathcal{D}$ but also consider the

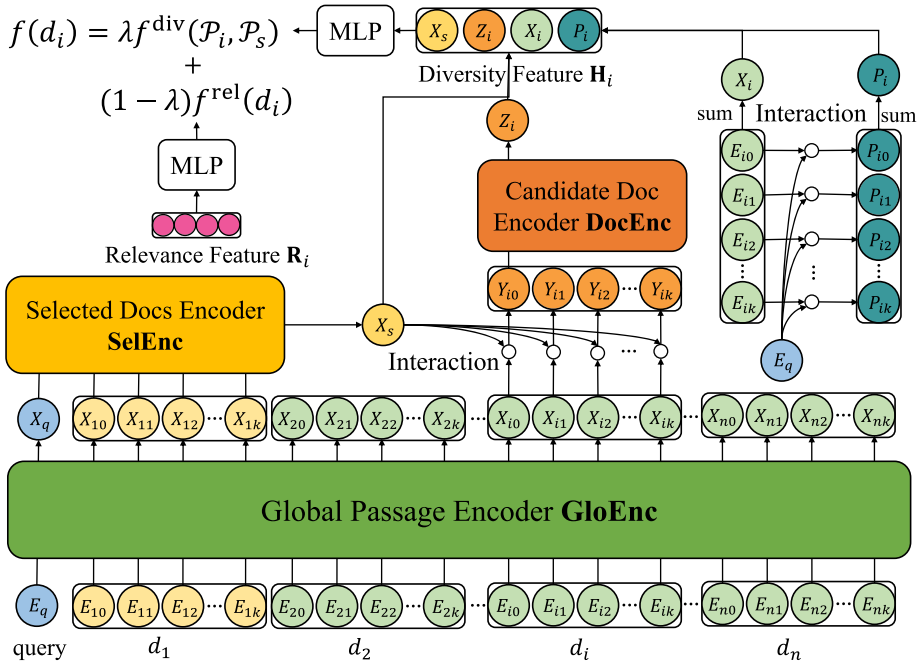

Fig. 2. The Architecture of PAD. Supposing document d_1 is the selected document (in yellow) at the current state, the diversity feature $\mathbf{H}_i$ of candidate document (in green) d_i is generated based on passage interactions. The interaction operation "$\circ$" stands for the element-wise product of two vectors.

selected document sequence $\mathcal{S}$. Hence, the scoring function f can be formulated as $f(q, d_i, \mathcal{S})$. Note that $\mathcal{S} = \emptyset$ at the first step. Therefore, a search result diversification task is to figure out the function f that can well reflect the novelty of the document d considering the selected sequence $\mathcal{S}$ at each step. Different from previous methods, we model document novelty and result diversity at the passage level.

Formally, considering a document set $\mathcal{D}$ retrieved by a query q, PAD selects the most essential passage set $\mathcal{P}_i = \{p_{i1}, \ldots, p_{ik}\}$ from a passage set $\{p_{(i,1)}, \ldots, p_{(i,t)}\}$ for each document d_i using a passage classifier. At the step t, we can measures the diversity and relevance of each document d_i in the candidate set $\mathcal{C}$, considering the passages set $\mathcal{P}_j$ of each selected document d_j from the selected passages $\mathcal{P}_s$ ($\mathcal{P}_j \subset \mathcal{P}_s$). The list of passages from the selected document sequence $\mathcal{S}$ is $\mathcal{P}_s = \bigcup_j \mathcal{P}_j$ for $d_j \in \mathcal{S}$. The scoring function f could be described as $f(q, \mathcal{P}_i, \mathcal{P}_s)$, which can be derived from both the diversity score $f^{\text{div}}(\mathcal{P}_i, \mathcal{P}_s)$ and the relevance score $f^{\text{rel}}(q, d_i)$.

3.2 The Architecture of PAD

As shown in Figure 2, PAD comprises three main modules: a GloEnc, a SelEnc), and a candidate DocEnc. The global passage encoder performs the global interactions of all the passages and the query, while the selected documents encoder generates the representation of selected passages at each greedy selection step. The candidate document encoder will finally aggregate the information within each candidate document based on the attention mechanism.

Given that document diversity is evaluated based on the relevance [10], PAD incorporates both diversity score and relevance score for final ranking as well as most research [23, 27, 34, 38, 40, 47, 48, 57]. Formally, the ranking score of the document d_i is calculated as the sum of diversity score

$f^{\text{div}}(q, d_i, \mathcal{S})$ and relevance score $f^{\text{rel}}(q, d_i)$ as follows:

$$f(q, d_i, \mathcal{S}) = \lambda f^{\text{div}}(q, d_i, \mathcal{S}) + (1 - \lambda) f^{\text{rel}}(q, d_i), \tag{2}$$

where λ and $(1 - \lambda)$ are the weights of diversity and relevance. Different from the most existing methods, our diversity score function $f^{\text{div}}(d_i, \mathcal{S}) = f^{\text{div}}(\mathcal{P}_i, \mathcal{P}_s)$ is derived from the selected documents sequence $\mathcal{S}$ at passage level.[1] For a fair comparison, we employ the same relevance features $\mathbf{R}_i$ as the previous work [23, 27, 34, 40]. The relevance score $f^{\text{rel}}(d_i)$ is obtained from the relevance feature $\mathbf{R}_i$ that includes BM25 and TF-IDF,

$$f^{\text{rel}}(d_i) = \text{MLP}(\mathbf{R}_i). \tag{3}$$

The diversity score $f^{\text{div}}(\mathcal{P}_i, \mathcal{P}_s)$ is calculated from the passage-aware diversity feature $\mathbf{H}$,

$$f^{\text{div}}(\mathcal{P}_i, \mathcal{P}_s) = \text{MLP}\left(\mathbf{H}_i\right), \tag{4}$$

where the document d_i's diversity feature $\mathbf{H}_i = [\mathbf{X}_s; \mathbf{Z}_i; \mathbf{X}_i; \mathbf{P}_i]$ is generated from several passage-level features. The diversity features $\mathbf{H}_i$ contain the currently selected state $\mathbf{X}_s$, document-level representation $\mathbf{Z}_i$, features from the original passage representations $\mathbf{X}_i$, and features $\mathbf{P}_i$ from the passage representations after interacted with query q. At each time step t, $\mathbf{H}_i$ is dynamically changed according to the selected passage set $\mathcal{P}_s$ and the passage set $\mathcal{P}_i$ of document d_i. Note that we omit the notation t to reduce redundancy. To capture the dynamic satisfaction degree concerning the multiple query intents and select the next novel document, the selected state representation $\mathbf{X}_s$ and the document representation $\mathbf{Z}_i$ are dynamically captured by PAD (more details in Section 3.5).

The key components of our diversified scoring process are briefly introduced as follows:

(1) *Passage Selection.* Considering that not all the passages are relevant and the irrelevant passages can influence the novelty measure of documents, it is necessary to distinguish the relevant passages from the irrelevant ones in the search result diversification task. Therefore, we develop a passage classifier to judge which passage is relevant to the given query q. For each passage, we use the concatenation of query and passage as the input of the passage classifier. For a long document d_i that has t passages in total, we can obtain top-k essential passages for the downstream diversity ranking model, which is helpful to reduce noise brought by irrelevant content (illustrated in Section 3.3).

(2) *Passage-aware Diversification.* Since we use the passage-level document representations in this article, we also devise a corresponding passage-aware diversification framework that can automatically compare different passages and aggregate passage-level similarity to document-level diversity features. To accomplish this, we leverage the attention mechanism to automatically learn the diversity features of the document d_i. More specifically, the global passage encoder will model the relationship of all passages, while the selected documents encoder is expected to capture the dynamic state of the query considering that the information needs will be partially satisfied by the selected passages. The candidate document encoder will generate the context-aware document presentations by considering all intra-document passages and the current state of the query (elaborated in Section 3.4).

3.3 Passage Selection

As discussed in Section 1, passages can offer a more fine-grained view for the model to figure out the contents that answer different query intents. From our perspective, it is unnecessary to offer the user documents that are irrelevant to the given query. Therefore, search result diversification

[1]We omit the query q in all equations for notation convenience.

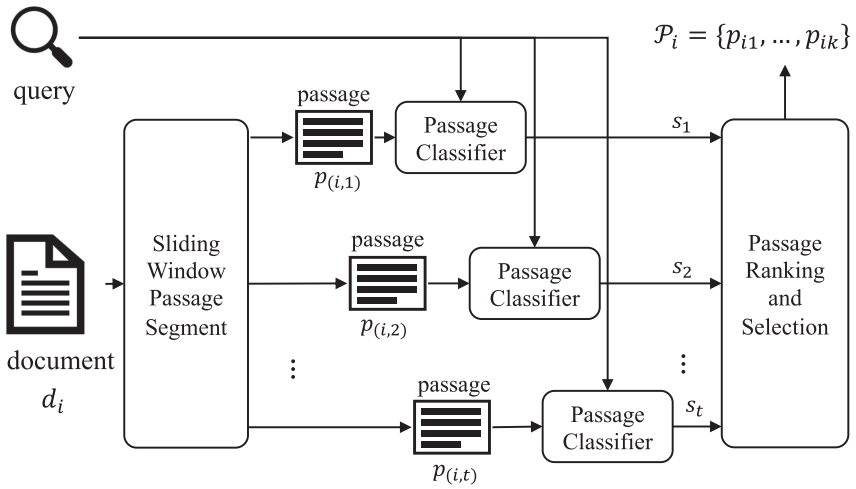

Fig. 3. The selection process of the passages.

should be considered under the relevance restriction. Similarly, irrelevant passages will bring more noise to confuse the diversified ranking model. For the convenience of denoising and capturing the real query intents contained in the passages, we select the most essential and representative passages for diversified ranking. We will introduce the process of passage segment, passage classifier, passage representation, and passage evaluations next.

