基于双星型自注意力网络的搜索结果多样化方法

秦绪博[1], 窦志成[2,*], 朱余韬[3], 文继荣[4,5]
[1]中国人民大学信息学院
[2]中国人民大学高瓴人工智能学院
[3]蒙特利尔大学
[4]大数据管理与分析方法研究北京市重点实验室
[5]数据工程与知识工程教育部重点实验室
qratosone@live.com, dou@ruc.edu.cn

摘要

相关研究指出，用户提交给搜索引擎的查询通常为短查询。由于自然语言本身的特点，短查询通常具有歧义性，同一个查询可以指代不同的事物，或同一事物的不同方面。为了让搜索结果尽可能满足用户多样化的信息需求，搜索引擎需要对返回的结果进行多样化排序，搜索结果多样化技术应运而生。目前已有的基于全局交互的多样化方法通过全连接的自注意力网络捕获全体候选文档间的交互关系，取得了较好的效果。但由于此类方法只考虑文档间的相关关系，并没有考虑到文档是否具有跟查询相关的有效信息，在训练数据有限的条件下效率相对较低。该文提出了一种基于双星型自注意力网络的搜索结果多样化方法，将全连接结构改为星型拓扑结构，并嵌入查询信息以高效率地提取文档跟查询相关的全局交互特征。相关实验结果显示，该模型相对于基于全连接自注意力网络的多样化方法，具备显著的性能优势。

关键词： 搜索结果多样化 ； 文档全局交互 ； 自注意力网络

Search Result Diversification Framework Based on Dual Star-shaped Self-Attention Network

Xubo Qin[1], Zhicheng Dou[2,*], Yutao Zhu[3], Ji-Rong Wen[4,5]
[1] School of Information, Renmin University
[2] Gaoling School of Artificial Intelligence, Renmin University of China
[3] Université de Montréal
[4] Beijing Key Laboratory of Big Data Management and Analysis Methods,
[5] Key Laboratory of Data Engineering and Knowledge Engineering, MOE
qratosone@live.com, dou@ruc.edu.cn

Abstract

Research shows that most queries issued by users are short queries. Those queries may be ambiguous or broad for specifying users' actual information need. The same query issued by different users may lead to different information needs. In order to satisfy different users' needs, the search engines need to rank the search result document diversely. This leads to the search result diversification technology. For those previous diversification approaches based on global document interactions, fully-connected self-attention networks are used for capturing the interactions among all the candidate documents, leading to well performance. However, those models only take the document interactions into consideration, ignoring the relevance signals of candidate documents. As the annotated datasets are not enough for model training, the effectiveness of the model will be limited. In this paper we prove a new implicit diversification framework based on self-attention network in the shape of dual-star. A star-shaped topological structure is deployed with an external query node, in order to effectively extract the global interaction signals relevant to the query. Using self-attention to learn

第二十届中国计算语言学大会论文集，第280页-第292页，呼和浩特，中国，2021年8月13日至15日。

the similarity between each document, this model can be trained easily, and perform significantly better than the state-of-the-art implicit approach indicating the potential of implicit methods.

Keywords: Search Result Diversification , Global Document Interactions , Self-Attention Networks

1 引言

目前主流商业搜索引擎的工作方式通常为接受用户提交的查询，按照查询获取一系列相关文档，排序后将文档列表返回给用户。相关研究指出(Silverstein et al., 1999)，大多数用户在使用搜索引擎时习惯于提交短查询，而由于自然语言所具有的歧义性，同一个查询可能指代不同的事物。例如对于查询“苹果”，既可以表示“苹果手机”，也可以表示水果“苹果”。而针对同一个事物，用户也可能有不同方面的信息需求，例如搜索“人民大学”的用户可能会想要获得关于“人民大学招生”，“人民大学信息学院”或“人民大学教务处”等不同方面的结果。若搜索引擎只对文档进行相关性排序，则返回的结果可能会具有较大的冗余性，无法有效命中用户意图，满足用户多样化的信息需求，进而对用户体验产生负面影响。为了解决这一问题，需要应用搜索结果多样化技术，对返回的结果文档进行多样化排序，使得排序靠前的文档尽量覆盖较多的用户信息需求。例如在购物网站搜索“苹果”，搜索引擎返回的靠前的结果应当既包括“苹果手机”相关结果，也包括“水果苹果”相关结果，以满足不同用户的潜在信息需求。

近年来随着Learning-to-Rank相关技术的发展，学者们开始将监督式学习相关技术引入到搜索结果多样化任务，用以取代原有的人工指定的排序函数(Jun and Yanyan, 2016)。按照这一标准，可以将多样化模型分为启发式模型和学习式模型（也叫非监督式和监督式模型）。而按照是否显式地区分用户意图，多样化方法也可以分为隐式多样化方法与显式多样化方法，以上两种分类的方法是正交的，同一个模型既可以是启发式或学习式方法，也可以是隐式或显式方法(Dou et al., 2019)。

较早期的隐式多样化方法是最大边界相关性（Maximal Margin Relevance，MMR）(Carbonell and Goldstein, 1998)模型，后续的各种隐式与显式多样化方法大部分都继承了它的基本思想和方法。对于基于MMR思想的隐式多样化模型而言，一个候选文档的新颖性体现在它与已选中文档的不相似性，候选文档与已选中的文档越不相似，则文档新颖性越高。相对于隐式多样化方法，显式多样化方法主要关注候选文档所对应的用户意图（通常用子话题表示），一个好的候选文档应当尽可能地覆盖此前没有被覆盖到的子话题。

以MMR为代表的隐式多样化方法通常使用人工指定的多样化特征和评分函数，随着Learning-to-rank相关技术的发展，有学者将监督式学习方法引入多样化排序任务(Zhu et al., 2014)，在人工指定特征的基础上尝试使用机器学习方法自发地学习出最优化的评分函数；在此基础上，近年来也有学者提出利用深度学习相关技术，让模型可以自发地学习到最优化的多样化特征(Xia et al., 2016)和评分函数(Xia et al., 2015)。以上所有隐式多样化方法，都假设候选文档序列中的每一个候选文档彼此间都相互独立，仅考虑单个候选文档与已选定文档间的相关关系，通过贪心选择算法反复选出当前最佳的候选文档，进而得到全局最佳的候选文档序列。但是相关研究已经证明(Feng et al., 2018)，候选文档之间并不是彼此相互独立的，当一个候选文档被选中之后，其他候选文档带来的边际信息收益也将随之改变，忽略候选文档间的相关关系将难以得到全局最优解。

