

Learning to Diversify Search Results via Subtopic Attention

Zhengbao Jiang^{1,2}, Ji-Rong Wen^{1,2,3}, Zhicheng Dou^{1,2},
Wayne Xin Zhao^{1,2}, Jian-Yun Nie⁴, Ming Yue^{1,2}

¹School of Information, Renmin University of China

²Beijing Key Laboratory of Big Data Management and Analysis Methods, China

³Key Laboratory of Data Engineering and Knowledge Engineering, MOE, China

⁴DIRO, Université de Montréal, Québec

rucjzb@163.com, jirong.wen@gmail.com, dou@ruc.edu.cn,
batmanfly@gmail.com, nie@iro.umontreal.ca, yomin@ruc.edu.cn

ABSTRACT

Search result diversification aims to retrieve diverse results to satisfy as many different information needs as possible. Supervised methods have been proposed recently to learn ranking functions and they have been shown to produce superior results to unsupervised methods. However, these methods use implicit approaches based on the principle of Maximal Marginal Relevance (MMR). In this paper, we propose a learning framework for explicit result diversification where subtopics are explicitly modeled. Based on the information contained in the sequence of selected documents, we use attention mechanism to capture the subtopics to be focused on while selecting the next document, which naturally fits our task of document selection for diversification. The framework is implemented using recurrent neural networks and max-pooling which combine distributed representations and traditional relevance features. Our experiments show that the proposed method significantly outperforms all the existing methods.

KEYWORDS

search result diversification; subtopics; attention

ACM Reference format:

Zhengbao Jiang^{1,2}, Ji-Rong Wen^{1,2,3}, Zhicheng Dou^{1,2}, Wayne Xin Zhao^{1,2}, Jian-Yun Nie⁴, Ming Yue^{1,2}. 2017. Learning to Diversify Search Results via Subtopic Attention. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 7–11, 2017*, 10 pages. DOI: <http://dx.doi.org/10.1145/3077136.3080805>

1 INTRODUCTION

In real search scenario, queries issued by users are usually ambiguous or multi-faceted. In addition to being relevant to the query, the retrieved documents are expected to be as diverse as possible in order to cover different information needs. For example, when users issue “apple”, the underlying intents could be the IT company or the fruit. The retrieved documents should cover both topics to increase the chance to satisfy users with different information needs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '17, Shinjuku, Tokyo, Japan

© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5022-8/17/08...\$15.00
DOI: <http://dx.doi.org/10.1145/3077136.3080805>

Traditional approaches to search result diversification are usually unsupervised and adopt manually defined functions with empirically tuned parameters. Depending on whether the underlying intents (or subtopics) are explicitly modeled, they can be categorized into implicit and explicit approaches [28]. Implicit approaches [6] do not model intents explicitly. They emphasize novelty, i.e. the following document should be “different” from the former ones based on some similarity measures. Instead, explicit approaches [1, 12, 13, 16, 27, 35] model intents (or subtopics) explicitly. They aim to improve intent coverage, i.e. the following document should cover the intents not satisfied by previous ones. Intents or subtopics can be determined by techniques such as query reformulation [2, 14, 34, 38] and query clustering based on query logs and other types of information. Existing studies showed that explicit approaches have better performance [12, 13, 16, 27, 35] than implicit approaches due to several reasons: on the one hand, they provide a more natural way to handle subtopics than implicit approaches; on the other hand, their ranking functions are closer to the diversity evaluation metrics which are mostly based on explicit subtopics. Furthermore, most similarity measures used in the implicit approaches, e.g., those based on language model or vector space model, are determined globally on the whole documents, regardless of possible search intents. This might be problematic for search result diversification: two documents could contain similar words and considered globally similar, but this similar part may be unrelated to underlying search intents.

To avoid heuristic and handcrafted functions and parameters, a new family of research work using supervised learning is proposed. They try to learn a ranking function automatically. Their major focus lies in the modeling of diversity, including structural prediction [36], rewarding functions for novel contents [39], measure-based direct optimization [32], and neural network based method [33]. Regardless of diversity modeling and optimization methods, all these solutions inherit the spirit of MMR which is an implicit approach and do not take intents into consideration. Although the learning methods may result in a better similarity measure, they are hindered by the gap between reducing document redundancy and improving intent coverage. They suffer from similar problems with implicit unsupervised approaches. Without modeling subtopics explicitly, they can't directly improve intent coverage. Hence, there is a need to incorporate explicit subtopic modeling into supervised diversification methods.

To address the above issue, we propose to model subtopics in a general supervised learning framework. Our framework combines

Table 1: Subtopic relevance example.

doc\subtopic	i_1	i_2	i_3
d_1	√	√	×
d_2	√	√	×
d_3	×	×	√
d_4	×	√	×

the strengths of both explicit unsupervised approaches and (implicit) supervised approaches. First, subtopics are explicitly modeled, allowing us to improve intent coverage in a proactive way. Second, it automatically learns the diversification ranking function, and is able to capture complex interaction among documents and subtopics. We call this framework Document Sequence with Subtopic Attention (DSSA). More specifically, to select the next document, we first model the sequence of selected documents in order to capture their contents as well as their relationship with the subtopics. Then based on the information contained by previous documents, attention mechanism is used to determine the undercovered subtopics to which we have to pay attention in selecting the next document. Attention mechanism has been successfully used to deal with various problems in image understanding [24] and NLP [3, 21]. This mechanism corresponds well to the document selection problem in search result diversification: attention on subtopics changes along with the addition of a document in the result list. For example. Assume that we have 3 subtopics and 4 documents whose relevance judgments are shown in Table 1. Given that we have selected d_1 and d_2 , which cover subtopics i_1 and i_2 , the attention for next choice should incline to i_3 which is not covered, thus d_3 is a better choice than d_4 at this position. We will show that the DSSA framework is general enough to cover the ideas of previous unsupervised explicit methods.

We then propose a specific implementation of DSSA using recurrent neural networks (RNN) and max-pooling to leverage both distributed representations and traditional relevance features, which we call DSSA-RNNMP. Experimental results on TREC Web Track data show that our method outperforms the existing methods significantly. To our knowledge, this is the first time that a supervised learning framework with attention mechanism is used to model subtopics explicitly for search result diversification.

2 RELATED WORK

2.1 Implicit Diversification Approaches

The basic assumption of implicit diversification approaches is that dissimilar documents are more likely to satisfy different information needs. The most representative approach is MMR [6]:

$$S_{\text{MMR}}(q, d, C) = (1 - \lambda)S^{\text{rel}}(d, q) - \lambda \max_{d_j \in C} S^{\text{div}}(d, d_j), \quad (1)$$

where S^{rel} and S^{div} model document d 's relevance to the query q and its similarity to a selected documents d_j respectively. To gain high ranking score, a document should not only be relevant, but also be dissimilar from the selected documents. The definition of measures for relevance and document similarity is crucial, which is done manually in this approach.

Table 2: Categorization of diversification approaches.

	unsupervised	supervised
implicit	MMR	SVM-DIV, R-LTR, PAMM, NTN
explicit	IA-Select, xQuAD, PM2, TxQuAD, TPM2, HxQuAD, HPM2, 0-1 MSKP	DSSA (our approach)

Recently, machine learning methods have been leveraged to learn score functions. Yue and Joachims [36] proposed SVM-DIV which uses structural SVM to learn to identify a document subset with maximum word coverage. However, word coverage may be different from intent coverage. Optimizing the former may not necessarily lead to optimizing the latter. Similar to MMR, Zhu et al. [39] proposed relational learning-to-rank model (R-LTR) which learns to score a document based on both relevance and novelty automatically, in order to maximize the probability of optimal rankings. Based on R-LTR score function, Xia et al. [32] proposed a perceptron algorithm using measures as margins (PAMM) to directly optimize evaluation metrics by enlarging the score margin of positive and negative rankings. They further proposed to use a neural tensor network (NTN) [33] to measure document similarity automatically from document representations, which avoids the burden to define handcrafted diversity features.