Passage Segment. The passage segment and selection procedure are shown in Figure 3. For the convenience of processing and the input limitation of the document encoder, we leverage a fixed-length sliding window to split the document into several passages.[2] Supposing the total token number of the input document d_i is $|d_i|$, the window size is w, and the overlap between two neighbor passages in the document is o ($o < w$), then the total passage number $t = 1 + \lceil \frac{|d_i|-w}{w-o} \rceil$. After the passage segment, the documents are split into several passages with the same length of w (short ones will be padded). Hence, we can obtain the initial passage set $\{p_{(i,1)}, \ldots, p_{(i,t)}\}$ to present the document d_i. Since the whole passage set have irrelevant passages, the passages will be further evaluated in the next steps.

Passage Classifier. Considering the powerful semantic understanding capability of pre-trained language models (e.g., BERT [17]), we develop our passage classifier based on the popular pre-trained model BERT. Given that the diversity ranking datasets lack passage-level relevance annotations, we turn to training our passage relevance classifier on external resources. Specifically, we fine-tune BERT on the MS MARCO passage ranking dataset to obtain an effective passage classifier. Consistent with the previous studies [14, 45], we use the concatenation of the query and the passage as the input of the classifier and leverage the representation of "[CLS]" token for classification. Supposing the token sequence of each passage p is $[t_1, \ldots, t_w]$, the token sequence of query q is $[q_1, \ldots, q_m]$, and the input of the passage classifier is the concatenation of query and passage tokens, namely $[[\text{CLS}], q_1, \ldots, q_m, [\text{SEP}], t_1, \ldots, t_w, [\text{SEP}]]$. The relevance score s_i of the passage $p_{(i,c)}$ is calculated as follows:

$$s_c = \sigma\left(\text{MLP}\left(\underset{[\text{CLS}]}{\text{BERT}}(q, p_{(i,c)})\right)\right), \tag{5}$$

[2]The experiments will show that this passage splitting method is simple yet effective. Note that our method is compatible with other advanced passage splitting algorithms. We leave this exploration in our future work.

where σ is the non-linear activation function (e.g., sigmoid) and the output score s_i is a non-negative real number that scales from 0 to 1. The relevance score s_i is derived from the vector of the [CLS] token in the input sequence. The passage $p_{(i,c)}$ trends to be irrelevant to the query q when the s_c is closer to zero, while passage $p_{(i,c)}$ is more relevant to the query if the score is closer to one.

Passage Representation. Similarly to the Passage Classifier, the representation of the passage p_i is obtained from the BERT. The representation $\mathbf{E}_i$ of the passage $p_i = [t_{i1}, \ldots, t_{iw}]$ is the "[CLS]" token representation of the BERT with the input sequence of the passage p_i as follows:

$$\mathbf{E}_i = \underset{[\text{CLS}]}{\text{BERT}}([\text{CLS}]; t_{i1}; \ldots; t_{iw}; [\text{SEP}]), \tag{6}$$

where the input length of the passage p_i is smaller than the max input limitation of BERT (512) to avoid information loss. With the encoder BERT, the semantic features of the passage p_i are contained in the vector $\mathbf{E}_i$, which will be further used by our passage-aware diversity ranking framework PAD.

Passage Evaluation. The input of the passage selection procedure are the initial passages set $\{p_{(i,1)}, \ldots, p_{(i,t)}\}$ of the document d_i and the passage number k, and the output is the set of the most essential k passages $\mathcal{P}_i = \{p_{i1}, \ldots, p_{ik}\}$. Note that we use all the passages for the short documents that $t < k$, namely $\mathcal{P}_i = \{p_{(i,1)}, \ldots, p_{(i,t)}\}$. The set $\mathcal{P}_i$ is the passage set used by the diversity ranking model. Considering that the passage classifier judges the relevance of passage p_i at the semantic view, we also adopt traditional retrieval metrics (e.g., BM25) to select the most essential passages. Specifically, we can derive the relevance score $\{s_1, \ldots, s_t\}$ of each query and passage pair from the passage classifier. The BM25 score of each passage can be calculated as $\{b_1, \ldots, b_t\}$. Specifically, the BM25 score used for selecting passages is normalized via the maximal BM25 score within the same query. Therefore, both the classifier score and BM25 score scale from zero to one. In our implementation, we use the sum of the scores from the passage classifier and BM25 as the final passage selection scores. The output of the passage selection module is the passage set $\mathcal{P}_i = \{p_{i1}, \ldots, p_{ik}\}$ used by the following diversity ranking module.

3.4 Passage-aware Diversification

Modeling the document's relationship is a crucial task in search result diversification. As introduced in Section 3.2, we model the document's diversity at the passage level. Therefore, how to derive the document relationship from the passages is our focus in this section. There are several differences from the existing document-level diversification approaches: (1) Some passages within the same document may cover different query intents, (2) the satisfaction degree of query information needs can be sensed at the passage level, and (3) the document's novelty is determined by its passages' novelty considering the selected passage list.

To accommodate these query-specific characteristics, we designed a passage interaction framework that can (1) automatically aggregate the passage information to form the document representation, (2) capture the multiple intents that have been covered by the passages of the selected document sequence, and (3) flexibly encode passage's novelty via context-aware passage features. We implement these functions with three components of PAD: the GloEnc, SelEnc, and candidate DocEnc. They will be introduced as follows.

Supposing the number of the initial document set $\mathcal{D}$ is n for the given query q, we can obtain top-k essential passages $\mathcal{P}_i = \{p_{i1}, \ldots, p_{ik}\}$ for each document $d_i \in \mathcal{D}$ after the passage selection. The input of the PAD is $\mathbf{E} = [\mathbf{E}_q, \mathbf{E}_{10}, \ldots, \mathbf{E}_{nk}] \in \mathbb{R}^{S \times D}$, which includes the initial query representation $\mathbf{E}_q$ and passages representations $[\mathbf{E}_{i0}, \mathbf{E}_{i1}, \ldots, \mathbf{E}_{ik}]$ for each document d_i. Hence, $S = 1 + n \times (k + 1)$ is the input sequence length and D is the dimension of the distributed representations of

passages and query. Note that we add a document-specific representation $\mathbf{E}_{i0}$ for each document $d_i \in \mathcal{D}$ from random initialization to identify the document. We apply Transformer Encoder [42] to model the interactions of the passages. Besides, we also add position embeddings and segment embeddings for the input sequence. GloEnc, SelEnc, and the candidate DocEnc are several layers of the InterLayer$(\cdot)$. The interaction layer InterLayer$(\cdot)$ is intended for the model to automatically sense the diversity features from the passage representations. The layer numbers of these three encoders are L_g, L_s, and L_d, respectively.

(1) **Global Passage Encoder.** The global passage encoder GloEnc is designed to conduct global interactions of all passages. Given the initial input $\mathbf{E} = [\mathbf{E}_q, \mathbf{E}_{10}, \ldots, \mathbf{E}_{nk}]$, the representations are updated by GloEnc as follows:

$$\mathbf{X} = \text{GloEnc}(\mathbf{E}), \tag{7}$$

where $\mathbf{X} = [\mathbf{X}_q, \mathbf{X}_{10}, \ldots, \mathbf{X}_{nk}]$ is the updated representations of the query and passages after L_g layers of the InterLayer$(\cdot)$. With the global encoder GloEnc, the passages can have interactions with the query q and passages from other documents, which will provide a global view of all documents' diversity at the passage level. Concretely, the GloEnc is a stack of L_g passage interaction layers as follows:

$$\text{GloEnc}(\mathbf{E}) = \text{InterLayer}_G^{L_g}(\cdots \text{InterLayer}_G^1(\mathbf{E})), \tag{8}$$

where the interaction layer InterLayer$(\cdot)$ is implemented via attention mechanism. The details of the passage interaction calculations will be described later in this section. It is worth noting that we add segment embeddings and position embeddings to distinguish passages from different documents and different positions. Concretely, for passage p_{ic}, the document d_i's cth passage, we add the document-specific segment embedding $[\text{D}_i]$ and position-specific embeddings $[\text{Pos}_c]$ to the original BERT representation of passage p_{ic} and get the input representation $\mathbf{E}_{ic}$. The segment embeddings and position embeddings are generated via random initialization, which is used to identify different passages.

(2) **Selected Document Encoder.** The selected document encoder is expected to capture the states of the query intents covered by the selected documents. Hence, the input $\mathbf{X}_{sp}$ of the SelEnc is part of the updated representation $\mathbf{X}$, containing $\mathbf{X}_q$, and all the passages representations $\mathbf{X}_{jt}, t \in [0, \ldots, k], d_j \in \mathcal{S}$. The query and the passages of the selected documents are used to generate the current state representation $\mathbf{X}_s$ via L_s layers InterLayer,

$$\mathbf{X}_s = \underset{q}{\text{SelEnc}}(\mathbf{X}_{sp}), \tag{9}$$

where $\mathbf{X}_s$ is the updated query representation of $\mathbf{X}_q$ generated by the SelEnc. Similarly to GloEnc, the SelEnc is implemented with several interaction layers with an attention mechanism. Specifically, the SelEnc consists of L_s layers of InterLayer,

$$\text{SelEnc}(\mathbf{X}_{sp}) = \text{InterLayer}_S^{L_s}(\cdots \text{InterLayer}_S^1(\mathbf{X}_{sp})), \tag{10}$$

After L_s layer interactions, the passages information contained by the selected document sequence $\mathcal{S}$ is sensed and absorbed by the query representation $\mathbf{X}_s$, which is used to represent the current selection state in the iterative selection. Because the selection state features $\mathbf{X}_s$ are generated from the selected document sequence $\mathcal{S}$, the subtopic information contained by the passages from $\mathcal{S}$ will be encoded in the selection state $\mathbf{X}_s$, while the representations of the passages that cover less-satisfied subtopics will be more different from the state $\mathbf{X}_s$.

