计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(7):204−209 [doi: 10.15888/j.cnki.csa.008114]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于对抗式数据增强的深度文本检索重排序①

陈丽萍, 任俊超

(东北大学 理学院, 沈阳 110819)
通讯作者: 任俊超, E-mail: renjc@mail.neu.edu.cn

摘　要: 在信息检索领域的排序任务中, 神经网络排序模型已经得到广泛使用. 神经网络排序模型对于数据的质量要求极高, 但是, 信息检索数据集通常含有较多噪音, 不能精确得到与查询不相关的文档. 为了训练一个高性能的神经网络排序模型, 获得高质量的负样本, 则至关重要. 借鉴现有方法 doc2query 的思想, 本文提出了深度、端到端的模型 AQGM, 通过学习不匹配查询文档对, 生成与文档不相关、原始查询相似的对抗查询, 增加了查询的多样性, 增强了负样本的质量. 本文利用真实样本和 AQGM 模型生成的样本, 训练基于 BERT 的深度排序模型, 实验表明, 与基线模型 BERT-base 对比, 本文的方法在 MSMARCO 和 TrecQA 数据集上, *MRR* 指标分别提升了 0.3% 和 3.2%.

关键词: 神经网络排序模型; 稠密表征; 信息检索; 对抗式数据增强; 生成模型

引用格式: 陈丽萍,任俊超.基于对抗式数据增强的深度文本检索重排序.计算机系统应用,2021,30(7):204–209. http://www.c-s-a.org.cn/1003-3254/8114.html

Deep Text Retrieval Re-Ranking Based on Adversarial Data Augmentation

CHEN Li-Ping, REN Jun-Chao

(School of Science, Northeastern University, Shenyang 110819, China)

Abstract: The neural network ranking model has been widely used in the ranking task of the information retrieval field. It requires extremely high data quality; however, the information retrieval datasets usually contain a lot of noise, and documents irrelevant to the query cannot be accurately obtained. High-quality negative samples are essential to training a high-performance neural network ranking model. Inspired by the existing doc2query method, we propose a deep and end-to-end model AQGM. This model increases the diversity of queries and enhances the quality of negative samples by learning mismatched query document pairs and generating adversarial queries irrelevant to the documents and similar to the original query. Then, we train a deep ranking model based on BERT with the real samples and the samples generated by the AQGM model. Compared with the baseline model BERT-base, our model improves the *MRR* index by 0.3% and 3.2% on the MSMARCO and TrecQA datasets, respectively.

Key words: neural network ranking models; dense representation; information retrieval; adversarial data augmentation; generative model

1　引言

近年来, 随着互联网技术的不断发展, 信息检索领域的相关研究也取得了巨大突破. 信息检索是用户获取查询信息的主要方式. 信息检索主要是解决从大量候选信息集中获取与所需信息相关的信息资源, 返回的相关信息通常根据某种相关性概念进行排名, 排名的结果至关重要. 因此, 对于信息检索领域中排序模型的研究成为一大热点.

① 收稿时间: 2020-11-09; 修改时间: 2020-12-12, 2021-02-02; 采用时间: 2021-02-08; csa 在线出版时间: 2021-06-30

在过去的几十年中, 研究人员提出了许多不同的排序模型, 包括向量空间模型[1], 概率模型[2] 和 LTR (Learning To Rank) 模型[3]. 最高效的检索方法是使用向量空间模型, 其方法包括 TF-IDF 关键词权重匹配[4], 这些方法是基于词的匹配, 更容易受到关键词的限制. 例如, 候选段落集合为: “这个女明星如此漂亮”, “莉莉”. 查询问题为: “这个女明星叫什么名字? ”, 如果采用 TF-IDF 关键词匹配方法, 这个查询的最相关答案是“这个女明星如此漂亮”. 显然, 此方法只能获得与查询相似的段落, 无法得到语义的匹配信息. 随着机器学习的发展, *LTR* 模型已经取得了巨大的成功, 其主要取决于特征的选取, 特征的质量和数量决定了模型的质量, 一定程度上缓解了向量空间模型带来的不足, 但是依然无法获取连续的语义信息. 近年来, 神经网络发展快速并且在自然语言处理领域取得了重大突破, 神经网络模型被应用到信息检索领域中, 神经网络信息检索模型[5] 被证明可以有效地从原始输入中学习文本的潜在语义信息, 并能一定程度上解决了词语不匹配问题. 越来越多的研究者探索双重编码结构[6], 该结构使用原始查询和检索到的段落作为输入文本, 通过 one-hot 编码, Word2Vec[7] 和 sub-word components[8] 等词嵌入方法来表示文本. 神经网络模型在信息检索应用中得到了广泛推广, 例如段落检索、文档检索和段落排序.

信息检索领域中的段落排序任务, 其主要框架如图 1 所示. 段落排序任务通常分为两个阶段: 第 1 阶段是使用简单高效的检索模型来快速检索候选段落集合中的 Top-*k*; 第 2 阶段是使用相对复杂的排序模型对 Top-*k* 候选集进行重排序. 本文主要针对第 2 阶段, 研究查询结果重排序问题. 段落重排序不需要二次检索, 仅基于原始检索结果进行重排序. 当前主流搜索引擎: Google, Baidu, Bing, 在输入查询时将返回一系列查询结果, 但用户提交的查询词太短或太长, 查询字词无法准确表示用户的意图. 查询和查询结果之间存在语义上的不匹配, 并且无法获得用户所需的查询结果. 此外, 用户通常只关注排名最高的搜索结果, 而排名较高的搜索结果可能包含不相关的文档. 该排序结果直接影响用户体验. 因此, 如何提高排名的准确性并提高用户对查询结果的满意度一直是搜索引擎中的重点研究问题.