近年来，序列模型在排序学习领域取得了一系列应用。此前已有学者开始将序列模型应用于相关性排序学习领域，例如基于循环神经网络（RNN）和注意力（Attention）机制的DLCM(Ai et al., 2018)。随着Vaswani et al. (2017)首先提出了完全基于自注意力（Self-Attention）机制的Transformer模型并取得成功，学者们受到启发，开始将自注意力网络（Self-Attention Network，SAN）引入基于序列的相关性排序任务中。在相关性排序领域，较为典型的工作包括SetRank(Pang et al., 2020)和DIN(Pasumarthi et al., 2020)等，此后Qin et al. (2020)首先将基于自注意力网络的序列模型用于搜索结果多样化任务中，提出了基

于Transformer编码器的DESA框架。与已有的基于贪心选择的算法不同，该框架接受整个候选文档序列作为输入，全局地捕获文档间的交互关系并返回全体文档的排序评分。相对于已有的基于贪心选择的多样化算法，该模型可以捕获全体候选文档彼此之间的全局交互关系，避免每一次选择局部最优解的局限性。

目前基于全局交互关系的多样化排序模型主要依赖于Transformer编码器，即基于全连接的SAN结构。这一类方法存在以下局限性：一方面，已有的基于SAN结构的多样化排序模型本质上继承了隐式多样化排序的MMR思想，即认为一个文档的多样性主要由文档的新颖性决定。此种方法的问题在于多样化排序的根本目的在于满足用户意图，而此前的工作在捕获文档全局交互特征时并没有将文档本身与查询之间的相关性纳入考量。显然，一个具备新颖性，但是与查询不相关的搜索结果文档是不能满足用户的实际需求的，除了显式的相关性得分加权之外，多样化模型本身也应当考虑到文档与查询的相关性。另一方面，文档在输入的初始排序中的位置信息本身也表征了文档对查询的相关性特征，而基于Transformer结构的SAN模型本身具有排序无关性(Pang et al., 2020)，没有包含显式引入的归纳偏置信号，只能完全依靠位置编码来表征文档在初始排序中的位置信息。由于搜索结果多样化任务可用训练数据相对较少，基于Transformer结构的SAN模型在缺乏大量数据训练的情况下，很难有效地利用文档初始排序的位置信息带来的先验的文档相关性特征信号。因此，需要提出一种紧凑高效的网络结构，可以有针对性地捕获文档间与查询相关的全局交互信号，并充分利用文档的初始排位信息所包含的相关性特征。

本文在Guo et al. (2019)提出的Star-Transformer模型的启发下，提出了一种基于双星形SAN结构的搜索结果多样化方法，称为Query-Transformer。该方法在星型SAN结构中加入一个新的查询结点，形成双星形结构。相对于此前基于全连接SAN结构的多样化方法，该模型可以在衡量文档全局交互信息的过程中，充分地考虑到每一个候选文档本身与查询的相关性特征，进而高效而准确地捕获候选文档间与查询相关的有效交互信息。此外，相比全连接SAN结构，该模型使用的星型拓扑结构本身即可充分利用序列的局部组合性（Local Compositionality）带来的先验归纳偏置，进而显式地增强先验的文档位置信息带来的相关性特征信号，更加充分地利用规模较小的多样化标注数据集。基于TREC Web Track数据集的相关实验结果证明，该模型性能显著地领先于此前业界最佳（state-of-the-art）的基于全连接SAN的隐式多样化方法，与业界最佳的显式多样化方法相当。

2 相关研究

2.1 基于贪心选择的搜索结果多样化方法

按照是否显式考虑用户意图，主流的搜索结果多样化方法可以分为隐式多样化方法与显式多样化方法。隐式多样化方法通常聚焦于当前候选文档和已选定文档间的关系，认为一个候选文档与已选定的文档越不相似，则它的新颖性（Novelty）越高，对当前结果文档序列的多样化就越有利。Carbonell and Goldstein (1998)首先提出了MMR（Maximal Margin Relevance）的思想，将文档的多样化评分定义为文档对查询的相关性与文档的新颖性评分的线性组合，这一思想可以用下述公式表示：

$$\mathrm{MMR} = \arg\max \left[\lambda P(d_i|q) - (1-\lambda)\max_{d_j \in S}\left[\sum P(d_i|d_j)\right]\right], \quad (1)$$

式中$P(d_i|q)$表示当前文档d_i与查询q相关的概率，$P(d_i|d_j)$表示d_i与d_j相似的概率，$\lambda \in [0,1]$是调节两者比重的参数。对于MMR模型，当前文档与已选中的文档越不相似，则其新颖性越好，文档评分越高。后续的一系列隐式多样化模型通常继承了MMR的思想，即使用相关性与新颖性的线性组合表征文档的评分，主要将改进的着眼点放在如何更好地表征文档间的相关关系上。

随着Learning-to-rank相关技术的发展，学者们开始着眼于引入机器学习技术，使用排序学习方法抽取文档间的不相似性。典型的隐式多样化模型包括R-LTR(Zhu et al., 2014), SVM-DIV(Yue and Joachims, 2008)，PAMM(Xia et al., 2015)和PAMM-NTN(Xia et al., 2016)等。相对于隐式多样化方法，显式多样化方法关注文档对具体的用户意图的满足情况。显式多样化方法的基本思想是，多样化排序的根本目的是尽可能让排序靠前的文档满足用户的潜在信息需求，因此算法应当关注文档对不同的用户意图的满足情况，这类用户意图通常用子话题（Subtopic）表示。对于显式多样化模型，一个好的候选文档应该尽可能地覆盖当前还没有