Because we hope to obtain the context-aware passage representations, we implement the interactions between the selected state $\mathbf{X}_s$ and each candidate passage representation $\mathbf{X}_{ic}$ of p_{ic}

($p_{ic} \in \mathcal{P}_i, d_i \in C$) with an element-wise product operation. They are calculated as follows:

$$\mathbf{Y}_{ic} = \mathbf{X}_s \circ \mathbf{X}_{ic}, \tag{11}$$

where $\circ$ stands for the element-wise product of two vectors, $c \in [0, \ldots, k]$. Then the passages representation $\mathbf{Y}_{ic}$ will contain the information from the current state $\mathbf{X}_s$ at each selection step. After the interactions with the current selection state $\mathbf{X}_s$, the passage representations of the document d_i are updated to $\mathbf{Y}_i = [\mathbf{Y}_{i0}, \mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{ik}]$ from $\mathbf{X}_i = [\mathbf{X}_{i0}, \mathbf{X}_{i1}, \ldots, \mathbf{X}_{ik}]$. Compared with $\mathbf{X}_i$, the representations $\mathbf{Y}_i$ are context-aware passage-level document diversity features, which will be further used to generate document-level diversity features by the candidate DocEnc.

(3) **Candidate Document Encoder.** With the context-aware passage representations $\mathbf{Y}_i$ of the candidate document d_i, we can evaluate the novelty of the candidate documents at the passage level. Considering that some passages will become more redundant with respect to the selected document sequence $\mathcal{S}$, we leverage a candidate DocEnc to automatically aggregate context-aware information from k passage representations $[\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{ik}]$ for the candidate document d_i. The document-level diversity features Z_i of the document d_i is calculated as follows:

$$\mathbf{Z}_i = \underset{d_i}{\text{DocEnc}}(\mathbf{Y}_i), \tag{12}$$

where $\mathbf{Y}_i = [\mathbf{Y}_{i0}, \mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{ik}]$, $\mathbf{Z}_i$ is the representation of document d_i's identifier, namely the updated representation of $\mathbf{Y}_{i0}$ by the DocEnc. The task of the DocEnc is to fuse the most novel information contained by the k passages of the candidate document d_i. Different from directly adding up all the passage representations, we hope the DocEnc can focus more on the novel passages in the search result diversification task. Therefore, we do not adopt the linear combination of all the passages to represent the documents. Similarly with the previous two encoders, the DocEnc leverages the self-attention mechanism to automatically discover the novel passages within the candidate document as follows:

$$\text{DocEnc}(\mathbf{Y}_i) = \text{InterLayer}_D^{L_d}(\cdots \text{InterLayer}_D^1(\mathbf{Y}_i)), \tag{13}$$

where the candidate DocEnc consists of L_d interaction layers. After the interactions of DocEnc, the document-level diversity features $\mathbf{Z}_i$ are encoded with the representative information from the k relevant passages of the document d_i concerning the selection states $\mathbf{X}_s$.

(4) **Interaction Layers.** The interaction layers adopted by the GloEnc, SelEnc, and DocEnc are $\text{InterLayer}_G(\cdot)$, $\text{InterLayer}_S(\cdot)$, and $\text{InterLayer}_D(\cdot)$, respectively. These interaction layers are similar to the Transformer encoder layers, which are based on a self-attention mechanism. The subscripts G, S, and D of these three types of layers are used to identify different encoders. In other words, we do not share the parameters of these three interaction layers. Given the ith InterLayer output $\mathbf{O}^{(i)}$, we can obtain the $(i+1)$-th layer output $\mathbf{O}^{(i+1)}$ from the $(i+1)$-th interaction layer as follows:

$$\mathbf{O}^{(i+1)} = \text{InterLayer}^{i+1}(\mathbf{O}^{(i)}), \tag{14}$$

where the input of the first global interaction layer is the original passage representations, namely $\mathbf{O}^{(0)} = \mathbf{E}$. The interaction layers are implemented with multi-head self-attention mechanism and they can be calculated as follows:

$$\mathbf{O}^{(i+1)} = \text{LayerNorm}(\mathbf{U}^{(i)} + \text{FFN}(\mathbf{U}^{(i)})), \tag{15}$$

$$\mathbf{U}^{(i)} = \text{LayerNorm}(\mathbf{O}^{(i)} + \text{MultiHead}(\mathbf{O}^{(i)})), \tag{16}$$

where $\text{FFN}(\cdot)$ is a fully connected feed-forward network with activation function (i.e., ReLU), LayerNorm stands for the layer normalization operation [4]. Since $\mathbf{Q} = \mathbf{K} = \mathbf{V}$ in the self-attention,

the multi-head self-attention function MultiHead($\mathbf{Q}, \mathbf{K}, \mathbf{V}$) is denoted as MultiHead($\mathbf{V}$) for convenience. Given the head number h, the output of MultiHead($\mathbf{V}$) is the concatenation of each head's output $\mathbf{a}_j, j \in [1, h]$,

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{a}_1; \ldots; \mathbf{a}_h], \tag{17}$$

where $\mathbf{a}_j = \text{Attn}(\mathbf{Q}\mathbf{W}_j^Q, \mathbf{K}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V)$, $\mathbf{W}_j^Q, \mathbf{W}_j^K$, and $\mathbf{W}_j^V$ are all trainable parameters. The self-attention used in this article is the scaled dot-product attention function,

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}}{\sqrt{d_k}})\mathbf{V}. \tag{18}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are the query, key and value matrices and d_k is the feature dimension of the input $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$.

With the self-attention mechanism, the output passage representations will be updated with the sequence information. Hence, for the three encoders GloEnc, SelEnc, and DocEnc, passage representations will be encouraged to focus on the whole document sequence $\mathcal{D}$, the selected passage sequence $\mathcal{P}_s$, and the candidate document d_i's passage sequence $\mathcal{P}_i$. Therefore, we can generate the diversity features of each candidate document d_i from the passage representations with the global view, the selected document view, and the candidate document view.

3.5 Diversity and Relevance Features

As shown in Equation (2), the final ranking score of the document d_i consists of the relevance score $f^{\text{rel}}(q, d_i)$ and diversity score $f^{\text{div}}(q, d_i, \mathcal{S})$. The diversity features $\mathbf{H}_i$ and the relevance features $\mathbf{R}_i$ of the document d_i used for generating these scores are introduced as follows.

(1) **Diversity Features.** The diversity feature $\mathbf{H}_i$ for each candidate document d_i is the concatenation of the features from the context-aware passage interactions shown in Figure 2. More specifically, $\mathbf{H}_i = [\mathbf{X}_s; \mathbf{Z}_i; \mathbf{X}_i; \mathbf{P}_i]$.

$\mathbf{X}_s$: The current state presentation at each selection step. Considering that some query intents (or subtopics) are partly satisfied when some documents are selected, the query information needs to change dynamically during the selection procedure. Therefore, it is necessary to represent the current status of query satisfaction when considering document diversity. $\mathbf{X}_s$ is generated by the SelEnc, which considers all the passages contained by the selected documents. $\mathbf{X}_s$ helps capture the passage-level dynamic needs of the query during the selection procedure.

$\mathbf{Z}_i$: The document representation from the candidate DocEnc after interacting with the selected status $\mathbf{X}_s$. $\mathbf{Z}_i$ is automatically generated via PAD with the self-attention mechanism, which contains the information aggregated from the essential passages belonging to document d_i. Given that $\mathbf{Z}_i = \text{DocEnc}(\mathbf{Y}_i)$ and $\mathbf{Y}_i$ is the context-aware representations of the document d_i's passages, $\mathbf{Z}_i$ is a dynamic representative novelty feature of the document d_i according to the selected document sequence $\mathcal{S}$.

$\mathbf{X}_i$: The origin representation of the document d_i from the passage set $\mathcal{P}_i$. Considering that $\mathcal{P}_i$ is the essential passage set concerning query q, it is necessary to take their representations into account. Specially, we adopt $\mathbf{X}_i = \sum_{c=0}^{k} \mathbf{E}_{ic}$, where $\mathbf{E}_{ic}$ is the initial representation of passage $p_{ic} \in \mathcal{P}_i$. Compared with $\mathbf{Z}_i$, $\mathbf{X}_i$ offers an original passage aspect for the ranking model.

$\mathbf{P}_i$: The interaction representation of document d_i. Given that we need to focus on the relevance of the passages in the search result diversification task, we use the element-wise product to implement the interactions of query and passages. Specifically, $\mathbf{P}_i = \sum_{c=0}^{k} \mathbf{P}_{ic}$, where $\mathbf{P}_{ic} = \mathbf{E}_{ic} \circ \mathbf{E}_q$. The interaction representation $\mathbf{P}_i$ of the document d_i is a complement to the document diversity features at the passage level.

(2) **Relevance Features.** As illustrated in Equation (3), we use the relevance feature $\mathbf{R}_i$ to derive the relevance score of the document d_i. Consistent with the previous work [23, 27, 34, 40], $\mathbf{R}_i$ is an 18-dimension traditional relevance feature, including BM25, PageRank, and TF-IDF. For a fair comparison, the relevance features are kept the same with DSSA.

3.6 Training and Optimization

Our approach involves two training processes: the training of the passage relevance classifier and the training of the diversification. Each process will be discussed in detail below.