The above supervised approaches are shown to outperform the unsupervised counterparts. However, they are all implicit approaches without using subtopics. In this paper, we propose a learning-based explicit approach which models subtopics explicitly.

2.2 Explicit Diversification Approaches

Explicit approaches model subtopics underlying a query, aiming at returning documents covering as many subtopics as possible. These approaches leverage external resources to explicitly represent information needs in subtopics. IA-Select [1] uses classified topical categories based on ODP taxonomy. xQuAD [27] is a probabilistic framework that uses query reformulations as intent representations. PM2 [13] tackles search result diversification problem from the perspective of proportionality. TxQuAD and TPM2 [12] represent intents by terms and transform intent coverage to term coverage. Hu et al. [16] proposed to use a hierarchical structure for subtopics instead of a flat list, which copes with the inherent interaction among subtopics. Two specific models, namely HxQuAD and HPM2, were proposed using hierarchical structure. Yu et al. [35] formulated diversification task as a 0-1 multiple subtopic knapsacks (0-1 MSKP) problem where documents are chosen like filling up multiple subtopic knapsacks. To tackle this NP-hard problem, max-sum belief propagation is used.

As summarized in Table 2, all existing explicit approaches are unsupervised and the functions and parameters are defined heuristically. In this paper, we use supervised learning to model the interaction among documents and subtopics simultaneously.

2.3 RNN with Attention Mechanism

RNN can capture the interdependency between elements in a sequence. Attention mechanism, which is usually built on RNN, mimics human attention behavior focusing on different local region of

the object (an image, a sentence, etc) at different times. In computer vision, Google DeepMind [24] used RNN with attention to extract information from an image by adaptively selecting a sequence of the most informative regions instead of the whole image. In NLP, attention mechanism is typically used in neural machine translation (NMT). Traditional encoder-decoder models encode the source sentence into a fixed-length vector from which the target sentence is decoded. Such fixed-length vector may not be powerful enough to reflect all the information of the source sentence. An attention-based model [3] was proposed to automatically pay unequal and varied attention to source words during decoding process. In particular, to decide the next target word, not only the fixed-length vector, but also the hidden states corresponding to source words relevant to the target word are used. Luong et al. [21] generalized the idea and proposed two classes of attention mechanism, namely global and local approaches. In this paper, attention mechanism is used on subtopics, which guides the model to emphasize different intents at different positions.

In the following section, we will first propose a general framework, then instantiate it with a specific implementation.

3 DOCUMENT SEQUENCE WITH SUBTOPIC ATTENTION FRAMEWORK

Given a query set \mathcal{Q} , a document set \mathcal{D}_q and a subtopic set \mathcal{I}_q for each query $q \in \mathcal{Q}$, the goal of explicit methods is to learn a ranking function $f(q, \mathcal{D}_q, \mathcal{I}_q)$ which is expected to output a ranking of documents in \mathcal{D}_q that is both *relevant* and *diverse*. The loss function could be written in the following general form:

$$\sum_{q \in \mathcal{Q}} L(f(q, \mathcal{D}_q, \mathcal{I}_q), \mathcal{Y}_q), \quad (2)$$

where L measures the quality gap between the ranking outputted by f and the best ranking \mathcal{Y}_q . Different from traditional retrieval tasks, diversity has to be considered in the ranking and evaluation process. Theoretically, diversity ranking is NP-hard [1, 7]. Hence, a common strategy is to make greedy selections [6, 27]: at the t -th position, we assume that $t - 1$ documents have been selected and formed a document sequence C_{t-1} . The task is to select a locally optimal document d_t from the remaining candidate documents based on a score function $S(q, d_t, C_{t-1}, \mathcal{I}_q)$. Note that implicit supervised methods correspond to the case where \mathcal{I}_q is an empty set.

To motivate our approach, we start with the ideas of the unsupervised explicit approaches, which can be formulated as the following general form:

$$\begin{aligned} S_{\text{unsupervised}}(q, d_t, C_{t-1}, \mathcal{I}_q) = & \\ (1 - \lambda)S^{\text{rel}}(d_t, q) + & \Rightarrow \text{relevance} \\ \lambda \sum_{i_k \in \mathcal{I}_q} S^{\text{div}}(d_t, i_k) \underbrace{A(C_{t-1}, \mathcal{I}_q)_k}_{\text{subtopic weights}}, & \Rightarrow \text{diversity} \end{aligned} \quad (3)$$

where $i_k \in \mathcal{I}_q$ is the k -th subtopic of q and S^{rel} and S^{div} calculate document d_t 's relevance to a query and to a subtopic respectively. The essence of diversity lies in the function A which calculates the weights for subtopics \mathcal{I}_q based on previous document sequence C_{t-1} . For xQuAD, $A(C_{t-1}, \mathcal{I}_q)_k = P(i_k|q) \prod_{d_j \in C_{t-1}} (1 - P(d_j|i_k))$ where $P(i_k|q)$ is the initial importance of subtopic i_k , $P(d_j|i_k)$ is

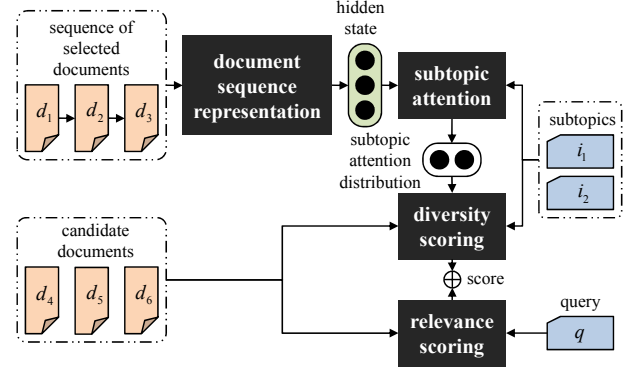


Figure 1: Illustration of DSSA framework.

Table 3: Notations in DSSA.

Notation	Definition
d_t, q, i_k, r	the t -th document, query, k -th subtopic, a ranking.
\mathbf{v}_{d_t}	representation of the document at the t -th position.
\mathbf{v}_q	representation of the query.
\mathbf{v}_{i_k}	representation of the k -th subtopic.
\mathbf{h}_t	hidden state of previous t documents.
$a_{t,k}$	attention on the k -th subtopic at the t -th position. $\sum_{k=1}^K a_{t,k} = 1, a_{t,k} \in [0, 1]$ where K is the number of subtopics. A large value means that this subtopic is less satisfied by previous $t - 1$ documents and thus needs more attention at the t -th position.
s_{d_t}	the final score of the document at the t -th position.

the probability that d_j is relevant to i_k . The weight of a subtopic is determined by the likelihood that previous documents are not relevant to this subtopic. PM2 mimics seats allocation of competing political parties to adjust subtopic weights after each selection, i.e. $A(C_{t-1}, \mathcal{I}_q)$ is estimated according to the difference between the subtopic's distributions in C_{t-1} and in \mathcal{I}_q . All these methods don't model the selected documents as a sequence. In addition, the functions and parameters are heuristically defined, which may not best fit the final goal.