覆盖的子话题，满足新的潜在用户意图。典型的方法包括非监督式的IA-Select(Agrawal et al., 2009)、PM2(Dang and Croft, 2012)、xQuAD(Santos et al., 2010)、HxQuAD/HPM2(Hu et al., 2015)等模型，以及监督式学习的DSSA(Jiang et al., 2017)和DVGAN(Liu et al., 2020)等模型。

上文提到的所有多样化方法都只考虑单个候选文档与已选中文档间的相似性关系，通过贪心选择算法反复选择当前最佳的候选文档，并不考虑全体候选文档间的相关关系。对于贪心选择算法，若子空间彼此相互独立，则每次都选择局部最优解最终可以导向全局最优解，但是实际上，相关研究(Feng et al., 2018; Qin et al., 2020)已经证明候选文档彼此并不是相互独立的关系，对一个候选文档的选择会对其他候选文档的新颖性效应产生影响，因此基于候选文档独立性假设的贪心选择算法不可能在合理的时间内达成全局最佳的排序。

2.2 自注意力网络

自注意力网络（Self-Attention Network，SAN）结构是近年来的研究重点——相对于传统的CNN和RNN等网络结构，SAN不仅更适合并行计算，而且可以更好地抽取长距离依赖关系，显著地克服CNN和RNN对于较长序列的处理能力不足的问题。Vaswani et al. (2017)首先提出了完全基于SAN的Transformer模型，其在神经机器翻译任务上的性能显著地领先于已有模型。随后，学者们提出了大量基于Transformer的模块化结构的模型。由于Transformer是一套完整的编码器-解码器（Encoder-Decoder）结构，后续的模型通常将Transformer的编码器或者解码器单独提取出来，作为单独的网络结构使用。典型的例子包括BERT(Devlin et al., 2019)、Transformer-XL(Dai et al., 2019)、ERNIE(Zhang et al., 2019)等。由于Transformer的编码器端本身规模较大，难以适用于小规模数据集，Guo et al. (2019)提出了基于星形拓扑结构的自注意力网络的Star-Transformer模型。传统的Transformer编码器端基于全连接SAN结构，参数总量较多。Star-Transformer则是一种围绕中心结点进行自注意力交互的星形SAN结构，周围结点仅与相邻结点和中心结点进行交互，从而显著降低了总体计算代价。相关实验已证明，Star-Transformer可以获得与全连接Transformer相仿的长距离依赖学习能力，其星型拓扑结构相对于全连接Transformer结构可以更加有效地利用序列的局部组合性（Local Compositionality）带来的先验归纳偏置，在小规模数据集上的效果优于全连接Transformer结构。

2.3 基于序列模型的多变量排序方法

传统的排序模型通常只单独地计算每一个候选文档的排序评分。对于相关性排序，每个候选文档的评分由文档对查询的相关性特征计算得到；而对于多样化排序，候选文档评分由单个候选文档相对于已选文档序列的新颖性特征计算得到。近年来学者们开始引入序列模型，在排序中引入候选文档之间的交互关系。较早期的成果是Ai et al. (2018)的DLCM，其基于RNN和注意力（Attention）机制。在Transformer取得成功之后，Pang et al. (2020)将基于Transformer编码器的SAN结构引入排序任务中，提出了SetRank模型。在多样化排序任务中，Qin et al. (2020)首先提出了DESA框架，利用SAN结构来衡量全体候选文档之间的全局交互关系，不依赖贪心选择过程即可获得多样化排序后的文档序列。相对于已有的基于贪心选择的多样化排序方法，DESA模型可以从全局角度衡量所有候选文档彼此之间的交互关系，克服贪心选择算法无法获得全局最优解的局限性。

在搜索结果多样化排序任务中，已有的多变量排序方法通常基于全连接SAN结构而实现。相对于RNN等已有的模型结构，全连接的SAN结构允许序列中的每一个候选文档都与其他候选文档进行直接交互，且不受长距离依赖问题的限制。但全连接SAN结构几乎完全没有引入对输入序列的任何先验的归纳偏置（Inductive Bias），充分发挥其效力需要较多的训练数据支撑——由于搜索结果多样化任务中可用的数据相对较少，全连接SAN结构在多样化排序任务中难以利用初始排序本身的位置信息，容易面临过拟合等问题。此外，已有的基于SAN的多样化方法使用文档的嵌入式（Embedding）向量作为网络的输入，通过SAN捕捉到表征文档新颖性的全局交互信号，这一全局交互信号是针对全部的文档内容的，缺乏对文档与查询相关的交互信号的针对性捕获。

3 基于双星形自注意力网络的多样化框架

本节将介绍基于星形自注意力网络的Query-Transformer方法的整体框架结构，及星形自注意力网络的具体细节。为了简化问题，此处我们将Query-Transformer实现为一个隐式的

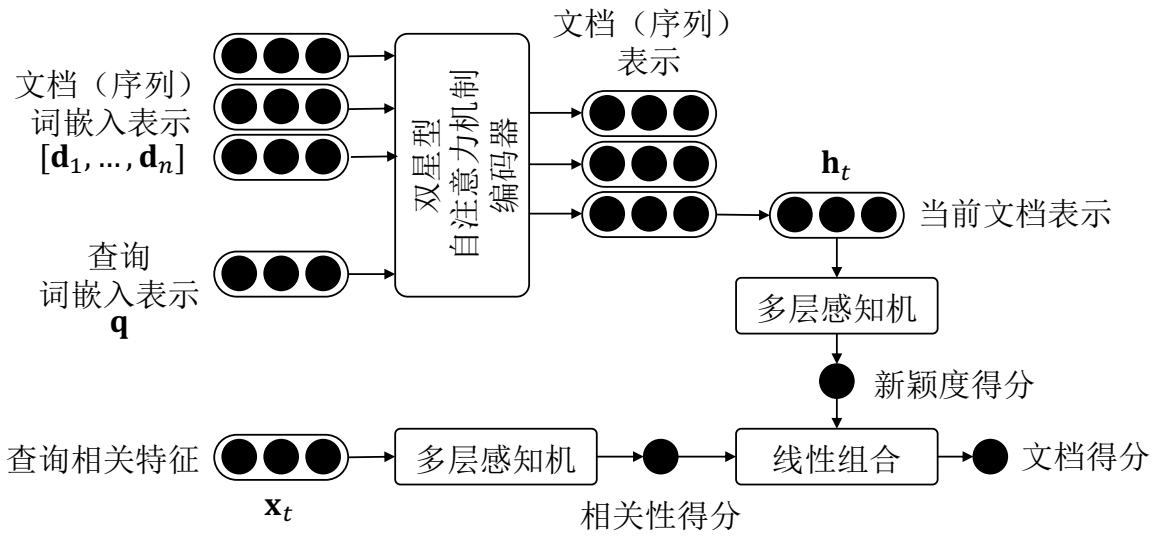


Figure 1: Query-Transformer整体框架结构

多样化模型，即不依赖于外部的子话题信息。在后续工作中，我们将参照DESA，把Query-Transformer框架扩展为同时融入显式与隐式多样化特征的融合式多样化方法。