(1) **Passage Classifier.** The training pair of the passage classifier is generated based on the binary label of the relevance judgment. For example, given the query q, passage p, and a relevance label $y_{(q,p)}$ of pair (q, p), we can derive the training sample $(q, p, 1)$ for positive sample, while $(q, p, 0)$ stands for a negative one. The passage classifier is trained with a binary cross-entropy loss based on the training samples. Supposing the training sample set $\mathcal{T}$ has $|\mathcal{T}|$ (q, p) training pairs and the relevance label set of these pairs is $\mathcal{Y}$ ($y_{(q,p)} \in \mathcal{Y}$). The training loss of the passage classifier can be derived as follows:

$$\mathcal{L}_{\mathrm{c}} = -\frac{1}{|\mathcal{T}|} \sum_{(p,q)\in\mathcal{T}} \left(y_{(q,p)} \times \log \frac{1}{1+e^{-r(p,q)}} + (1-y_{(q,p)}) \times \log \frac{e^{-r(q,p)}}{1+e^{-r(q,p)}} \right), \tag{19}$$

where $r(q, p)$ is the prediction score of the passage classifier. The classifier can be implemented via the pre-trained language models. For example, we can use BERT to model the relations of the query q and passage p as follows:

$$r(q,p) = \mathrm{MLP}(\underset{[\mathrm{CLS}]}{\mathrm{BERT}}([\mathrm{CLS}]; \mathrm{q}; [\mathrm{SEP}]; \mathrm{p}; [\mathrm{SEP}])), \tag{20}$$

where the outputs of the classifier are generated via an MLP layer with the input of "[CLS]" token representations. The input of the BERT classifier is the concatenation of the query q and passage p token sequence. The special token "[SEP]" is added to separate query tokens from the passage tokens.

(2) **Diversity Ranking.** The search result diversification can be considered with the greedy selection strategy. For example, given the query q ($q \in Q$), the corresponding document set $\mathcal{D}$, the current selected document sequence $\mathcal{S}$, and the remnant candidate document set C ($C = \mathcal{D} \backslash \mathcal{S}$), we can evaluate the diversity of the candidate documents concerning the selected document sequence $\mathcal{S}$. Specifically, leveraging the diversity metric function $M(\cdot)$ (e.g., ERR-IA [11]), we can select a positive document d^+ and a negative document d^- from the candidate document C, which means d^+ can bring more novel information to the search results than d^-. In other words, if we append the document d^+ to the sequence $\mathcal{S}$, then the ranking sequence $[\mathcal{S}, d^+]$ is more diverse than the sequence $[\mathcal{S}, d^-]$. Formally, we can measure the importance of the positive-negative sample with a weight w derived from the diversity metric function $M(\cdot)$ as follows:

$$w = |M([\mathcal{S}, d^+]) - M([\mathcal{S}, d^-])|, \tag{21}$$

As elaborated on in Equation (2), the ranking score of the document d_i is $f(q, d_i, \mathcal{S})$. The output scores of the training pair $(q, \mathcal{S}, d^+, d^-, w)$ are $f(q, d^+, \mathcal{S})$ and $f(q, d^-, \mathcal{S})$. The loss of the diversity ranking models can be calculated with list-pairwise loss [23],

$$\mathcal{L}_{\mathrm{div}}(q, \mathcal{S}, d^+, d^-, w) = -w \log \left(\frac{1}{1+e^{-(s^+ - s^-)}} \right). \tag{22}$$

where $s^+ = f(q, d^+, \mathcal{S})$ and $s^- = f(q, d^-, \mathcal{S})$ are the output scores of our diversity ranking model PAD. Supposing the training sample set of query q is O_q, the optimization of the model can be

calculated as follows:

$$f = \arg\min \sum_{q \in Q} \sum_{o \in O_q} \mathcal{L}_{\text{div}}(q, \mathcal{S}_o, d_o^+, d_o^-, w_o), \tag{23}$$

where $o = (q, \mathcal{S}_o, d_o^+, d_o^-, w_o)$ is a training sample of O_q and f is the document diversity scoring function of the PAD. With Equation (23), the parameters of the PAD are optimized through the generated training samples.

4 EXPERIMENTS

4.1 Dataset and Evaluation

The experiments are conducted on the widely used ClueWeb09 [8] dataset, which contains about 50 million web pages in English. The topic (query) sets come from TREC Web Track 2009 to 2012.[3] There are 200 queries in total, and 198 of them are used for search result diversification, because query #95 and query #100 do not have diversity judgments. The human-labeled subtopic number of these ambiguous queries ranges from 3 to 8, and the subtopics within a query have the same weights. We also found a passage-level dataset TREC Dynamic Domain Track [52] related to our method. However, for a fair comparison, we still conduct our experiments on ClueWeb09. We will consider more datasets for future exploration.

The evaluation metrics in this article have been widely used in previous methods [23, 27, 34, 40, 47, 48, 51, 57], including α-nDCG [12], ERR-IA [11], NRBP [13], and S-rec [56].

We adopt the metric α-nDCG to select the best model in our experiments. α-nDCG is a classical diversity ranking metric proposed by Clarke et al. [12]. Supposing the query q has m subtopics and n candidate documents related to these subtopics. The human-assessed relevance label of document d_i concerning subtopic t is $J(i,t)$, which can be 0 or 1. The α discounted cumulative gain (α-DCG) of the ranking sequence is calculated as follows:

$$\alpha\text{-DCG} = \sum_{i=1}^{n} \sum_{t=1}^{m} \frac{J(i,t)(1-\alpha)^{C(i,t)}}{\log_2(1+r_i)}, \tag{24}$$

where $\alpha \in (0,1]$ reflects the possibility of assessor error, r_i is the ranking position of the document d_i, and $C(i,t) = \sum_{j:r_j<r_i} J(j,t)$ is the number of documents that cover subtopic t and rank higher than document d_i. For the convenience of the comparison, we can normalize α-DCG by dividing the metric of the ideal ranking,

$$\alpha\text{-nDCG} = \frac{\alpha\text{-DCG}}{\alpha\text{-DCG}_{\text{ideal}}}. \tag{25}$$

We can obtain the result α-nDCG@k of the ranking list by only counting the top-k documents. It is worth noting that α-$\text{DCG}_{\text{ideal}}$ is the ideal metric of the entire ranking list rather than the top-k results.

4.2 Experiment Settings

For a fair comparison, the experiment settings are consistent with the previous studies [23, 27, 34, 40]. The diversity judgments of the documents are treated as binary. The initial document rankings are retrieved via the Lemur service.[4] The top 50 documents provided by Lemur are used for search result diversification, and all the metrics are derived from the top 20 documents in the result list output by the model. Fivefold cross-validation is used in the experiment, and the reported metrics are the average metrics of the five folds. The subtopics used by explicit methods, such as

[3] https://trec.nist.gov/data/web09.html
[4] Lemur service: http://boston.lti.cs.cmu.edu/Services/clueweb09_batch/

DSSA [23], DESA [34], and DVGAN [27], are the query suggestions from the commercial search engines, which are released by Hu et al. [22].

4.3 Baseline Approaches

To evaluate the effectiveness of our method, we compare PAD with four types of baseline methods: (1) non-diversified methods, (2) explicit methods, (3) ensemble methods, and (4) implicit methods.

Lemur and ListMLE. These are two ranking methods that do not consider the diversity of the documents. The results of the Lemur are generated by the Lemur service. Besides, ListMLE [46] is a listwise learning-to-rank approach.

xQuAD, HxQuAD, TxQuAD, PM2, HPM2, and TPM2. Theese are representative explicit heuristic methods. xQuAD [38] is a probability-based framework that measures the diversity from the subtopic coverage distribution of the documents, while PM2 [15] manipulates the diversified document ranking list according to the popularity of the subtopics. HxQuAD and HPM2 [22] are two methods that use hierarchically organized subtopic lists to improve xQuAD and HPM2, while TxQuAD and TPM2 [16] are term-level approaches based on xQuAD and PM2. DSSA [23] is a supervised explicit method that leverages RNNs to encode the selected document sequence during the greedy selection.

DESA and DVGAN. These are supervised ensemble approaches that utilize both the document features and subtopic features. DESA [34] leverages the subtopics and document representations to generate the ensemble features with a decoder. DVGAN [27] uses a generator to produce the explicit subtopic features and exploits a discriminator to learn document similarity features.

R-LTR, PAMM, DALETOR, NTN, and Graph4DIV. Theese are several competitive supervised implicit methods. R-LTR [57] is an LTR approach that leverages document similarity to model document diversity. PAMM [47] and DALETOR [51] are two supervised methods that propose approximate loss functions based on the metric functions. The neural tensor network [48] can be applied to R-LTR and PAMM, respectively. Graph4DIV [40] is an implicit method that models document relationships by a graph. We reproduced DALETOR based on its paper. More specifically, we adopt a document interaction network with a α-DCG loss with latent cross.

4.4 Implementation Details

Given that the ClueWeb09 dataset lacks passage-level relevance labels, we fine-tune the BERT [17] model on the MS MARCO [31] passage dataset to obtain the passage relevance classifier. The classifier is trained on the training dataset and tuned on the validation dataset of the MS MARCO dataset. Since the classifier is not trained or finetuned on the ClueWeb09 dataset, the results from the classifier are used by all the folds in cross-validation in the diversification, which will not lead to a data leakage problem.