To tackle the above problems, we extend Equation (3) to the following general learning framework:

$$\begin{aligned} S_{\text{DSSA}}(q, d_t, C_{t-1}, \mathcal{I}_q) = s_{d_t} = & \\ (1 - \lambda)S^{\text{rel}}(\mathbf{v}_{d_t}, \mathbf{v}_q) + & \Rightarrow \text{relevance} \\ \lambda S^{\text{div}}\left(\mathbf{v}_{d_t}, \mathbf{v}_{i_{(c)}} \underbrace{\mathcal{A}\left(\mathcal{H}([\mathbf{v}_{d_1}, \dots, \mathbf{v}_{d_{t-1}}]), \mathbf{v}_{i_{(c)}}\right)}_{\text{subtopic attention}}\right), & \Rightarrow \text{diversity} \end{aligned} \quad (4)$$

where documents, queries, and subtopics are denoted by their representations, as explained in Table 3. In this paper, we focus on

learning a ranking function only and assume that these representations are given and will not be modified. There are three main components, namely (1) **document sequence representation** component \mathcal{H} , (2) **subtopic attention** component \mathcal{A} , and (3) **scoring** component \mathcal{S}^{rel} and \mathcal{S}^{div} , which are also illustrated in Figure 1. This framework is inspired from the attention models used in image understanding [24] and neural machine translation [3, 21], however adapted to our diversification task.

Next, we briefly describe the three components. The document sequence representation component \mathcal{H} encodes the information contained in document sequence C_{t-1} into a fixed-length hidden state h_{t-1} , which could consider the interaction and dependency among these documents. h_{t-1} could be viewed as a comprehensive and high-level representation of C_{t-1} . The subtopic attention $a_{t,(\cdot)}$ is calculated by the subtopic attention component \mathcal{A} using h_{t-1} and subtopic representations v_{i_k} . The attention evolves from the first to the last ranking position, driving the model to emphasize different subtopics based on previous document sequence. Finally, the scoring components \mathcal{S}^{rel} and \mathcal{S}^{div} calculate relevance and diversity scores respectively. Notice that \mathcal{S}^{div} is not limited to be a weighted sum over all subtopics as Equation (3). It can incorporate more complex interaction among subtopics.

The essence of this framework can be summarized as follows. Along with the selection of more documents, we encode the information of previous document sequence, and the attention mechanism will monitor the degree of satisfaction for each subtopic. High scores are assigned to the documents relevant to less covered subtopics. Finally, multiple subtopics would be well covered by adaptively learning the attention. In this way, our framework builds an intuitive approach to explicitly model subtopics. We name the framework **Document Sequence with Subtopic Attention (DSSA)**. DSSA is a unified architecture that takes both relevance and diversity into consideration, and diversity is achieved by modeling the interaction among documents and subtopics.

4 RESULT DIVERSIFICATION USING DSSA

In this section, we instantiate DSSA to a concrete form and articulate the training and prediction algorithms. The main idea of DSSA is to dynamically capture accumulative relevance information of previous document sequence, so as to calculate subtopic attention. Inspired by the recent progress on sequence data modeling, we adapt RNN to capture the information of previous document sequence based on distributed representations of documents. However, the effectiveness of distributed representation heavily depends on a large amount of training data. Typically, the representation is built automatically using the data to optimize an objective function [17]. We do not have such large data and we can only use unsupervised methods (e.g. doc2vec) to create representation, of which the effectiveness could be suboptimal. Indeed, our preliminary experiments using only the distributed representation created by unsupervised methods yield low effectiveness. To compensate this weakness, we also use traditional relevance features such as BM25 score, which are proven useful, to calculate subtopic attention and final score. Such a combination of distributed representations and features has been used in several previous works [29, 33]. In addition to RNN, we also adopt the way using max-pooling [33],

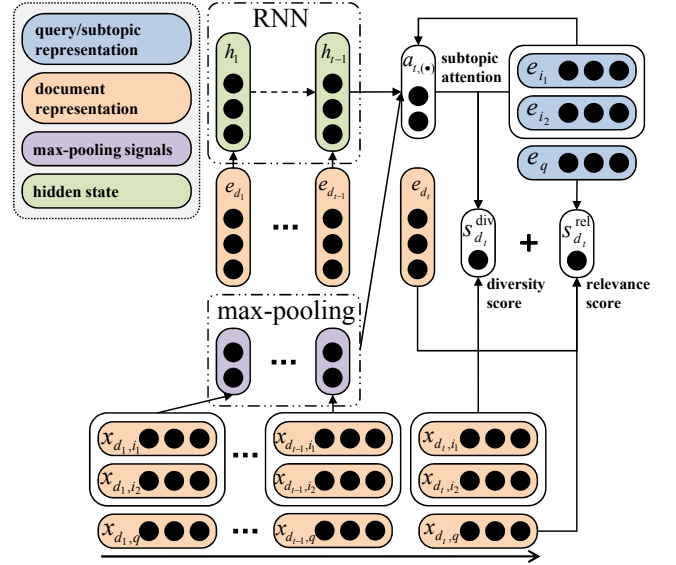


Figure 2: Architecture of DSSA-RNNMP. Previous $t - 1$ documents are encoded into h_{t-1} from distributed representations $e_{d_1}, \dots, e_{d_{t-1}}$. Attention on the k -th subtopic $a_{t,k}$ is then calculated based on (1) hidden state h_{t-1} and subtopic representation e_{i_k} (2) max-pooling on relevance features $x_{d_1, i_k}, \dots, x_{d_{t-1}, i_k}$.

Table 4: Parameters in DSSA-RNNMP.

Notation	Definition
W^n, b^n	parameters of RNN with vanilla cell.
W^a, w^p	parameters used in subtopic attention.
W^s, w^r	parameters used in scoring.

which has been shown effective, to implement subtopic attention mechanism. We call this model DSSA-RNNMP (DSSA model using RNN and Max-Pooling), as illustrated in Figure 2. In addition, we also propose a list-pairwise approach for optimization, which is different from the existing studies.

4.1 A Neural Network Implementation

We first describe the constitution of representations, namely v_{d_t} , v_q , and v_{i_k} , then elaborate how we implement **document sequence representation**, **subtopic attention**, and **scoring** components. The parameters to be learned are listed in Table 4.

v_{d_t} : the representation of a document is composed of two parts: distributed representations and relevance features. Distributed representation can be constructed in different ways. In this paper, we consider three methods: SVD, LDA [4], and doc2vec [18]. Relevance features are those used in traditional IR, such as BM25 score etc. Suppose that we have a distributed representation of size E_d , K subtopics, and R relevance features, the total size of v_{d_t} would be $E_d + R + KR$. We use $e_{d_t} \in \mathbb{R}^{E_d}$, $x_{d_t, q}$ and $x_{d_t, i_k} \in \mathbb{R}^R$ to denote distributed representation, relevance features for a query and a subtopic respectively.

$\mathbf{v}_q, \mathbf{v}_{i_k}$: we first retrieve top Z documents using some basic retrieval model (such as BM25). These documents are concatenated as a pseudo document, then similar to \mathbf{e}_{d_t} , a distributed representation of size E_q is generated. For consistency, we also use \mathbf{e}_q and $\mathbf{e}_{i_k} \in \mathbb{R}^{E_q}$ to represent these representations.