与MMR的基本思想一致，Query-Transformer模型也将文档的评分视为相关性和新颖性的线性组合。文档的相关性评分由一系列文档对查询的传统信息检索相关特征通过线性Learning-to-rank函数计算得到，新颖性评分由自注意力网络生成的文档多样性相关表示经过线性函数计算得到，将两者线性组合即得到文档评分。自注意力网络接受查询向量和全体候选文档向量组成的序列作为输入，返回相同长度的的隐藏状态（Hidden State）向量，作为全体候选文档的多样性相关的文档表示。对于每一个文档，将其文档表示输入给线性Learning-to-rank函数，即可获得其多样性评分。将新颖性评分与多样性评分线性组合之后，即可获得文档的最终得分。图 1展示了Query-Transformer模型的整体结构。对于模型的训练，模型将整个文档序列中所有文档求和，作为文档序列的评分，用于计算损失函数。当执行排序任务时，模型直接返回所有文档的评分用于排序。图 1展示了Query-Transformer模型的整体结构。

3.1 总览

作为隐式多样化框架，Query-Transformer模型接受长度为n的候选文档序列的嵌入式表示$\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_n]$，其与查询的相关性特征$[\mathbf{x}_1, \cdots, \mathbf{x}_n]$，以及给定查询的嵌入式向量$\mathbf{q}$作为输入，计算返回每一个文档的排序评分$s_i (i \in [1, n])$。参照Jiang et al. (2017)的工作，我们使用以下方式构造文档和查询的表示：相关性特征$\mathbf{x}_i$使用18个典型的传统信息检索相关性特征指标，例如BM25，TF-IDF等；文档嵌入式表示$\mathbf{d}_i$使用doc2vec(Le and Mikolov, 2014)在全体数据集上生成；查询嵌入式表示$\mathbf{q}$通过伪文档方式构造，即使用传统信息检索系统按指定查询在全体数据集上搜索，将相关性排名前Z个的文档首尾相连，最后使用doc2vec生成文档向量，作为查询的嵌入式向量使用。受限于篇幅，此处我们不展开介绍这些特征的细节，相关内容可在(Jiang et al., 2017)中查阅。如上文所述，Query-Transformer模型继承了MMR的基本思想，将每一个文档的多样化评分定义为文档的相关性评分和新颖性评分的线性组合，可用如下公式表示：

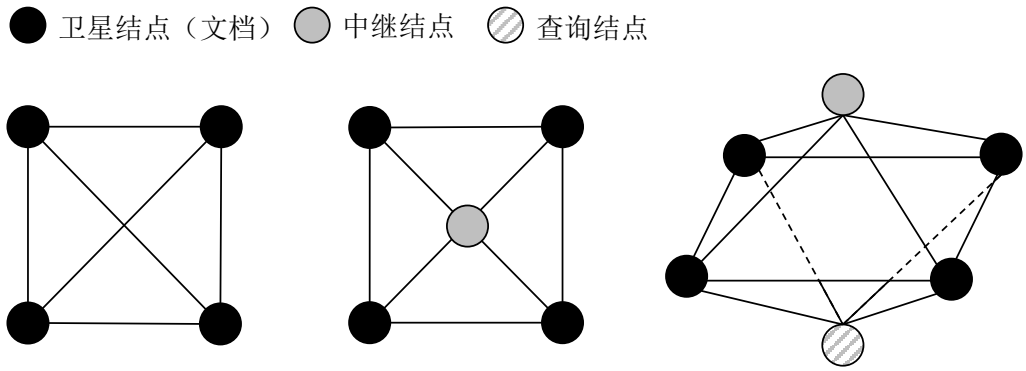
$$s_i = \lambda \mathbf{x}_i{}^\top \mathbf{W}_r + (1 - \lambda) \mathbf{h}_i{}^\top \mathbf{W}_h, \tag{2}$$

式中$\mathbf{x}_i$为文档d_i相对于查询q的相关性特征，$\mathbf{h}_i$表示文档d_i经由双星形SAN结构生成的文档表示，$\mathbf{W}_r$和$\mathbf{W}_h$均为可训练的参数。下文将详细介绍$\mathbf{h}_i$的计算过程。

3.2 嵌入查询的双星形自注意力网络

3.2.1 星形自注意力网络结构

本节将介绍Query-Transformer模型使用的双星形自注意力网络（SAN）结构。对于长度为n的文档序列，网络中包含1个中继结点(Relay Node)，1个查询结点(Query Node)和n个卫星



(1) 自注意力机制　(2) 星形自注意力机制　(3) 双星形自注意力机制

Figure 2: 三种自注意力网络结构对比

结点(Satellite Node)，此处每一个卫星结点对应一个输入的文档表征向量。n个卫星结点首尾相连形成环形结构，每一个卫星结点只与中继结点、查询节点和自己直接相邻的两个卫星结点进行交互，中继结点的作用是让彼此间不相邻的卫星结点也可以进行交互，查询节点的作用是提取文档中跟查询相关的交互特征。图2展示了Transformer、Star-Transformer和Query-Transformer结构所对应的三种自注意力网络结构的对比。在Query-Transformer网络结构中，全体卫星结点与中继节点构成一个星形拓扑的网络结构，卫星结点与查询结点构成另一个星型结构，因此得名"双星形自注意力网络"。