To ensure a fair comparison, we have closely followed the experimental settings of prior works, including DSSA [23], DESA [34], DVGAN [27], and Graph4DIV [40]. In explicit methods, the subtopics we use are derived from Google query suggestions, which can be found in Reference [22]. Consistent with the previous approaches [23, 34, 35], we adopted the first-level subtopics. It is important to note that all subtopics are given equal weight, and the human-annotated relevance for each subtopic is treated as a binary value in our experiments. Given that our primary focus in this article is on modeling document diversity, we have utilized the widely accepted relevance features that were made available by Jiang et al. [23] on GitHub, which have also been adopted in previous research.[5] The baseline methods, such as DESA,[6] Graph4DIV,[7] are reproduced based on their

[5]https://github.com/jzbjyb/dssa
[6]https://github.com/qratosone/GDESA
[7]https://github.com/su-zhan/Graph4DIV

Table 3. Performance Comparison of All Approaches

Category	Method	α-nDCG	ERR-IA	NRBP	S-rec
Ad hoc	Lemur	.369*	.271*	.232*	.621*
	ListMLE	.387*	.287*	.249*	.619*
Explicit	xQuAD	.413*	.317*	.284*	.622*
	TxQuAD	.410*	.308*	.272*	.634*
	HxQuAD	.421*	.326*	.294*	.629*
	PM2	.411*	.306*	.267*	.643*
	TPM2	.399*	.291*	.250*	.639*
	HPM2	.420*	.317*	.279*	.645*
	DSSA	.456*	.356*	.326*	.649*
	DSSA (BERT)	.457*	.352*	.319*	.656*
Ensemble	DESA	.464*	.363*	.332*	.653*
	DVGAN	.465*	.367*	.334*	.660
Implicit	R-LTR	.403*	.303*	.267*	.631*
	PAMM	.411*	.309*	.271*	.643*
	R-LTR-NTN	.415*	.312*	.275*	.644*
	PAMM-NTN	.417*	.311*	.272*	.648*
	DALETOR	.397*	.305*	.271*	.607*
	Graph4DIV	.468	.370	.338	.666
	PAD (ours)	**.482**	**.386**	**.357**	**.670**

The best result is in bold. The symbol $\star$ indicates significant improvements obtained by PAD in t-test with p-value< 0.05.

public code repositories. The training samples are generated by the list-pairwise method proposed by Jiang et al. [23], which is widely used by many methods [34, 35, 40, 41]. Besides, our dataset division is also kept the same with these baseline methods for a fair comparison. The diversity metrics have been calculated for the top 20 documents within the initial sequence of 50 documents across the fivefold results.

The training batch size of the passage relevance classifier is 32, and we adopt the optimizer AdamW [29] with the warm-up mechanism. The layer numbers L_g, L_s, and L_d of the passage interaction layers are set at 1, 3, and 3 based on the experimental results. The head number of the multi-head attention is 8. Our model is selected based on the α-nDCG@20. At each fold, four subsets are used for training, and the rest is used for testing. The results are the average over the five folds. The parameter λ is tuned from $[0.1, \ldots, 0.9]$, and we use $\lambda = 0.5$ according to the results. Our diversity ranking model is trained with batch size 8, learning rate 1e-3, and dropout rate 0.5. The learning rate is tuned from 1e-6 to 1e-3. More details about our method can be found at https://github.com/su-zhan/PAD.

4.5 Experimental Results

The overall results of our model and the baseline methods are shown in Table 3. According to the results, PAD outperforms all the diversification baselines in terms of α-nDCG, ERR-IA, and NRBP, which clearly demonstrates the effectiveness of our model. We also have the following observations.

(1) PAD outperforms all implicit methods in terms of α-nDCG, ERR-IA, and NRBP. R-LTR-NTN and PAMM-NTN are two representative implicit methods that learn a document's novelty function automatically. The advantages acquired by PAD demonstrate that the attention mechanism is more suitable to capture the dynamic information needs of the query during the selection. DALETOR is a recent competitive method with a novel loss function that directly optimizes the diversity metrics.

However, it seems to be sensitive to the initial document ranking in our experiments. Compared to DALETOR, the large improvement gained by PAD shows the effectiveness of greedy selection and passage representations. The combination of passages and selected documents encoder can model the sequence of the selected documents more precisely, which helps select the novel passages at each step. Graph4DIV is a competitive implicit approach that models the relation of the documents based on the signals from a relation classifier. Although PAD does not acquire significant improvement over Graph4DIV, it still demonstrates superior performance across all evaluation metrics. It is important to highlight some notable details in this comparison. First, PAD outperforms Graph4DIV by a substantial margin, particularly with a 1.4% improvement in terms of alpha-nDCG@20. The p-value of the two-tailed t-test between the two methods is 0.05836, slightly exceeding the significance threshold of 0.05. Second, Graph4DIV relies on human judgment for document relations in the training period, which contributes to its impressive performance. In contrast, PAD does not depend on additional information like Graph4DIV, making it applicable in situations where subtopic annotations are unavailable. This indicates the benefits of leveraging multiple essential passages as document representation in search result diversification.

(2) PAD outperforms the explicit methods by a large margin. Concretely, PAD significantly outperforms the unsupervised explicit approaches, such as xQuAD, and PM2, which demonstrates the superiority of supervised methods. Compared with DSSA, PAD achieves better performance without using subtopics, showing the effectiveness of enhancing document representations with passages. We adopt BERT embeddings as passage representations on PAD for BERT is a recent popular encoder. To demonstrate the effects of BERT embeddings, we use them in DSSA and the results are reported in Table 3. The performance of DSSA with BERT embeddings is denoted as DSSA (BERT). Given that DSSA with BERT embeddings achieves similar performance compared with the original DSSA, PAD can still outperform DSSA with different document representations, which indicates that BERT embeddings do not appear to have a substantial impact on the final ranking performance. Furthermore, since obtaining the real search intents of the users is still a challenging task, how to mine the information from the documents could be a promising direction for implicit search result diversification approaches. Hence, fully utilizing the passages and their relationships may overcome the disadvantages of lacking subtopics.

(3) PAD can also outperform two ensemble methods. DESA is an ensemble framework that leverages both implicit (document) features and explicit (subtopic) features for direct diversity ranking. Different from the direct ranking adopted by DESA, the greedy selection of PAD acquires higher scores in terms of all metrics, which implies that capturing the information needs of the query is an essential part of search result diversification. DVGAN is the most advanced ensemble method that combines the features of a generator and a discriminator. Compared with DVGAN, the advantages of PAD lie in the more precise modeling of the document's content.

It is worth noting that the subtopics from a search engine are suboptimal to explicit/ensemble methods. However, offering the human-annotated search intents as subtopics to the explicit and ensemble models will result in a data leakage problem. All the methods, including implicit models and explicit models, cannot perceive the real subtopics in the inference period, which is reasonable in practice (mining the exact subtopics of a given query is a challenging task). Therefore, both implicit and explicit approaches are fairly compared. Both of them have their own advantages: implicit approaches do not depend on external subtopic resources like Google Suggestions, while leveraging better subtopics in explicit methods could potentially lead to higher performance.

Since the p-value of PAD and Graph4DIV is slightly more than 0.05, we add a further wins and losses analysis in the experiments. The comparison results are shown in Table 4. The experimental results support that PAD has advantages over Graph4DIV in more query cases, especially in faceted queries, which is consistent with the example shown in Figure 1.

Table 4. Wins and Loses Analysis of PAD Compared with Graph4DIV

Category	Win	Tie	Lose
All queries	94	20	84
Ambiguous queries	25	8	24
Faceted queries	69	12	60

Table 5. Performance of PAD with Different Settings

	α-nDCG	ERR-IA	NRBP	S-rec
PAD	**.482**	**.386**	**.357**	**.670**
w/o GloEnc	.471	.373	.342	.667
w/o $\mathbf{X}_i$	.462	.363	.332	.660
w/o $\mathbf{P}_i$	.462	.366	.333	.658
w/o $\mathbf{X}_s$	.466	.368	.337	.665
w/ Head Passages	.467	.371	.340	.660
w/ Random Passages	.471	.373	.341	.669
w/ doc2query	.468	.374	.345	.658

Best results are in bold.

4.6 Effects of Different Settings

To figure out the effects of different settings in our method, we conduct the ablation study by removing the components one by one from the entire model. The performance of the PAD with different settings is shown in Table 5.

(1) **Global Interactions in PAD.** The performance of PAD without a global passage encoder (denoted as *w/o* GloEnc) degrades across all metrics, which illustrates the importance of global interactions. Moreover, the results of PAD without GloEnc are still superior to most baselines in Table 3. It demonstrates that modeling the relationship between the selected and candidate documents is an essential part of search result diversification. Meanwhile, the good results show that PAD is robust as the SelEnc and candidate DocEnc can still work well even without the GloEnc.

(2) **Ablation of diversity features.** Since the diversity feature $\mathbf{H}_i$ of Section 3.5 is the concatenation of multiple representations, it is necessary to investigate the effects of these features. Given that $\mathbf{H}_i = [\mathbf{X}_s; \mathbf{Z}_i; \mathbf{X}_i; \mathbf{P}_i]$, $\mathbf{Z}_i$ is the fundamental representation for identifying document d_i in the selection process, we only remove the other three features: $\mathbf{X}_i$ (*w/o* $\mathbf{X}_i$), $\mathbf{P}_i$ (*w/o* $\mathbf{P}_i$), and $\mathbf{X}_s$ (*w/o* $\mathbf{X}_s$). All the metrics decline when any feature is eliminated, validating the usefulness of these features in PAD. Compared with $\mathbf{Z}_i$, $\mathbf{X}_i$ and $\mathbf{P}_i$ contain passage information of document d_i with only local interactions (as shown in Figure 2). Together with $\mathbf{Z}_i$, $\mathbf{X}_i$, and $\mathbf{P}_i$, our method is able to sense both local and global features at passage level. In Figure 2, passage representations of documents are first processed by a GloEnc in our method PAD. The GloEnc will provide a global view of each passage within all passages from different documents. And the passage vectors will interact with other passages from other documents.