4.1.1 Document Sequence Representation. \mathcal{H} is instantiated using RNN to encode the information of previous document sequence. Several types of RNN cell can be used, ranging from the simple vanilla cell, GRU cell [9], to LSTM cell [15]. For simplicity, we only show the vanilla cell here. At the t -th position, we derive the (accumulative) document sequence representation as follows:

$$\mathbf{h}_t = \tanh(\mathbf{W}^n[\mathbf{h}_{t-1}; \mathbf{e}_{d_t}] + \mathbf{b}^n), \quad (5)$$

where $\mathbf{W}^n \in \mathbb{R}^{U \times (U+E_d)}$ (U is the size of the hidden state), $\mathbf{b}^n \in \mathbb{R}^U$ and $[\cdot; \cdot]$ is a concatenation. The cell transforms previous hidden layer \mathbf{h}_{t-1} and current document distributed representation \mathbf{e}_{d_t} to another space, where a bias \mathbf{b}^n is added and a non-linear activation (i.e. \tanh) then happens, producing the next hidden layer \mathbf{h}_t . \mathbf{h}_0 is initialized as a vector of zeros. The vanilla cell can be easily replaced by GRU and LSTM cells, whose results will be report in Section 6.2.

4.1.2 Subtopic Attention. By looking at \mathbf{h}_{t-1} which stores the information of previous $t-1$ documents and \mathbf{e}_{i_k} which represents the meaning of each subtopic, we are capable of discovering which intents are not satisfied and thus need to be emphasized at the t -th position. To capture this idea, we use $\mathcal{A}'(\mathbf{h}_{t-1}, \mathbf{e}_{i_k})$ to measure the (unnormalized) importance of the k -th subtopic at the t -th position, which could be implemented in many ways. We consider the following two ways similar to [21]:

$$\mathcal{A}'(\mathbf{h}_{t-1}, \mathbf{e}_{i_k}) = \begin{cases} \mathbf{h}_{t-1}^\top \mathbf{W}^a \mathbf{e}_{i_k}, & (\text{general}) \\ -\mathbf{h}_{t-1}^\top \cdot \mathbf{e}_{i_k}, & (\text{dot}) \end{cases} \quad (6)$$

where $\mathbf{W}^a \in \mathbb{R}^{U \times E_q}$. The “general” operation uses bilinear tensor product to relate two vectors multiplicatively through its nonlinearity [30]. The “dot” product requires both vectors to be in the same space. Similar \mathbf{h}_{t-1} and \mathbf{e}_{i_k} mean that previous documents are likely to satisfy this subtopic, and thus a lower attention score will be attributed to it. The above way mainly relies on distributed representations, which may not always be effective, especially under limited data.

Hence, we further leverage relevance features to enhance the subtopic attention. \mathbf{x}_{d_t, i_k} directly reflects the degree of satisfaction for a subtopic-document pair and is combined linearly using \mathbf{w}^p to form an explicit signal. To derive the accumulative information of the document sequence, we adopt commonly used max-pooling to select the most significant signal from previous documents:

$$\mathcal{A}''(\mathbf{x}_{d_1, i_k}, \dots, \mathbf{x}_{d_{t-1}, i_k}) = \max([\mathbf{x}_{d_1, i_k}^\top \cdot \mathbf{w}^p, \dots, \mathbf{x}_{d_{t-1}, i_k}^\top \cdot \mathbf{w}^p]), \quad (7)$$

where $\mathcal{A}''(\mathbf{x}_{d_1, i_k}, \dots, \mathbf{x}_{d_{t-1}, i_k})$ measures the degree of satisfaction of the k -th subtopic based on relevance features through max-pooling. Lower value indicates that the previous documents are more likely to be relevant to this subtopic. Note that if we view the signals produced by max-pooling (i.e. the vectors in “max-pooling” section of Figure 2) as a part of the general hidden states, our concrete implementation fit in DSSA framework.

We adopt an additive way to integrate both parts and then use softmax to produce (normalized) attention distribution:

$$\begin{aligned} a'_{t,k} &= \mathcal{A}'(\mathbf{h}_{t-1}, \mathbf{e}_{i_k}) + \mathcal{A}''(\mathbf{x}_{d_1, i_k}, \dots, \mathbf{x}_{d_{t-1}, i_k}), \\ a_{t,k} &= \frac{w_{i_k} \exp(a'_{t,k})}{\sum_{j=1}^K w_{i_j} \exp(a'_{t,j})} \quad (w_{i_j} \geq 0, \forall j). \end{aligned} \quad (8)$$

softmax is modified to include the initial subtopic importance w_{i_k} , which encodes our intuition that an important subtopic is more likely to gain attention than unimportant ones.

4.1.3 Scoring. The final score consists of relevance score $s_{d_t}^{\text{rel}}$ and diversity score $s_{d_t}^{\text{div}}$, which are combined by a coefficient λ :

$$s_{d_t} = (1 - \lambda)s_{d_t}^{\text{rel}} + \lambda s_{d_t}^{\text{div}} \quad (0 \leq \lambda \leq 1). \quad (9)$$

The relevance score and diversity score are calculated as follows:

$$\begin{aligned} s_{d_t}^{\text{rel}} &= \mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_q) + \mathbf{x}_{d_t, q}^\top \cdot \mathbf{w}^r, \\ s_{d_t}^{\text{div}} &= \mathbf{a}_{t, (\cdot)}^\top \cdot \begin{bmatrix} \mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_{i_1}) + \mathbf{x}_{d_t, i_1}^\top \cdot \mathbf{w}^r \\ \vdots \\ \mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_{i_K}) + \mathbf{x}_{d_t, i_K}^\top \cdot \mathbf{w}^r \end{bmatrix}, \end{aligned} \quad (10)$$

where $\mathbf{w}^r \in \mathbb{R}^R$ and $\mathbf{a}_{t, (\cdot)}$ is the attention derived from subtopic attention component. The diversity score is calculated as a weighted combination of the document’s relevance to each subtopic by attention distribution. We use the same way to calculate document’s relevance to a query and to its subtopics using both distributional representations and relevance features, although different ways can be used. Specifically, d_t ’s relevance to a query q (or a subtopic i_k) is calculated based on both the similarity between two distributed representations $\mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_q)$ (or $\mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_{i_k})$) and relevance features $\mathbf{x}_{d_t, q}$ (or \mathbf{x}_{d_t, i_k}). \mathcal{S}' intends to produce a matching score between two representations and \mathbf{w}^r linearly combines features. Similar to \mathcal{A}' , \mathcal{S}' could also be implemented as:

$$\mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_{i_k}) = \begin{cases} \mathbf{e}_{d_t}^\top \mathbf{W}^s \mathbf{e}_{i_k}, & (\text{general}) \\ \mathbf{e}_{d_t}^\top \cdot \mathbf{e}_{i_k}, & (\text{dot}) \end{cases} \quad (11)$$

where $\mathbf{W}^s \in \mathbb{R}^{E_d \times E_q}$. Then the score of a ranking r is calculated by summing up all the $|r|$ documents’ scores:

$$s_r = \sum_{t=1}^{|r|} s_{d_t}. \quad (12)$$

Vector interaction operations \mathcal{A}' and \mathcal{S}' could be implemented using more complex models, such as multilayer perceptron (MLP), to model the interaction between two vectors more accurately. We could also use convolutional neural network (CNN) instead of RNN to model the interaction among a sequence of documents and encode their information. We deliberately choose to use simple mechanisms in this implementation in order to show that the general framework is capable of capturing the essence of diversification even without complex operations. More complex implementations will be examined in future work.

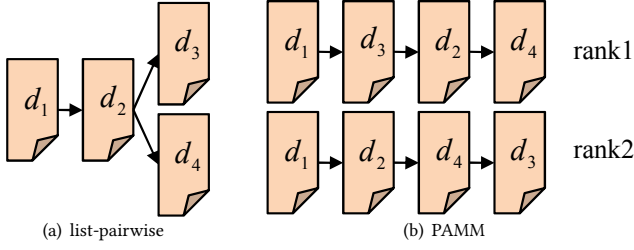


Figure 3: Pair sample examples of (a) list-pairwise and (b) PAMM. Both samples are positive.