3.2.2 嵌入查询的双星形SAN结构的实现

(1) 缩放点积注意力函数（Scaled Dot Product Attention Function）：自注意力网络的核心部分是自注意力函数，其形式表现为缩放点积注意力（Scaled Dot product）函数。该函数接受一个查询（Query）向量[0]，并与键-值（Key-Value）向量分别进行计算，最终得到输出。由于注意力函数是基于序列计算的，因此上述向量计算均以矩阵计算的形式进行。缩放点积注意力函数的公式表述如下：

$$\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}})\mathbf{V}. \tag{3}$$

此处d为向量的维度数，$\sqrt{d}$为缩放因子（Scaled Factor），用于控制点积计算的值的数量级。

(2) 多头注意力机制（Multi-Head Attention）：基于Transformer的经验，类似于CNN的多通道机制，引入多头注意力（Multi-Head Attention，MHA）可以更有效地学习文档的不同方面。对于k个头的多头注意力，可用以下公式表述：

$$\mathrm{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = (\mathbf{a}_1;\cdots;\mathbf{a}_k)\mathbf{W}^{O}, \tag{4}$$

$$\mathbf{a}_j = \mathrm{Attn}(\mathbf{Q}\mathbf{W}_{\mathrm{j}}^{\mathrm{Q}}, \mathbf{K}\mathbf{W}_{\mathrm{j}}^{\mathrm{K}}, \mathbf{V}\mathbf{W}_{\mathrm{j}}^{\mathrm{V}}), \quad \mathrm{j} \in [1,\mathrm{k}]. \tag{5}$$

此处所有$\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V, \mathbf{W}^O$均为可学习的参数，[;]为向量连接操作。对于自注意力网络，此处有$\mathbf{Q}=\mathbf{K}=\mathbf{V}$，即查询-键-值向量在经过线性变换之前的原始输入均相等。在搜索结果多样化任务中，$\mathbf{Q}=\mathbf{D}$，此处$\mathbf{D}$为整个文档序列。即函数接受一个文档序列作为输入，返回每一个文档对应的自注意力生成隐藏状态（Hidden State）组成的矩阵。

(3) 位置编码机制：Query-Transformer模型是一个重排序模型，它接受一个纯相关性排序的搜索结果序列，返回多样化重排序结果。由于自注意力网络并不会显式地编码文档的顺序信息，而初始的候选文档顺序直接反映了文档的相关性，因此基于自注意力网络的模型通常会引入额外的位置编码（Positional Encoding）。尽管Query-Transformer的星型拓扑结构引入了针对局部组合性的归纳偏置，可以一定程度上增强输入文档序列的位置信息，但这一归纳偏置仍

[0]需要说明的是，此处所述的"查询"指的是注意力函数的查询向量，与信息检索领域用户提交的查询无关。

然无法完全取代位置编码信号。因此我们参照Star-Transformer的做法，引入一个训练式的嵌入向量，用于表征不同的位置信息。该位置编码向量将直接与输入向量相加。

(4) 结点更新：模型接受原始的文档向量$\mathbf{d}_i(i \in [1,n])$作为卫星结点$\mathbf{h}_i$的初始值，中继结点初始化为全体文档向量的平均值，即：

$$\mathbf{h}_i^0 = \mathbf{d}_i, \tag{6}$$

$$\mathbf{s}^0 = \text{mean}(\mathbf{d}_1, \cdots, \mathbf{d}_\text{n}), \tag{7}$$

符号中的上标数字t表示第t层的结点表示，$t=0$表示初始化时对应的结点值。与全连接SAN结构相同，星形SAN结构也可以拥有多个自注意力网络层，以增强模型的学习能力。

对于每一层网络，各结点的更新方式如下：

- 对于每一个卫星结点，定义一个上下文（Context）矩阵$\mathbf{C}_i^t$，然后对当前结点和上下文矩阵应用多头注意力机制，并使用ReLU激活函数和层标准化（Layer Normalization）机制：

$$\mathbf{C}_i^t = [\mathbf{h}_{i-1}^{t-1}; \mathbf{h}_i^{t-1}; \mathbf{h}_{i+1}^{t-1}; \mathbf{s}^{t-1}; \mathbf{q}; \mathbf{d}_i], \tag{8}$$

$$\mathbf{h}_i^t = \text{LayerNorm}(\text{ReLU}(\text{MHA}(\mathbf{d}_i, \mathbf{C}_i^t, \mathbf{C}_i^t))), \tag{9}$$

  此处$\boldsymbol{C}_i^t$包含以下内容：当前卫星结点在上一层的值$\mathbf{h}_i^{t-1}$；与之左右相邻的卫星结点$\mathbf{h}_{i-1}^{t-1}$和$\mathbf{h}_{i+1}^{t-1}$的值；上一层的中继结点值$\mathbf{s}^{t-1}$；查询结点值$\mathbf{q}$；以及当前卫星结点的原始嵌入向量$\mathbf{d}_i$。中继结点的作用是允许非相邻的卫星结点通过中继结点发生交互，而查询结点则提取文档中跟查询相关的有效内容，避免不相关文档的特征对全局交互结果产生干扰。

- 所有卫星结点更新完成之后，对中继结点进行更新：

$$\mathbf{s}^t = \text{LayerNorm}(\text{ReLU}(\text{MHA}(\mathbf{s}^{t-1}, [\mathbf{s}^{t-1}; \mathbf{H}^t], [\mathbf{s}^{t-1}; \mathbf{H}^t]))), \tag{10}$$

$$\mathbf{H}^t = [\mathbf{h}_1^t; \cdots; \mathbf{h}_n^t], \tag{11}$$

  此处$\mathbf{H}^t$表示由全体卫星结点$\mathbf{h}_i^t$构成的矩阵。只有中继结点会随卫星结点更新而更新，查询结点$\mathbf{q}$的值保持不变。