(3) **Passage Selection.** Given that k passages are used to represent a document in the diversity ranking process, the selection of the passages will also impact the experimental results. Therefore, we compare the results of the PAD with different passage selection strategies. Considering that the positions of the passages may contain additional information, we select k passages at the beginning of the documents (denoted as *w/* Head Passage). On the contrary, we randomly sample k passages from inside the documents (denoted as *w/* Random Passage) for comparison. According to the results, these two strategies are incapable of achieving the same level of performance as

Table 6. Effects of Different Passage Segment and Selection

Category	Parameters	α-nDCG	ERR-IA	NRBP	S-rec
Origin	(5, 256, 16)	**.482**	**.386**	**.357**	**.670**
First Passage	(1, 128, -)	.432	.327	.287	.665
	(1, 256, -)	.440	.340	.304	.655
	(1, 512, -)	.444	.343	.308	.659
Window Size	(5, 128, 16)	.473	.376	.345	.662
	(5, 512, 16)	.476	.379	.349	.669
Overlap	(5, 256, 0)	.475	.380	.349	.668

the original selection strategy (illustrated in Section 3.3), which indicates the superiority of our selection strategy. Specifically, compared with the head-passages selection strategy, the random selection strategy gets better metrics. From our perspective, content that covers different subtopics may scatter in different parts of the document. Therefore, the random selection strategy has more chances to obtain passages that answer different query intents. It also implies the necessity of selecting representative passages for diversity ranking.

Notably, the accuracy and F_1 of the passage classifier are 0.749 and 0.736, respectively. Since the passage-level relevance is absent in the ClueWeb dataset, the results of the passage classifier are evaluated on the validation set of MS MARCO. Given the large space of the classifier's performance, we expect that a better classifier can further enhance our PAD.

(4) **Passage Extension.** Apart from using the original passage content from the documents, we also explore the effects of PAD with the passage extension methods like doc2query [19]. Different from the traditional methods, doc2query initially employs the document content to predict potential queries, and then these predicted queries are appended to the documents before ranking. Similarly, we exploit docT5query to generate the queries for each passage used in PAD and append them to the passages. The docT5query is an excellent approach that uses T5 [36] to generate predicted queries. For the convenience of comparison, we use the T5 model released on Github in our experiments.[8] The results of PAD with predicted queries are denoted as *w/* doc2query in Table 5. Although equipped with a powerful generated model like T5, the results are not as good as the original PAD. One possible explanation for this is that the doc2query models are not fine-tuned for the ClueWeb dataset, and they may not distinguish subtle differences between various subtopics, potentially introducing more noise to the passage content.

5 DISCUSSION

In this section, we further investigate the effects of different passage segment approaches and different passage numbers in search result diversification. We focus on these questions: (1) What is the influence of different passage segments and representations? and (2) How many passages are necessary to represent a document?

5.1 Effect of Passage Segment and Selection

The experimental results of the PAD with different passage segments and selection are shown in Table 6. We mainly focus on the *passage number* k, *segment window size* w, and *overlap* o of two neighbor passages. The experimental results reported in Table 3 are conducted with parameters $(k, w, o) = (5, 256, 16)$.

(1) **Effect of the First Passage.** Different from most search result diversification methods, our PAD models the document's diversity based on multiple passages. To demonstrate the ben-

[8] https://github.com/terrierteam/pyterrier_doc2query

efits brought by passage-level modeling, we compare PAD with the single representation baseline models. As shown in Table 6, we use the first passage to represent the document with the window size as 128, 256, and 512 (denoted as (1,128,−), (1,256,−), and (1,512,−), respectively). PAD outperforms the single passage baseline models by a large margin, which validates the effectiveness of leveraging multiple passages to model document diversity. Furthermore, the baseline model with the parameter $(1, 512, -)$ is slightly better than the other two in terms of α-nDCG, ERR-IA, and NRBP, which may result from more sufficient information brought by the larger passage size. This could also account for the improvement brought by the multiple passages.

(2) **Influence of the Window Size.** To figure out the effect of the window size w of the passages, we fix the passage number k and overlap o and tune the window size from 128 to 512. As shown in Table 6, the models with a window size of 256 and 512 perform better than the one with a window size of 128. A possible reason is that more content is used to generate passage representations. Even with different passage sizes, PAD can still outperform the strongest baseline Graph4DIV in terms of α-nDCG, ERR-IA, and NRBP, which demonstrates the robustness of our model. Interestingly, PAD with passage size $w = 256$ is slightly higher than that with $w = 512$. The potential reason is that a large passage may cover more than one subtopic, which supports our assumption that different passages may cover different subtopics. Therefore, selecting a suitable passage size leads to a better document representation in search result diversification.

(3) **Effect of the Overlap.** Given that finding the complete passage structure within the document could be a more complicated task, we simply divide the documents into equal-length passages. To avoid that some essential token sequences may be split, we set an overlap parameter o to control the overlap length of two neighbor passages. The experimental results decline with the parameter $(5, 256, 0)$, which shows the effects of the overlap mechanism.

We conduct our experiments in the same dataset with many previous baseline methods, such as DSSA, DVGAN, and DESA. We agree that it is a good idea to use natural paragraphs. However, the baseline methods remove the HTML tags of the web pages and treat the content as a whole. To keep the same experiment settings with them, we follow the same pre-processing procedure. We understand that HTML tags could offer extra information to segment passages. However, the actual situations are much more complicated, because different web pages are organized in different styles. Therefore, considering the convenience and universal usage, we adopt the current passage segment method in PAD. Moreover, a more complicated passage segment method is also compatible with our method. Besides, the performance of PAD is expected to be better with natural paragraphs, which can be a promising direction for future study.

5.2 Effect of Passage Number

Since we leverage multiple passages to represent the documents in search result diversification, the passage number is an essential factor in the diversity ranking process. We tune the passage number k from 1 to 10 under the same settings. In general, the model's performance with more passages is better than the model with fewer passages according to the results. Specifically, the results of the baseline model with parameter $(1, 256, 16)$ are much lower than the original one, which demonstrates the potential drawback of using insufficient document representations.

The distributions of the subtopic number and passage number are shown in Figure 4(a) and Figure 4(b), respectively. Since the subtopics contained in the documents are much fewer than the passages of the documents. It is necessary to distinguish which passage is relevant to the query. Note that the subtopic in Figure 4(a) is the real search intents labeled by humans, different from the subtopics (e.g., Google Suggestions) used by explicit search result diversification models. Moreover, the document number with three passages is the highest in Figure 4(b), and the majority of the documents have more than three passages. Hence, the model cannot cover the most content with a

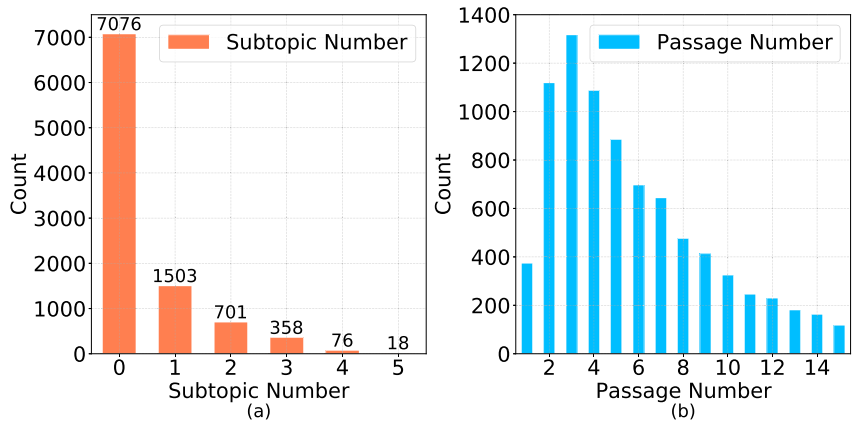

Fig. 4. The number distributions of subtopics and passages are shown in (a) and (b), respectively.

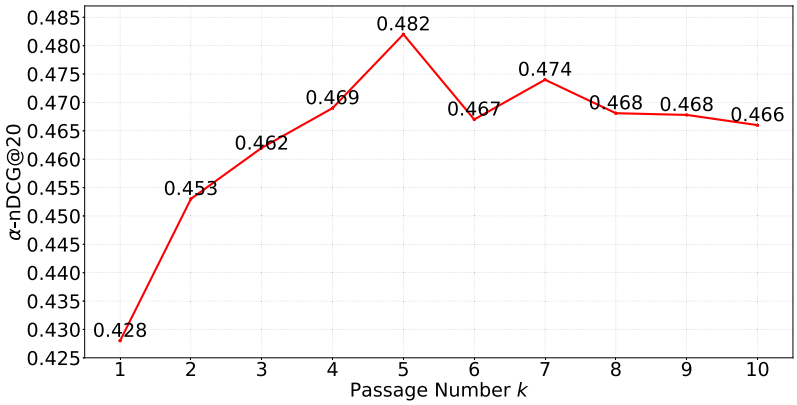

Fig. 5. The effects of the PAD with different passage number.

passage number less than three, which demonstrates the necessity of leveraging multiple passages to better represent documents.