4.2 A List-pairwise Approach for Optimization

Liu [19] classifies LTR approaches into three categories: pointwise, pairwise, and listwise. Search result diversification is naturally a listwise problem because the score of a document depends on the previous documents. Take Table 1 as an example, under no previous documents, d_2 is better than d_3 because d_2 covers one more subtopic (subtopics are of equal weight). However, given that we have selected d_1 , which is similar to d_2 while dissimilar to d_3 , d_3 becomes superior because it provides additional information.

4.2.1 List-pairwise Training. We propose a list-pairwise training approach. We call it list-pairwise because a sample in our algorithm consists of a pair of rankings (r_1, r_2) : r_1 and r_2 are totally identical except the last document. The sample can be written as (C, d^1, d^2) , where C is the shared previous document sequence. The pairwise preference ground-truth is generated based on an evaluation metric M , such as α -nDCG. If $M(r_1) > M(r_2)$, it is positive, otherwise it is negative. Our approach is similar to pairwise approaches because it aims to compare a pair of documents, but this is done within some context. Similarly to pairwise, the loss function can be defined as binary classification logarithmic loss:

$$L_{\text{list-pairwise}} = \sum_{q \in Q} \sum_{o \in O_q} w^{(o)} \left(y^{(o)} \log \left(P(r_1^{(o)}, r_2^{(o)}) \right) + (1 - y^{(o)}) \log \left(1 - P(r_1^{(o)}, r_2^{(o)}) \right) \right), \quad (13)$$

where O_q is all the pair samples of query q , $y^{(o)} = 1$ indicates positive and 0 for negative, and $P(r_1^{(o)}, r_2^{(o)})$ is the probability of being positive calculated by $\frac{1}{1 + \exp(s_{r_2^{(o)}} - s_{r_1^{(o)}})}$. To enhance effectiveness,

we weight pairs with $w^{(o)} = |M(r_1^{(o)}) - M(r_2^{(o)})|$, which means that the bigger the metric score gap, the more important the pair.

Because DSSA calculates document d 's score s_d^C based on previous document C , we could also use Maximum Likelihood Estimation (MLE) or PAMM to optimize our model. We use Plackett-Luce model [22] to estimate the probability of a ranking r :

$$P(r) = \prod_{i=1}^{|r|} \frac{\exp(s_{d_i}^{r[:i-1]})}{\sum_{j=i}^{|r|} \exp(s_{d_j}^{r[:i-1]})}, \quad (14)$$

where $r[:i-1]$ means the top $i-1$ documents of ranking r . Then the loss functions could be written as:

$$L_{\text{MLE}} = \sum_{q \in Q} -\log(P(r_q^+)), \quad (15)$$

Algorithm 1 A List-pairwise Approach For Optimization

```

1: procedure LIST-PAIRWISE TRAINING
   input: loss function  $L$ , learning rate  $r$ , epochs  $V$ , query set  $Q$ ,
   document set  $\mathcal{D}$ , evaluation metric  $M$ , random permutation
   count  $N$ 
   output: DSSA with trained parameters  $\theta$ 
2:   initialize  $\theta$ 
3:   for  $i$  from 1 to  $V$  do
4:     for batch  $b \in \text{GetSamples}(Q, \mathcal{D}, M, N)$  do
5:        $g \leftarrow \text{GetGradient}(L(b, \theta))$ 
6:        $\theta \leftarrow \theta - rg$ 
   return DSSA $_{\theta}$ 
7: procedure GETSAMPLES
   input: query set  $Q$ , document set  $\mathcal{D}_q$  for each query  $q$ , evaluation
   metric  $M$ , random permutation count  $N$ 
   output: a set of ranking pairs with weight and preference
    $\{(q^{(1)}, C^{(1)}, d_1^{(1)}, d_2^{(1)}, w^{(1)}, y^{(1)}), (q^{(2)}, C^{(2)}, d_1^{(2)}, d_2^{(2)}, w^{(2)}, y^{(2)}), \dots\}$ 
   include:  $\text{GetPerms}(\mathcal{D}_q, l, N, M)$  return a best ranking (under
   metric  $M$ ) and  $N$  random permutations of length  $l$ .
    $\text{GetPairs}(q, \mathcal{D}_q, C, M)$  samples pairs of documents  $(d_1, d_2)$ 
   from  $\mathcal{D}_q \setminus C$  under context  $C$  if and only if they lead to different
   metric scores. Let  $r_1 \leftarrow [C, d_1]$ ,  $r_2 \leftarrow [C, d_2]$ ,  $w = |M(r_1) - M(r_2)|$  and  $y = \mathbb{1}[M(r_1) > M(r_2)]$ .
8:    $\mathcal{R} \leftarrow \emptyset$ 
9:   for query  $q$  in  $Q$  do
10:    for  $l$  from 0 to  $|\mathcal{D}_q| - 1$  do
11:      for perm  $C$  in  $\text{GetPerms}(\mathcal{D}_q, l, N, M)$  do
12:         $\mathcal{R} \leftarrow \mathcal{R} \cup \text{GetPairs}(q, \mathcal{D}_q, C, M)$ 
   return  $\mathcal{R}$ 

```

$$L_{\text{PAMM}} = \sum_{q \in Q} \sum_{r_q^+, r_q^-} \mathbb{1}[P(r_q^+) - P(r_q^-) \leq M(r_q^+) - M(r_q^-)], \quad (16)$$

where $\mathbb{1}[\text{condition}]$ is 1 if the condition is satisfied, 0 otherwise, MLE maximizes the probability of positive rankings, and PAMM enlarges the probability margin between positive and negative rankings according to an evaluation metric. For MLE, the number of best rankings is usually small if we only have hundreds of queries, which may not be enough to train adequately the parameters. PAMM uses preferences between very different rankings that are not comparable (see Figure 3(b)). In contrast, list-pairwise method only allows the last document to be different (Figure 3(a)). This corresponds better to the decision-making situation in which we have to choose a document under a given context. It is expected that such a pair sample allows us to better train the ranking function. Experiments will show that our approach works better.

As shown in Figure 2, our architecture is a unified neural network and the attention function is continuous, so the gradient of the loss function can be backpropagated directly to train the model. We use mini-batch gradient descent to facilitate training process.

Unfortunately, it is impossible to acquire all the list-pairwise samples, which has in total $|\mathcal{D}_q|!$ ($|\mathcal{D}_q|$ is the number of candidate documents) different permutations. So we develop a sampling strategy similar to negative sampling [23] as described in Algorithm 1: for each query q , we sample a large number of pairs of rankings, whose length ranges from 1 to $|\mathcal{D}_q|$. We first obtain some contexts C from both best rankings and randomly sampled

negative rankings (rankings that are not optimal). Then under each C , a pair of documents (d^1, d^2) are sampled from the remaining documents $\mathcal{D}_q \setminus C$ if and only if they lead to different metric scores.

4.2.2 Prediction. In prediction stage, for each query, we sequentially and greedily choose the document with the highest score and append it to the ranking list. Specifically, the first document is selected under initial subtopic importance from the whole candidate set \mathcal{D}_q . Once the top $t - 1$ documents have been selected (i.e. $|C| = t - 1$), we feed each document in $\mathcal{D}_q \setminus C$ into DSSA at the t -th position one by one and choose the one with the highest s_{d_t} . This process continues until all the documents in \mathcal{D}_q are ranked.