- 重复以上过程，直到每一层网络的权重值都更新完毕。

3.3 模型训练与推断

3.3.1 List-Pairwise优化方法

为了解决训练样本不足的问题，我们使用Jiang et al. (2017)提出的List-pairwise方法来构造训练样本并定义损失函数。List-pairwise方法的基本思想是，使用前缀文档序列C，在其末尾附加正例文档d_1和负例文档d_2，得到正例序列$[C, d_1]$和负例序列$[C, d_2]$。此处简要描述训练样例的采样过程：C为包含一系列不同长度的前缀文档序列，既包括基于人工标注数据生成的理想排序，也包括随机生成的排序。在已选文档序列的基础上，依次遍历全体候选文档并将候选文档一一附加到前缀文档序列的末尾。当且仅当正例文档d_1和负例文档d_2生成的排序可以导向不同的评价指标时，将正例序列$[C, d_1]$和负例序列$[C, d_2]$纳入为一对训练样本。

模型训练的优化目标与Pairwise损失函数相同，即为尽可能增大正例和负例间的差距，提高模型区分新颖文档和冗余文档的能力。损失函数可用以下公式表示：

$$\mathcal{L} = \sum_{o \in O} |\Delta M| [y^o \log(P(r_1^o, r_2^o)) + (1 - y^o) \log(1 - P(r_1^o, r_2^o))], \tag{12}$$

式中O表示全体训练样例的集合，$y^o = 1$表示正例，$y^o = 0$表示负例，$P(r_1^o, r_2^o) = \text{sigmoid}(r_1^o - r_2^o$表示正例评分$r_1^o$大于负例评分$r_2^o$的概率。此处文档序列的评分$r$为该序列中所有文档评分之和。$|\Delta M|$表示这一对正负例的权重，计算方式为正例与负例的评价指标之差，通常选用α-nDCG作为评价指标。

需要说明的是，对于前缀C中的文档$d_c \in C$，当序列末尾附加的文档改变时，d_c的注意力权重变化较小，生成的文档表示$\mathbf{h}_c$受到的影响也相对较小。因此，正负例序列$[C, d_1]$和$[C, d_2]$的评分之差$(r_1 - r_2)$将主要体现为附加文档(d_1, d_2)的评分之差$(s_1 - s_2)$，即对于正负样本(r_1^o, r_2^o)，可以近似地认为$P(r_1^o, r_2^o) \approx P(d_1^o, d_2^o)$。

综上所述，尽管(d_1, d_2)的多样性评分受到前缀序列C的影响，但List-pairwise方法的损失函数仍然属于针对正负例文档对的Pairwise损失函数，而非Listwise损失函数。从损失函数的角度来看，模型优化的目标是尽可能地提升模型区分新颖文档和冗余文档的能力，让模型可以分辨出候选文档序列中的每一个文档相对于全体其他候选文档的新颖性。

3.3.2 文档排序过程

在训练阶段，模型将返回文档序列中所有文档的和，用于计算损失函数。模型预测阶段，模型将返回全体文档的评分，用于对文档进行排序。继承了DESA的全局交互多样化排序的思路，Query-Transformer模型可以同步地完成全体候选文档的排序，不需要进行贪心文档选择。

由于自注意力网络具有排序无关性的特点，在相关性排序领域，SetRank等工作具备不依赖于初始排序信息的“集合（Set）到序列”能力。但Query-Transformer不具备此类能力。这是由于Query-Transformer使用星形拓扑结构的SAN结构。与全连接的SAN结构不同，星形拓扑的SAN结构本身会显式地衡量卫星结点彼此间的顺序，并且使用训练式位置编码来显式地衡量输入文档序列的位置信息——即Query-Transformer显式地依赖于输入的初始排序的位置信息。在多样化排序任务中，输入的初始序列是经过相关性排序的文档序列，因此此类初始位置信息可以直接表征文档对查询的相关性。在Query-Transformer模型中，此类基于位置信息的相关性信号起到的作用与查询结点类似，即初始文档序列的输入顺序表征了文档的相关性特征，将相关性特征融入文档全局交互过程，可以进一步捕获文档与查询相关的有效信息，改善多样化效果。本文第4.2节将展示相关实验结果以证明这一结论。

4 实验结果与分析

4.1 实验设置

本文使用的实验数据为TREC WebTrack 2009到2012的多样化评测数据集，共198个带子话题层面相关性标注的查询可用。由于时间有限，本文使用的部分基线实验结果来自于已发表的文献，这些实验结果与本文基于完全相同的数据集。为便于比较，本文使用Jiang et al. (2017)在GitHub上公布的相关性特征与文档嵌入式表示，包括18个传统信息检索指标（例如BM25、TD-IDF等）和使用doc2vec(Le and Mikolov, 2014)生成的文档向量。在今后的工作中，我们将尝试引入近年来被提出的基于深度文本匹配的相关性特征抽取工具，如K-NRM(Xiong et al., 2017)和BERT-IR 等。对于DSSA模型，为了进行公正的比较，此处我们使用相同的doc2vec作为文档的嵌入表示，而非Jiang et al. (2017)报告的通过LDA生成的文档向量，这一结果标记为DSSA （doc2vec）。本文使用的doc2vec文档向量维度为100，设置多头注意力的头（Head）数为5，即每一个头的维度为20。

本文使用的评价指标包括WebTrack官方的评价指标ERR-IA(Chapelle et al., 2009)、α-nDCG(Clarke et al., 2008)、NRBP(Baeza-Yates et al., 2005)、Pre-IA(Agrawal et al., 2009)和S-rec(Zhai et al., 2015)，以前20个文档的排序得分为准。模型在全体查询上进行5-折交叉验证，并使用α-nDCG@20为主要评价指标进行调参。我们使用一块NVIDIA Titan V GPU进行训练，完整的训练和5-折交叉验证过程可以在3小时内完成。

本文使用的基线模型包括Lemur服务提供的基于语言模型的非多样化搜索结果（实验表格中标注为Lemur），典型的非监督式显式多样化模型xQuAD(Santos et al., 2010)，PM2(Dang and Croft, 2012)和HxQuAD/HPM2(Hu et al., 2015)，典型的监督式隐式多样化模型R-LTR(Zhu et al., 2014)，PAMM(Xia et al., 2015)，PAMM-NTN(Xia et al., 2016)，以及监督式显式多样化模型DSSA(Jiang et al., 2017)。以上方法都是基于贪心文档选择的搜索结果多样化方法。对于基于全局交互的搜索结果多样化方法，我们使用DESA的隐式方法作为对照基线，按Qin et al. (2020)在其工作中提及的隐式方法参数对模型进行设置。在我们的实验结果列表中，该模型标注为DESA（Imp.）