The changing tendency of α-nDCG@20 with different passage numbers is shown in Figure 5. The performance of other metrics has a similar trend. In our experiments, the model with five passages obtains the best results in terms of α-nDCG@20. Additionally, α-nDCG increases while the passage number increases from 1 to 5. However, models with a passage number of more than 5 could achieve competitive performance. The tendency shown in Figure 5 reveals the general effects of passages that using more passages benefits the diversity ranking model than only using one passage. Considering that the subtopic number covered by most documents is less than 5 (from Figure 4(a)), using more passages (more than 5) may also introduce more irrelevant passages, which is a possible reason why the metric decline with passage number more than 5.

5.3 Effects of Different Query Types

To investigate the effects of our method on different types of queries, we calculate the ranking metrics of PAD on ambiguous queries and faceted queries, respectively. The results are shown in Table 7. In the ClueWeb dataset, PAD gets higher performance in faceted queries (0.522 in terms

Table 7. Effects of PAD on Different Types of Queries

Category	Number	α-nDCG	ERR-IA	NRBP	S-rec
All queries	198	.482	.386	.357	.670
Ambiguous queries	57	.385	.277	.243	.590
Faceted queries	141	**.522**	**.433**	**.407**	**.702**

Table 8. Effects of Only Modeling Relevance in Search Result Diversification Task

Category	Methods	α-nDCG	ERR-IA	NRBP	S-rec
Diversity-aware Method	PAD	**.482**	**.386**	**.357**	**.670**
Relevance Methods	selected passages mean (BERT)	.334	.231	.188	.600
	selected passages max (BERT)	.364	.265	.227	.610
	selected passages mean (E5)	.290	.189	.145	.552
	selected passages max (E5)	.299	.195	.149	.573

of α-nDCG) than ambiguous queries (0.385 in terms of α-nDCG). In the case of faceted queries, relevant documents can cover different facets of the general queries, in which our passage-aware diversification method PAD can model different parts of the document content. Credit to the capability of modeling passage-level relationships, PAD can acquire better performance on faceted queries than ambiguous queries. Since most queries are faceted queries, it is worthwhile modeling the relations within documents, such as passage relations.

Apart from the faceted queries, our method PAD is also compatible with ambiguous queries. For example, in a long document relevant to an ambiguous query, our method PAD can sense the most relevant parts related to the query credit to a passage selection module, which is our advantage compared with document-level methods. Based on the passage relevance and relations of different passages, our passage-aware method can better model document-level relevance and novelty.

5.4 Effects of Passage Relevance

In the framework of PAD, we first leverage a classifier to filter irrelevant passages and model document diversity in the following diversified ranking component. Therefore, PAD models both passage relevance and diversity. Hence, we investigate the performance of the model with only passage relevance scores. Therefore, we compared our diversity-aware method PAD with several relevance-only methods that use passage relevance scores to rank documents. The results are shown in Table 8. It is worth noting that the four relevance methods are different from our method PAD, because they could not sense the diversity relations of the passages. More specifically, the four compared methods use the mean or maximum of the relevance scores of the selected passages as the documents' ranking scores. The relevance scores of the passages are calculated as the cosine similarity of the passage embeddings and query embeddings. We adopt two models, BERT [17] and E5 [43], to generate these embeddings. The E5 model used in our experiments is downloaded from HuggingFace.[9] In general, the four relevance-modeling baseline methods get much lower performance than PAD. According to the results, the effects of E5 are slightly lower than BERT, which is already used in the selection procedure of PAD. The experimental results are shown in Table 8. Compared with the original PAD, only modeling passage relevance gets much lower scores in all diversity metrics, which implies that only modeling passage relevance is not enough in the search result diversification task.

[9]https://huggingface.co/intfloat/e5-base-v2

Table 9. Effects of Different Passage Selection Strategies in PAD

	α-nDCG	ERR-IA	NRBP	S-rec
PAD	**.482**	**.386**	**.357**	.670
w/ BERT selection	.473	.376	.346	.663
w/ BM25 selection	.475	.378	.347	.667
w/ MMR selection	.478	.381	.352	**.671**

5.5 Effects of Different Selection Strategies

The inclusion of a passage selection stage in our method is indeed critical for several reasons. First, the length of documents can vary significantly across different instances, resulting in a varying number of passages. This attribute makes uniform processing in a ranking model less practical. Second, not every passage within a document is relevant to the query, and neglecting to select the relevant passages could introduce considerable noise into downstream ranking models. Last, using all passages extracted from a document could overwhelm a ranking model, resulting in unnecessary computational costs. Hence, selecting the most essential top-k passages is a necessary step in our method.

To further investigate the effects of different passage selection strategies, we also examine the performance of PAD with three different selection methods. Considering that the BERT-based classifier focuses on modeling the passage's semantic relevance while BM25 reflects the token matching degree, we adopt both scores to select passages. In this section, we demonstrate the effects of only using BERT classifier scores (denoted as *w/* BERT selection) or BM25 scores (denoted as *w/* BM25 selection), respectively. As shown in Table 9, merely leveraging BM25 scores or BERT scores cannot acquire the same good results as PAD does. Moreover, both strategies used in PAD can still have high performance, which demonstrates the robust framework of our method.

What if we consider passage diversity in the passage selection procedure? To answer this question, we apply an MMR module to PAD. More specifically, we not only leverage passage relevance scores but also passage similarity scores from cosine similarity of passage BERT embeddings to filter representative passages. The experimental results (denoted as *w/* MMR selection) are shown in Table 9. The S-rec metric of PAD with an MMR selection (0.671) is a little higher than the original PAD version (0.670), while the other three metrics decline with an MMR selection. A possible reason is that an MMR selection strategy is beneficial to obtain more diverse passages that cover more subtopics (reflected in the S-rec metric). However, the side effect of this strategy is that more irrelevant passages are also chosen and passed to the next phase, which accounts for the decrease of α-nDCG, ERR-IA, and NRBP. According to the experimental results, the effects of passage relevance are greater than diversity. Considering the additional amount of computation and effects brought by MMR, we choose the original selection strategy of PAD.

5.6 Runtime Efficiency

To demonstrate the efficiency of PAD, we record the inference time of our ranking model with different passage numbers k. As shown in Figure 6, the average processing time of each query is 16.7 ms for PAD with only one passage, while PAD with 10 passages needs 61.2 ms to diversify the results of one query. The process time of PAD with $k = 10$ is only 3.66 times that of PAD with $k = 1$. However, PAD achieves best results in terms of α-nDCG@20 with 5 passages, which needs 32.4 ms (only 94.01% more in process time compared with one passage) to process a query. Compared with the inference time of Graph4DIV (21.4 ms with 0.468 in terms of α-nDCG), a 5-passage PAD achieves 0.482 in terms of α-nDCG (1.4% improvement) with 32.4 ms. According to the experimental results, we can balance effectiveness and time cost by leveraging five passages in PAD.

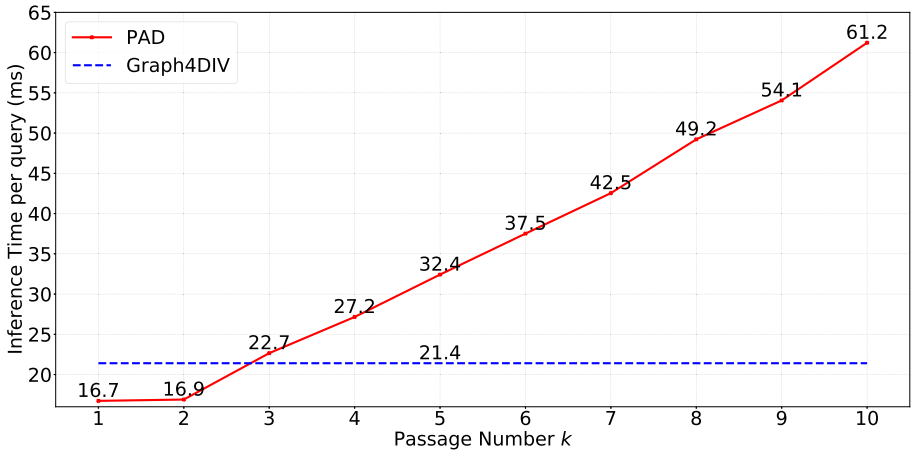

Fig. 6. The process time influenced by the passage number k.

Since introducing too many passages can also bring more noise, a high-performance PAD with a few passages does not come with a lot of extra time overhead. In general, PAD could achieve large-effectiveness (Figure 5) improvement with an acceptable degree of time cost (Figure 6). Therefore, leveraging multiple passages to model document diversity is both effective and efficient in practice.

6 CONCLUSIONS

In this article, we propose an implicit approach PAD to model the document's diversity through multiple passage interactions. To obtain the representative passages of the documents, we leverage a passage relevance classifier to select the top-k passages. Furthermore, we model the passage's global interactions via the GloEnc. Then the selected document state will be aggregated by the SelEnc. The document representation is automatically learned by the DocEnc from the passages that belong to it. Together with the context-aware features from the SelEnc and DocEnc, PAD selects the novel candidate document at each step. The experimental results show the efficiency and effectiveness of our model. In the future, we will explore more utilization of passages in explicit search result diversification.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the 2nd International Conference on Web Search and Web Data Mining (WSDM'09)*, Ricardo Baeza-Yates, Paolo Boldi, Berthier A. Ribeiro-Neto, and Berkant Barla Cambazoglu (Eds.). ACM, 5–14. https://doi.org/10.1145/1498759.1498766

[2] Qingyao Ai, Brendan O'Connor, and W. Bruce Croft. 2018. A neural passage model for ad-hoc document retrieval. In *Proceedings of the 40th European Conference on IR Research, Advances in Information Retrieval (ECIR'18), Lecture Notes in Computer Science, Vol. 10772*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer, 537–543. https://doi.org/10.1007/978-3-319-76941-7_41

[3] Joshua Ainslie, Santiago Ontañón, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding long and structured inputs in transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 268–284. https://doi.org/10.18653/v1/2020.emnlp-main.19

[4] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. arXiv:1607.06450. Retrieved from http://arxiv.org/abs/1607.06450

[5] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv:2004.05150. Retrieved from https://arxiv.org/abs/2004.05150

[6] Michael Bendersky and Oren Kurland. 2008. Re-ranking search results using document-passage graphs. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. Association for Computing Machinery, New York, NY, 853–854. https://doi.org/10.1145/1390334.1390539

[7] Michael Bendersky and Oren Kurland. 2008. Utilizing passage-based language models for document retrieval. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval (ECIR'08)*. Springer-Verlag, Berlin, 162–174.