4.2.3 Time Complexities. The training time complexity with vanilla cell and “general” operation is $O(V \cdot |Q| \cdot \Gamma \cdot |\mathcal{D}_q| \cdot \Theta)$ where V is the number of iterations, $|Q|$ is the number of training queries, $\Gamma = N \cdot |\mathcal{D}_q|^2$ is the number of sampled pairs where N is the number of random permutations, $|\mathcal{D}_q|$ is the number of candidate documents, and Θ is the complexity for each position:

$$\Theta = \underbrace{U(U + E_d)}_{\text{document sequence representation}} + \underbrace{KUE_q + KR + KE_dE_q + KR}_{\text{subtopic attention}}, \quad (17)$$

where the dominating terms are KUE_q and KE_dE_q which are proportional to the number of subtopics K . How to efficiently handle a large number of subtopics is our future work. The prediction complexity is $O(|\mathcal{D}_q|^2\Theta)$ for each query. We can limit $|\mathcal{D}_q|$ to a small number (say 50), so the prediction time can be reasonable.

5 EXPERIMENTAL SETTINGS

5.1 Data Collections

We use the same dataset as [16] which consists of Web Track dataset from TREC 2009 to 2012. There are 198 queries (query #95 and #100 are dropped because no diversity judgments are made for them), each of which includes 3 to 8 subtopics identified by TREC assessors. The relevance rating is given in a binary form at subtopic level. All experiments are conducted on ClueWeb09 [5] collection.

We use query suggestions of Google search engine as subtopics, which are released by Hu et al. [16] on their website¹. For DSSA, we only use the first level subtopics and leave the exploration of hierarchical subtopics to future work. Following the existing work [16], we simply use uniform weights for these subtopics.

5.2 Evaluation Metrics

We use ERR-IA [8], α -nDCG [10], and NRBP [11], which are official diversity evaluation metrics used in Web Track. They measure the diversity by explicitly rewarding novelty and penalizing redundancy. D \sharp -measures [26], the primary metric used in NTCIR Intent [25] and IMine task [20], is also included. In addition, we also use traditional diversity measures Precision-IA (denoted as Pre-IA) [1] and Subtopic Recall (denoted as S-rec) [37]. Consistent with existing works [32, 33, 39] and TREC Web Track, all these metrics are computed on top 20 results of a ranking. We use two-tailed paired t -test to conduct significance testing with p -value < 0.05 .

Table 5: Relevance features. Each of the first 3 features is applied to body, anchor, title, URL, and the whole documents.

Name	Description	#Features
TF-IDF	the TF-IDF model	5
BM25	BM25 with default parameters	5
LMIR	LMIR with Dirichlet smoothing	5
PageRank	PageRank score	1
#inlinks	number of inlinks	1
#outlinks	number of outlinks	1

Table 6: Diversity features. Each feature is extracted over a pair of documents.

Name	Description
subtopic diversity	euclidean distance based on SVD
text diversity	cosine-based distance on term vector
title diversity	text diversity on title
anchor text diversity	text diversity on anchor
link-based diversity	link similarity of document pair
URL-based diversity	URL similarity of document pair

5.3 Baseline Models

We compare DSSA² to various unsupervised and supervised diversification methods. The non-diversified baseline is denoted as **Lemur**. We use **xQuAD** [27], **PM2** [13], **TxQuAD**, **TPM2** [12], **HxQuAD**, and **HPM2** [16] as our unsupervised baselines. We use **ListMLE** [31], **R-LTR** [39], **PAMM** [32], and **NTN** [33] as our supervised baselines. Top 20 results of Lemur are used to train supervised methods. Top 50 (i.e. $|\mathcal{D}_q|$) results of Lemur are used for diversity re-ranking. To construct the representation of a query or a subtopic, we use the top 20 (Z) documents. We use 5-fold cross validation to tune the parameters in all experiments based on α -nDCG@20, which is one of the most widely used metrics. A brief introduction to these baselines is as follows:

Lemur. We use the same non-diversified results as [16] for fair comparison. They are produced by language model and retrieved using the Lemur service³ of which the spams are filtered. These results are released by Hu et al. [16] on the website¹.

ListMLE. ListMLE is a representative listwise LTR method without considering diversity.

xQuAD, PM2, TxQuAD, TPM2, HxQuAD, and HPM2. These are competitive unsupervised explicit diversification methods, as introduced in Section 2.2. All these methods use λ to control the importance of relevance and diversity. HxQuAD and HPM2 use an additional parameter α to control the weight of each layer of the hierarchical structure. Both λ and α are tuned using cross validation. They all require a prior relevance function to fulfill diversification re-ranking. Following [39], we use ListMLE.

R-LTR, PAMM, and NTN. For PAMM, we use α -nDCG@20 as the optimization metric. We optimize NTN based on both R-LTR and PAMM, denoted as R-LTR-NTN and PAMM-NTN respectively.

¹hierarchical search result diversification: <http://www.playbigdata.com/dou/hdiv>

²data and code available at: <http://www.playbigdata.com/dou/DSSA/>

³Lemur service: http://boston.lti.cs.cmu.edu/Services/clueweb09_batch/

To achieve optimal results, for R-LTR and PAMM, we tune the relational function $h_S(R)$ from minimal, maximal, and average. For PAMM, we tune the number of positive rankings τ^+ and negative rankings τ^- per query. For NTN, the number of tensor slices is tuned from 1 to 10. LDA is used to generate distributed representations of size 100 for NTN and DSSA. For all these supervised methods, the learning rate r is tuned from 10^{-7} to 10^{-1} . For DSSA, we have different settings possible. In our first set of results, we will use “general” as the implementation of vector interaction operations \mathcal{A}' and \mathcal{S}' , LSTM with hidden size of 50 as the cell of RNN. We set random permutation count as 10 in list-pairwise sampling. Similarly, λ of DSSA is tuned by cross validation. We also test the impact of different model settings and permutation counts on performance in Section 6.2 and Section 6.3 respectively.

Similar to [39], we implement 18 relevance features and 6 diversity features, as listed in Table 5 and 6 respectively. We collect the candidate and retrieved documents of all queries and subtopics to generate the distributed representations.

6 EXPERIMENTAL RESULTS

6.1 Overall Results

The overall results are shown in Table 7. We find that DSSA significantly outperforms all implicit and explicit baselines, including both unsupervised and supervised. The improvements are statistically significant (two-tailed paired t-test) for all metrics, except S-rec. The results clearly show the superiority of DSSA.

(1) DSSA vs. unsupervised explicit methods. **DSSA outperforms unsupervised explicit methods (xQuAD, PM2, TxQuAD, TPM2, HxQuAD, and HPM2) on all the measures.** The relative improvement over HxQuAD and HPM2, the best unsupervised explicit approaches, is up to 8.3% and 8.6% respectively in terms of α -nDCG. This comparison shows the great advantage of using supervised method for learning the ranking function.

(2) DSSA vs. supervised implicit methods. **DSSA also outperforms supervised implicit methods (R-LTR, PAMM, R-LTR-NTN, and PAMM-NTN) by quite large margins.** The improvement over R-LTR-NTN and PAMM-NTN, the best supervised implicit approaches is up to 9.9% and 9.4% respectively on α -nDCG. This result demonstrates the utility of taking into account subtopics explicitly in supervised approaches. The improvements are similar to those observed between explicit approaches and implicit approaches in unsupervised framework [12, 13, 16, 27]. The combination of the two observations suggests that explicit modeling of subtopics can improve result diversification, whether it is in a supervised or unsupervised framework.