4.2 实验结果和讨论

表1展示了主要模型的对照实验结果，加粗数字表示经过显著性检验$p < 0.05$，即

Table 1: 主要多样化模型性能对照

模型	ERR-IA	α-nDCG	NRBP	Pre-IA	S-rec
Lemur	.271	.369	.232	.153	.621
xQuAD	.317	.413	.284	.161	.622
PM2	.306	.411	.267	.169	.643
HxQuAD	.326	.421	.294	.158	.629
HPM2	.317	.420	.279	.172	.645
R-LTR	.303	.403	.267	.164	.631
PAMM	.309	.411	.271	.168	.643
R-LTR-NTN	.312	.415	.272	.166	.644
PAMM-NTN	.311	.417	.272	.170	.648
DSSA (doc2vec)	.350	.452	.318	.184	.645
DESA (Imp.)	.344	.445	.311	.177	.648
Query-Transformer	**.354**	**.454**	**.322**	.182	**.653**

Table 2: 自注意力网络层数及类型对模型性能的影响

模型	ERR-IA	α-nDCG	NRBP	Pre-IA	S-rec
Query-Transformer ($L = 1$)	**.354**	**.454**	**.322**	.182	**.653**
$L = 2$	.349	.450	.315	.177	.653
$L = 3$	.348	.447	.314	.169	.649
Transformer ($L = 3$)	.344	.446	.311	.177	.648
Star-Transformer ($L = 1$)	.348	.449	.316	.177	.650

本文提出的Query-Transformer显著地领先于当前业界最佳的隐式监督式模型，与显式监督式模型DSSA的性能相仿。此处所指"业界最佳的隐式监督式模型"包括基于文档选择的PAMM-NTN模型和基于全局文档交互的隐式DESA模型。对于最为典型的评价指标α-nDCG@20，Query-Transformer相对于PAMM-NTN性能领先9%，相对于隐式DESA模型性能领先2%。这一结果证明了我们的模型所具有的优越性。相对于基于贪心选择的PAMM-NTN模型，我们的模型可以有效地学习到全体候选文档间的全局交互特征，而相对于基于全局交互的隐式DESA模型，我们的模型相比全连接SAN结构，在有限的训练数据集下拥有更好的性能。

接下来，本文对自注意力网络的超参数和类型带来的影响进行了进一步的探索。表2上方展示了自注意力网络层数L的影响。理论上自注意力网络层数越多，模型分辨文档的新颖性的效果越好。但由于训练数据总量有限，若层数太多则可能出现过拟合的问题，实验表明当$L = 1$时Query-Transformer的效果达到最佳。作为对照，DESA在隐式条件下的最优层数$L = 3$，由此可以证明Query-Transformer使用的星型网络结构可以使用更低的计算代价，针对性地捕获文档间的全局交互关系。

表2下方展示了不同类型的自注意力网络模型在多样化任务上的表现，将框架中的自注意力网络结构由嵌入查询的Query-Transformer结构更换为其他自注意力网络结构，相关性特征部分保持不变。参与对比的包括原始的Transformer编码器结构（记为Transformer）和原始的Star-Transformer结构（记为Star-Transformer）。对于Transformer和Star-Transformer，根据其在五折交叉验证中的结果，分别设置$L = 3$和$L = 1$以获得最佳性能。实验结果显示，由于自注意力函数的性质与MMR的隐式多样化思想高度吻合，三者在多样化任务上都有较好的性能，显著地领先于完全基于文档贪心选择策略的PAMM-NTN模型。但是由于多样化数据集总量有限，星形的Star-Transformer和Query-Transformer网络结构性能领先于全连接的原始Transformer网络结构；在此基础上，嵌入查询的Query-Transformer性能领先于未经修改的Star-Transformer网络结构。这一结果可以证明本文提出的双星形网络结构中的查询结点在多

Table 3: 排序结果和覆盖的用户意图对照表，表头中的“(Trans.)”表示原始Transformer模型，“(Q-Trans.)”表示Query-Transformer模型

	#62		#106		#123	
排序位置	Q-Trans.	Trans.	Q-Trans.	Trans.	Q-Trans.	Trans.
1	i_3, i_4	-	i_2, i_3	-	i_1,i_2, i_3	-
2	i_3	i_3	-	-	-	i_1,i_2, i_3
3	-	-	-	i_3	i_1,i_2, i_3	-
4	i_3	i_3	-	-	i_1,i_2, i_3	-
5	i_3	i_3	-	-	-	-

Table 4: 输入文档序列顺序对不同自注意力网络的影响

模型	原始输入	打乱输入
Query-Transformer	.454	.438
Transformer	.446	.441

样化排序任务中能够起到积极的作用。

为了进一步探索Query-Transformer使用的双星形网络结构在多样化排序中发挥的作用，我们对照原始Transformer模型（即DESA隐式变种），对模型在测试集查询上的排序结果进行了逐查询的分析。我们将注意力放在最为典型的α-nDCG评价指标上，并重点关注排序前五个文档的结果。我们选取WebTrack数据集中的编号#62，#106和#123三个查询作为典型例子——表3列出了两种模型输出的排序结果中的前五个文档对子话题的覆盖情况，表中列举了候选文档经过Transformer和Query-Transformer两种模型的多样化重排序之后，位于列表前五的文档所覆盖的用户意图编号。 上述实验结果可见，相对于原始Transformer模型，Query-Transformer模型生成的重排序列表，覆盖用户意图的文档相对较多，且覆盖用户意图总数也较多。显然，若一个文档与查询相关，则该文档一定覆盖了至少一个用户意图，而一个没有覆盖任何用户意图的文档必然是不相关的文档。由于参与对照的模型都为隐式模型，没有外部输入的显式子话题信息，因此这一结果可以证明Query-Transformer使用的双星形SAN结构，可以更好地将文档的相关性特征融入文档全局交互中，提升多样化排序性能。