[8] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao". 2009. Clueweb09 Data Set. Retrieved from https://boston.lti.cs.cmu.edu/Data/clueweb09/

[9] James P. Callan. 1994. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*. Springer-Verlag, Berlin, 302–310.

[10] Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (Eds.). ACM, 335–336. https://doi.org/10.1145/290941.291025

[11] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy J. Lin (Eds.). ACM, 621–630. https://doi.org/10.1145/1645953.1646033

[12] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong (Eds.). ACM, 659–666. https://doi.org/10.1145/1390334.1390446

[13] Charles L. A. Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An effectiveness measure for ambiguous and underspecified queries. In *Advances in Information Retrieval Theory, 2nd International Conference on the Theory of Information Retrieval (ICTIR'09), Lecture Notes in Computer Science, Vol. 5766*, Leif Azzopardi, Gabriella Kazai, Stephen E. Robertson, Stefan M. Rüger, Milad Shokouhi, Dawei Song, and Emine Yilmaz (Eds.). Springer, 188–199. https://doi.org/10.1007/978-3-642-04417-5_17

[14] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, 985–988. https://doi.org/10.1145/3331184.3331303

[15] Van Dang and W. Bruce Croft. 2012. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson (Eds.). ACM, 65–74. https://doi.org/10.1145/2348283.2348296

[16] 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 (SIGIR'13)*, Gareth J. F. Jones, Paraic Sheridan, Diane Kelly, Maarten de Rijke, and Tetsuya Sakai (Eds.). ACM, 603–612. https://doi.org/10.1145/2484028.2484095

[17] 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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19), Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[18] Yue Feng, Jun Xu, Yanyan Lan, Jiafeng Guo, Wei Zeng, and Xueqi Cheng. 2018. From greedy selection to exploratory decision-making: Diverse ranking with policy-value networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*. Association for Computing Machinery, New York, NY, 125–134. https://doi.org/10.1145/3209978.3209979

[19] Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2Query-: When less is more. In *Advances in Information Retrieval, 45th European Conference on Information Retrieval (ECIR'23), Part II, Lecture Notes in Computer Science, Vol. 13981*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 414–422. https://doi.org/10.1007/978-3-031-28238-6_31

[20] Ankit Gupta and Jonathan Berant. 2020. GMAT: Global memory augmentation for transformers. arXiv:2006.03274. Retrieved from https://arxiv.org/abs/2006.03274

[21] Sebastian Hofstätter, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. 2021. *Intra-Document Cascading: Learning to Select Passages for Neural Document Ranking*. Association for Computing Machinery, New York, NY, 1349–1358. https://doi.org/10.1145/3404835.3462889

[22] Sha Hu, Zhicheng Dou, Xiao-Jie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15)*, James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu (Eds.). ACM, 63–72. https://doi.org/10.1145/2806416.2806455

[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 (SIGIR'17)*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 545–554. https://doi.org/10.1145/3077136.3080805

[24] Marcin Kaszkiel and Justin Zobel. 1997. Passage retrieval revisited. *SIGIR Forum* 31, SI (Jul. 1997), 178–185. https://doi.org/10.1145/278459.258561

[25] Eyal Krikon, Oren Kurland, and Michael Bendersky. 2011. Utilizing inter-passage and inter-document similarities for reranking search results. *ACM Trans. Inf. Syst.* 29, 1, Article 3 (Dec. 2011), 28 pages. https://doi.org/10.1145/1877766.1877769

[26] Shangsong Liang, Fei Cai, Zhaochun Ren, and Maarten de Rijke. 2016. Efficient structured learning for personalized diversification. *IEEE Trans. Knowl. Data Eng.* 28, 11 (2016), 2958–2973. https://doi.org/10.1109/TKDE.2016.2594064

[27] Jiongnan Liu, Zhicheng Dou, Xiao-Jie Wang, Shuqi Lu, and Ji-Rong Wen. 2020. DVGAN: A minimax game for search result diversification combining explicit and implicit features. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20)*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 479–488. https://doi.org/10.1145/3397271.3401084

[28] Xiaoyong Liu and W. Bruce Croft. 2002. Passage retrieval based on language models. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02)*. Association for Computing Machinery, New York, NY, 375–382. https://doi.org/10.1145/584792.584854

[29] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR'19)*. OpenReview.net.

[30] Elke Mittendorf and Peter Schäuble. 1994. Document and passage retrieval based on hidden markov models. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*. Springer-Verlag, Berlin, 318–327.

[31] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 Co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS'16) CEUR Workshop Proceedings, Vol. 1773*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org.

[32] Kai Ouyang, Xianghong Xu, Zuotong Xie, Hai-Tao Zheng, and Yanxiong Lu. 2023. Modeling global-local subtopic distribution with hypergraph to diversify search results. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'23)*. IEEE, 1–8. https://doi.org/10.1109/IJCNN54540.2023.10191529

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

[34] 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 (CIKM'20)*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 1265–1274. https://doi.org/10.1145/3340531.3411914

[35] Xubo Qin, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. GDESA: Greedy diversity encoder with self-attention for search results diversification. *ACM Trans. Inf. Syst.* 41, 2 (2023), 36 pages. https://doi.org/10.1145/3544103

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[37] Gerard Salton, J. Allan, and Chris Buckley. 1993. Approaches to passage retrieval in full text information systems *(SIGIR'93)*. Association for Computing Machinery, New York, NY, 49–58. https://doi.org/10.1145/160688.160693

[38] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti (Eds.). ACM, 881–890. https://doi.org/10.1145/1772690.1772780

[39] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2012. On the role of novelty for search result diversification. *Inf. Retr.* 15, 5 (2012), 478–502. https://doi.org/10.1007/S10791-011-9180-X

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

[41] Zhan Su, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2022. Knowledge enhanced search result diversification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22)*. Association for Computing Machinery, New York, NY, 1687–1695. https://doi.org/10.1145/3534678.3539459

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008.

[43] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. arXiv:2212.03533. Retrieved from https://arxiv.org/abs/2212.03533

[44] Mengqiu Wang and Luo Si. 2008. Discriminative probabilistic models for passage based retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. Association for Computing Machinery, New York, NY, 419–426. https://doi.org/10.1145/1390334.1390407

[45] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Jingtao Zhan, Yukun Zheng, Min Zhang, and Shaoping Ma. 2020. *Leveraging Passage-Level Cumulative Gain for Document Ranking*. Association for Computing Machinery, New York, NY, 2421–2431. https://doi.org/10.1145/3366423.3380305

[46] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference of Machine Learning (ICML'08), ACM International Conference Proceeding Series, Vol. 307*, William W. Cohen, Andrew McCallum, and Sam T. Roweis (Eds.). ACM, 1192–1199. https://doi.org/10.1145/1390156.1390306

[47] 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*, Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 113–122. https://doi.org/10.1145/2766462.2767710

[48] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling document novelty with neural tensor network for search result diversification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 395–404. https://doi.org/10.1145/2911451.2911498

[49] 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 (SIGIR'17)*. Association for Computing Machinery, New York, NY, 535–544. https://doi.org/10.1145/3077136.3080775

[50] Jun Xu, Zeng Wei, Long Xia, Yanyan Lan, Dawei Yin, Xueqi Cheng, and Ji-Rong Wen. 2020. Reinforcement learning to rank with pairwise policy gradient. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20)*. Association for Computing Machinery, New York, NY, USA, 509–518. https://doi.org/10.1145/3397271.3401148

[51] 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 (WWW'21)*. Association for Computing Machinery, New York, NY, USA, 127–136. https://doi.org/10.1145/3442381.3449831

[52] Grace Hui Yang and Ian Soboroff. 2016. TREC 2016 dynamic domain track overview. In *Proceedings of the 25th Text REtrieval Conference (TREC'16), NIST Special Publication, Vol. 500-321*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

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

[54] 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'22), the 34th Conference on Innovative Applications of Artificial Intelligence (IAAI'22), the 12th Symposium on Educational Advances in Artificial Intelligence (EAAI'22)*. AAAI Press, 10399–10407. https://doi.org/10.1609/aaai.v36i9.21282

[55] Yisong Yue and Thorsten Joachims. 2008. Predicting diverse subsets using structural SVMs. In *Machine Learning, Proceedings of the 25th International Conference (ICML'08), ACM International Conference Proceeding Series, Vol. 307*, William W. Cohen, Andrew McCallum, and Sam T. Roweis (Eds.). ACM, 1224–1231. https://doi.org/10.1145/1390156.1390310

[56] ChengXiang Zhai, William W. Cohen, and John D. Lafferty. 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*, Charles L. A. Clarke, Gordon V. Cormack, Jamie Callan, David Hawking, and Alan F. Smeaton (Eds.). ACM, 10–17. https://doi.org/10.1145/860435.860440

[57] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu. 2014. Learning for search result diversification. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)*, Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin (Eds.). ACM, 293–302. https://doi.org/10.1145/2600428.2609634

Received 23 March 2023; revised 21 February 2024; accepted 6 March 2024