6.2 Effects of Different Settings

We conduct experiments with different settings of DSSA to investigate whether the performance is sensitive to these settings. Different aspects of settings are listed follow. For simplicity, when investigating the impact of each aspect, we keep other aspects the same as the settings specified in Section 5.3.

- (1) Representation generation methods: SVD, LDA, and doc2vec with window size of 5.
- (2) Implementation of vector interaction operations \mathcal{A}' and \mathcal{S}' : “general” and “dot”.

Table 7: Performance comparison of all methods. The best result is in bold. Statistically significant differences between DSSA and baselines are marked with various symbols. ★ indicates significant improvement over all baselines.

Methods	ERR-IA	α -nDCG	NRBP	D $\#$ -nDCG	Pre-IA	S-rec
Lemur ^①	.271	.369	.232	.424	.153	.621
ListMLE ^①	.287	.387	.249	.430	.157	.619
xQuAD ^②	.317	.413	.284	.437	.161	.622
TxQuAD ^③	.308	.410	.272	.441	.155	.634
HxQuAD ^④	.326	.421	.294	.441	.158	.629
PM2 ^⑤	.306	.411	.267	.450	.169	.643
TPM2 ^⑥	.291	.399	.250	.443	.161	.639
HPM2 ^⑦	.317	.420	.279	.455	.172	.645
R-LTR ^⑧	.303	.403	.267	.441	.164	.631
PAMM ^⑧	.309	.411	.271	.450	.168	.643
R-LTR-NTN ^⑨	.312	.415	.275	.451	.166	.644
PAMM-NTN ^⑨	.311	.417	.272	.457	.170	.648
DSSA	.356★	.456★	.326★	.473★	.185★	.649^{⑩⑪}

Table 8: Effects of different settings.

Methods	ERR-IA	α -nDCG	NRBP	D $\#$ -nDCG	Pre-IA	S-rec
SVD	.348	.450	.315	.470	.184	.646
LDA	.356	.456	.326	.473	.185	.649
doc2vec	.351	.452	.318	.471	.184	.646
general	.356	.456	.326	.473	.185	.649
dot	.347	.450	.314	.470	.184	.647
vanilla	.354	.454	.322	.471	.184	.649
GRU	.357	.457	.326	.473	.185	.649
LSTM	.356	.456	.326	.473	.185	.649
DSSA-RNN	.342	.445	.306	.466	.172	.657
DSSA-RNNMP	.356	.456	.326	.473	.185	.649

(3) RNN cell: vanilla, GRU, and LSTM cell.

(4) Dimensionality: we test several representative settings on the size of distributed representations E_d and E_q , the size of hidden state U as (25, 10), (50, 25), (100, 50), (200, 100).

(5) Max-pooling: we experiment without using max-pooling (denoted as DSSA-RNN) in subtopic attention component.

The results are reported in Table 8. We can observe that DSSA does not heavily rely on specific settings. As for different representation generation methods, LDA has slightly better results. doc2vec could have been more appropriate if we had large datasets with more queries. The “general” operation yields slightly better results. A possible reason is that it is bilinear and thus is more powerful than “dot” to model the interaction. GRU and LSTM cells yield slightly better results than vanilla cell because of their ability of modeling long-term dependency. The difference is however small. This may be due to that with a limited number of training data, a model is unable to take advantage of its higher complexity to capture the fine-grained subtlety. Results with different size of distributed representation and hidden state shown in Figure 4(a) also

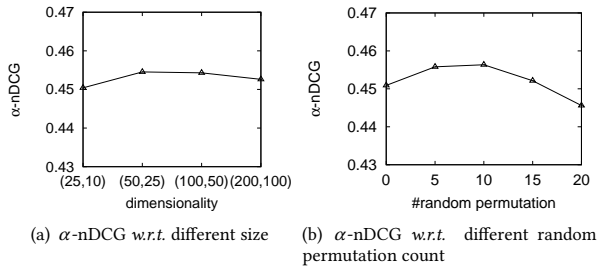


Figure 4: Performance tendency of different settings.

Table 9: Effects of different optimization methods.

Methods	ERR-IA	α -nDCG	NRBP	$D_{\#}$ -nDCG	Pre-IA	S-rec
MLE	.349	.446	.315	.462	.176	.644
PAMM	.348	.445	.315	.463	.175	.644
list-pairwise	.356	.456	.326	.473	.185	.649

indicate no strong correlation between performance and settings. α -nDCG remains above 0.45 using different sizes. The best performance is achieved using 100-dimensional representation and 50-dimensional hidden state. This suggests that high dimensionality may result in overfitting. Without using max-pooling, α -nDCG drops to 0.445, which demonstrates the usefulness of using max-pooling to enhance subtopic attention. The small differences between different settings suggest that DSSA is a stable and robust framework. Note that we use both distributed representations and relevance features, which are complementary to each other. This may be one of the reasons of the stability.

6.3 Effects of Different Optimization Methods

Results in Table 9 shows that list-pairwise is more effective than MLE and PAMM. This confirms our earlier intuition that list-pairwise optimization corresponds better to the situation of diversification ranking than the two other methods. Note that even using MLE or PAMM as optimization methods, DSSA could also achieve state-of-the-art performances, which confirms the effectiveness of our explicit learning framework from another perspective.

We vary the number of random permutations used in list-pairwise sampling from 0 to 20 to investigate its effect. As depicted in Figure 4(b), the performance does not heavily rely it. The best performance is achieved around 10. More permutations lead to lower effectiveness, which could be explained by model overfitting.

6.4 Visualization and Discussion

We visualize the ranking results of DSSA and the variation of subtopic attention to better understand why DSSA performs well.

We show the top 5 ranking results of query #58 and #182 in Figure 5 to illustrate why DSSA outperforms implicit learning methods. We choose PAMM-NTN as comparison method, which is the best existing learning method. In Figure 5, white means relevant and black means irrelevant. For query #58, DSSA ranks a document relevant to subtopics i_3 and i_4 first and a document relevant to i_1 and i_2 at the second position, while the first two documents of PAMM-NTN cover the same subtopics. Note that there is no

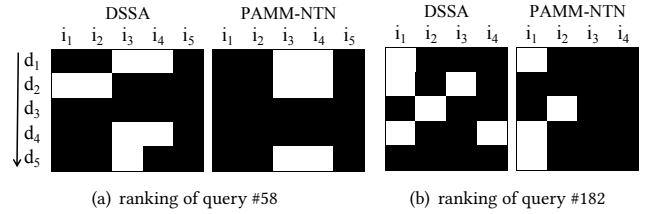


Figure 5: Case study for DSSA and PAMM-NTN. White means relevant and black means irrelevant.