除此之外，我们进一步探索了输入文档序列的顺序对模型性能带来的潜在影响。对于Query-Transformer和Transformer，我们将待排序的文档序列随机打乱之后输入到模型中，观察这一操作对模型性能（基于α-nDCG评价指标）的影响，实验结果如表 4。可以看到，打乱输入文档顺序使得Transformer的性能降低约%2，而Query-Transformer的性能降低约%4。这一结果可以证明Query-Transformer对输入的待排序文档序列的顺序更加敏感。正如我们在第 1节所述，全连接Transformer结构不包括对输入序列位置信息的归纳偏置，而Query-Transformer使用的星形拓扑结构则可以显式地利用局部组合关系，进而更有效地利用输入序列的位置信息。由于多样化排序任务是一个重排序任务，输入的文档序列是经过相关性排序的序列，其顺序表征了文档对查询的相关性，因此这一结果可以证明Query-Transformer可以更有效地利用输入文档序列的位置信息所表征的相关性特征信号，进而对多样化任务进行增强。

理论上多样化排序模型应当尽可能地覆盖不同用户意图的文档，但多样化排序的根本目的在于满足用户意图。由于实际参与排序的文档并不一定能恰好满足不同的用户意图，模型应当首先保证有尽可能多的文档覆盖尽可能多的用户意图，然后再让文档覆盖多样化的不同用户意图——一个“多样化”但不相关的排序结果并不能真正满足用户需求。引入相关性信号的目的即在于此：若一个文档满足了某一个用户意图，那么该文档一定是与查询相关的文档，融入相关性信号可以让多样化模型尽可能地对实际满足了某个用户意图的文档返回更高的评分。尽管Query-Transformer模型返回的排序结果依然存在用户意图冗余的情况，但由于排序结果提供的与用户意图相关的文档总数较多，且覆盖的用户意图数量也更多，因此生成的排序结果依然

优于原始Transformer模型。

5 结论

本文提出了一种基于双星形自注意力网络结构的隐式搜索结果多样化模型Query-Transformer，该模型是一种基于全局文档交互的多样化排序方法，通过引入查询结点和星型拓扑结构，可以将文档的相关性信号引入文档全局交互中，让模型可以更高效地衡量与查询相关的文档彼此间的全局交互特征。实验结果证明，相对于此前的基于全连接自注意力网络结构的方法，该模型结构紧凑，可以更加高效率地学习与查询相关的文档交互关系，并充分地利用输入文档序列的位置信息所表征的相关性特征信号，用于增强多样化任务。实验结果证明Query-Transformer模型性能显著地领先于此前的基于文档选择或文档全局交互的隐式多样化方法。未来潜在的改进方向之一是以现有的模型为基础，继续探索将子话题信息融入框架的方法，实现隐式多样化与显式多样化结合。另一方面，模型目前只考虑了候选文档彼此间的关系，没有考虑到候选文档与已选文档间的关系。未来的另一个改进方向是进一步优化模型结构，让模型可以综合考虑已选文档与候选文档间关系和候选文档彼此间关系。

致谢

感谢各位匿名评审老师提出的修改建议。本论文受国家自然科学基金会项目（编号61872370, 61832017）和北京市杰出青年科学基金项目（BJJWZYJH012019100020098）资助。

参考文献

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14.

Qingyao Ai, Keping Bi, Jiafeng Guo, and W. Bruce Croft. 2018. Learning a deep listwise context model for ranking refinement. In *The 41st International ACM SIGIR Conference*, SIGIR '18, page 135–144, New York, NY, USA. Association for Computing Machinery.

Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. 2005. Query recommendation using query logs in search engines. In Wolfgang Lindner, Marco Mesiti, Can Türker, Yannis Tzitzikas, and Athena I. Vakali, editors, *Current Trends in Database Technology - EDBT 2004 Workshops*, pages 588–596, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference*, SIGIR '98, pages 335–336.

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, pages 621–630.

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*, SIGIR '08, pages 659–666.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July. Association for Computational Linguistics.

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*, SIGIR '12, pages 65–74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June.

Zhicheng Dou, Xubo Qin, and Ji-Rong Wen. 2019. A survey on search result diversification. *Chinese Journal of Computer*, 42(12):2592–2613.

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 *The 41st International ACM SIGIR Conference*, SIGIR '18, page 125–134.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1315–1325, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 63–72.

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*, SIGIR '17, pages 545–554.

Xu Jun and Lan Yanyan. 2016. Diversification: A new direction of learning-to-rank. *Communications of CCF*, 12(7):50–52.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org.

Jiongnan Liu, Zhicheng Dou, Xiaojie Wang, Shuqi Lu, and Ji-Rong Wen. 2020. Dvgan: A minimax game for search result diversification combining explicit and implicit features. SIGIR '20, page 479–488, New York, NY, USA. Association for Computing Machinery.

Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen. 2020. Setrank: Learning a permutation-invariant ranking model for information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference*, SIGIR '20, page 499–508.

Rama Kumar Pasumarthi, Honglei Zhuang, Xuanhui Wang, Michael Bendersky, and Marc Najork. 2020. Permutation equivariant document interaction network for neural learning to rank. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ICTIR '20, page 145–148, New York, NY, USA. Association for Computing Machinery.

Xubo Qin, Zhicheng Dou, and Ji-Rong Wen, 2020. *Diversifying Search Results Using Self-Attention Network*, page 1265–1274. Association for Computing Machinery, New York, NY, USA.

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, pages 881–890.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, September.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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*, SIGIR '15, pages 113–122.

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*, SIGIR '16, pages 395–404.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 55–64, New York, NY, USA. Association for Computing Machinery.

Yisong Yue and Thorsten Joachims. 2008. Predicting diverse subsets using structural svms. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1224–1231.

ChengXiang Zhai, William W. Cohen, and John Lafferty. 2015. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. *SIGIR Forum*, 49(1):2–9, June.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. *CoRR*, abs/1905.07129.

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*, SIGIR '14, pages 293–302.