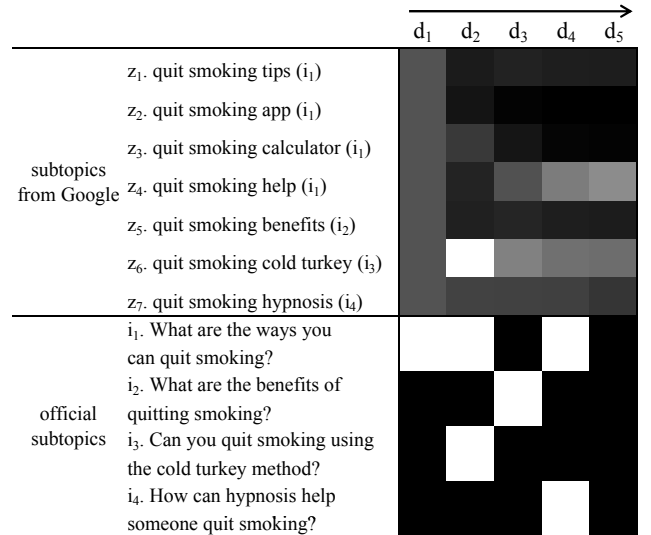


Figure 6: Subtopic attention variation of query #182. The top part is attention and the bottom part is relevance judgment.

document covering i_5 in the candidate set. For query #182, DSSA successively chooses documents that cover i_1 , i_3 , i_2 , and i_4 . One additional intent is satisfied at every position. PAMM-NTN, however, just covers i_1 and i_2 by top 5 documents, which is obviously not optimal. We see that the unequal and varied subtopic attention is capable of discovering unsatisfied subtopics at different positions, which eventually leads to more subtopic coverage.

To study attention mechanism, we further visualize the variation of subtopic attention of top 5 documents of query #182, namely “quit smoking”, which has 4 official subtopics (i_1 to i_4), as shown in Figure 6. The top part is subtopic attention variation and the bottom part is relevance judgment. For attention part, the darker the cell is, the lower the attention (weight) on this subtopic is. Note that we actually leverage query suggestions of Google (z_1 to z_7) to serve as subtopics, which do not match official ones exactly. We manually align subtopics mined from Google to official ones. At the beginning, all the subtopics have equal attention. The first selected document d_1 is relevant to i_1 , i.e. to the Goggle subtopics z_1 , z_2 , z_3 and z_4 . We see that the attention to these latter decreases at second position. Then the document d_2 is selected, which is relevant to uncovered i_3 . We see that the attention to the corresponding z_6 begins to diminish from the third position. d_3 and d_4 satisfy additional i_2 and i_4 respectively, which leads to the reduction

of attention on z_5 and z_7 at the following position. The subtopic attention, initialized as uniform distribution, ends up with more emphasis on z_4 , z_6 , and z_7 . This example illustrates how the unequal and varied attention drives the model to emphasize different subtopics at different positions, which is crucial in explicit diversification. This example also shows a potential problem inherent for any method using automatically discovered subtopics: those topics may be different from the ones defined by human assessors. Equal distribution is assumed on all the subtopics z_i . However, this implies an unequal distribution among the manually defined subtopics (more emphasis is put on i_1). Assuming an equal distribution at the beginning may not necessarily be the best approach. We will deal with this problem in our future work.

7 CONCLUSIONS

In this paper, we propose a general learning framework DSSA to model subtopics explicitly for search result diversification. Based on the sequence of selected documents, unequal and varied subtopic attention is calculated, driving the model to emphasize different subtopics at different positions. This is the first time that attention mechanism is used to model the process. We further instantiate DSSA using RNN and max-pooling to handle both distributed representations and relevance features, which outperforms significantly the existing approaches. The results confirm that modeling subtopics explicitly in a learning framework is beneficial and effective and this also avoids heuristically defined functions and parameters. However, accurately modeling the interaction among documents and subtopics is still challenging. There are many other more complex implementations besides our particular way, which will be investigated in future work. The proposed model contains a number of parameters to be learned. This requires a large number of training data. Collecting more training data to fully unlock the potential of the model is another direction. Finally, this work only deals with the learning of a ranking function, assuming that document and query representations have already been created. In practice, learning these representation is another interesting aspect, which could be incorporated into our framework, provided with sufficient training data.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author. This work was funded by the National Natural Science Foundation of China under Grant No. 61502501 and 61502502, the National Key Basic Research Program (973 Program) of China under Grant No. 2014CB340403, and the Beijing Natural Science Foundation under Grant No. 4162032.

REFERENCES

- [1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Jeong. 2009. Diversifying Search Results. In *WSDM*.
- [2] Ricardo A. Baeza-Yates, Carlos A. Hurtado, and Marcelo Mendoza. 2004. Query Recommendation Using Query Logs in Search Engines. In *Current Trends in Database Technology EDBT 2004 Workshops*.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014).
- [4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003).
- [5] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. Clueweb09 data set. <http://boston.lti.cs.cmu.edu/Data/clueweb09/>. (2009).
- [6] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *SIGIR*.
- [7] Ben Carterette. 2009. An Analysis of NP-Completeness in Novelty and Diversity Ranking. In *ICTIR*.
- [8] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *CIKM*.
- [9] Junyoung Chung, Çağlar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014).
- [10] 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 *SIGIR*.
- [11] Charles L. A. Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *ICTIR*.
- [12] Van Dang and Bruce W. Croft. 2013. Term Level Search Result Diversification. In *SIGIR*.
- [13] Van Dang and W. Bruce Croft. 2012. Diversity by Proportionality: An Election-based Approach to Search Result Diversification. In *SIGIR*.
- [14] Amac Herdagdelen, Massimiliano Ciaramita, Daniel Mahler, Maria Holmqvist, Keith Hall, Stefan Riezler, and Enrique Alfonseca. 2010. Generalized Syntactic and Semantic Models of Query Reformulation. In *SIGIR*.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997).
- [16] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search Result Diversification Based on Hierarchical Intents. In *CIKM*.
- [17] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*.
- [18] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *ICML*.
- [19] Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval* 3, 3 (2009).
- [20] Yiqun Liu, Ruihua Song, Min Zhang, Zhicheng Dou, Takehiro Yamamoto, Makoto P Kato, Hiroaki Ohshima, and Ke Zhou. 2014. Overview of the NTCIR-11 ITime Task. In *NTCIR*. Citeseer.
- [21] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP*.
- [22] John I Marden. 1996. *Analyzing and modeling rank data*. CRC Press.
- [23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.
- [24] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. In *NIPS*.
- [25] Tetsuya Sakai, Zhicheng Dou, Takehiro Yamamoto, Yiqun Liu, Min Zhang, Ruihua Song, MP Kato, and M Iwata. 2013. Overview of the NTCIR-10 INTENT-2 Task. In *NTCIR*.
- [26] Tetsuya Sakai and Ruihua Song. 2011. Evaluating Diversified Search Results Using Per-intent Graded Relevance. In *SIGIR*.
- [27] Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting Query Reformulations for Web Search Result Diversification. In *WWW*.
- [28] Rodrygo L.T. Santos, Craig Macdonald, Iadh Ounis, and others. 2015. Search result diversification. *Foundations and Trends® in Information Retrieval* (2015).
- [29] Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *SIGIR*.
- [30] Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *NIPS*.
- [31] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise Approach to Learning to Rank: Theory and Algorithm. In *ICML*.
- [32] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning Maximal Marginal Relevance Model via Directly Optimizing Diversity Evaluation Measures. In *SIGIR*.
- [33] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling Document Novelty with Neural Tensor Network for Search Result Diversification. In *SIGIR*.
- [34] Jeonghee Yi and Farzin Maghoul. 2009. Query Clustering Using Click-through Graph. In *WWW*.
- [35] Hai-Tao Yu and Fuji Ren. 2014. Search Result Diversification via Filling Up Multiple Knapsacks. In *CIKM*.
- [36] Yisong Yue and Thorsten Joachims. 2008. Predicting Diverse Subsets Using Structural SVMs. In *ICML*.
- [37] Cheng Xiang Zhai, William W. Cohen, and John Lafferty. 2003. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *SIGIR*.
- [38] Zhiyong Zhang and Olfa Nasraoui. 2006. Mining Search Engine Query Logs for Query Recommendation. In *WWW*.
- [39] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu. 2014. Learning for Search Result Diversification. In *SIGIR*.