利用神经网络排序模型对候选段落集合重新排序. 神经网络排序模型对于数据质量要求极高, 而信息检索数据集中含有较多噪音, 并且缺少大量的标签数据, 无法准确获取与查询不相关的文档. 训练一个能理解查询意图的深度学习模型是困难的.

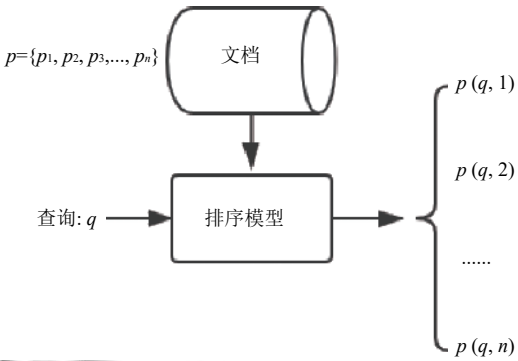

图 1 段落排序

本文目标是通过不匹配的查询文档对以提高排序模型的性能. 针对信息检索数据集不平衡且缺乏高质量的负样本这一问题, 我们提出一个深度的、端到端的生成模型, 利用不匹配的查询文档对, 生成与查询相似、文档不相关的对抗查询. 最后, 利用新构造的数据训练深度排序模型.

实验结果显示, 与基线模型 BERT-base 相比, MSMARCO 和 TrecQA 数据集利用本文方法 AQGM+BERT-base 在 MRR 指标上分别提升了 0.3%、3.2%. 由此可见, AQGM (Adversarial Query Generation Model) 算法通过生成的对抗查询, 增强了负样本质量, 使 BERT-base 分类模型更具鲁棒性.

2 相关工作

文本检索重排序任务早前的方法更多是利用简单的方法衡量查询和文档相关性, 最高效的衡量方法为 BM25[9], 其核心思想是对于查询中的每个词语, 计算和当前文档的相关性得分, 然后对查询中所有词语和文档的相关得分进行加权求和, 得到最终得分. 这种方法简单高效, 但是查询结果往往不那么准确. 研究人员对于初步相关性得分进一步探索, 其中一种思路是利用文档和文档之间的关系进行检索结果重排序, Plansangket 等[10] 提出基于文档分类实现检索结果重排序, 降低了一些分类分数较低的查询结果的等级. Balinski 等[11] 利用从文本、超链接获得的文档之间的距离关系来提高文档得分. 其次, 借助外部语料信息进行文档重排序, Qu 等[12] 利用每个主题构造单独的词库来生成文档向

量和查询向量.

随着深度学习的快速发展, 深度学习模型已经在各大领域展示了卓越的性能, 更多研究者将端到端的模型应用到排序任务中. 例如 DRMM[13], 在查询词级别时使用联合深层架构进行相关性匹配; matchpyramid[14], 基于 word-leval 级别交互的矩阵匹配思想, 用一个二维矩阵来代表 query 和 doc 中每个 word 的交互, 能捕获更精确的匹配信息, 以及目前最流行的预训练模型 BERT[15].

尽管深度模型性能强大, 但是对于数据的质量要求极高. 研究者主要从两方面对数据进行一系列加强: 一是查询扩展, Voorhees[16] 提出的词汇查询扩展, 利用同义词的信息,利用统计转换建模词语关系[17], 以及 Browm 大学提出的 query2query[18] 方法, 将原始查询利用生成查询进行扩充, 这些技术都是通过加强查询理解来匹配更好的文档; 二是文档扩充, doc2query[19], 利用文档生成其查询, 对文档进行扩充. 这些方法都是通过丰富文档信息, 来得到最好的匹配结果.

近年来, 对抗生成网络[20] 发展迅速, 应用到图像、语音和文本等领域, 从噪音数据中生成样本, 极大提高模型的鲁棒性. Wang 等人提出 IRGAN[21], 将对抗生成网络应用到信息检索排序任务中, 基于极小极大化博弈理论, 使得分类判别器将真正的样本与生成的对抗性样本尽可能准确区分. Bahuleyan 等人提出 VED 模型[22], 增加生成句子的多样性. Nguyen 等人提出 QUARTS 模型[23], 增强了模型的稳健性和泛化能力.

本文工作受到生成对抗性样本方法的启发, 通过生成对抗查询的方式, 增加了查询多样性, 并构造高质量的负样本对, 利用当前主流深度模型 BERT, 训练分类模型, 得到查询文档对的相关性得分.

3 方法

3.1 *AQGM* 模型

本文提出 AQGM (Adversarial Query Generation Model) 方法, 通过生成对抗查询, 得到高质量的负样本对, 对数据进行增强. 该方法由 3 部分组成: (1) 基于词的权重方法获得不相关的查询文档对. (2) 通过 VED (Variational Encoder-Decoder) 模型, 生成与查询相似、文档不相关的对抗查询, 模型结构如图 2. (3) 得到最终增强的样本$\{Q+(1-y)Q_gen, P, y\}$. 其中, y 代表查询文档对是否匹配. 当 y=1, (Q,P) 是真正的正例; y=0, $(Q+Q\_gen, P)$ 是通过对抗查询增强的负例.

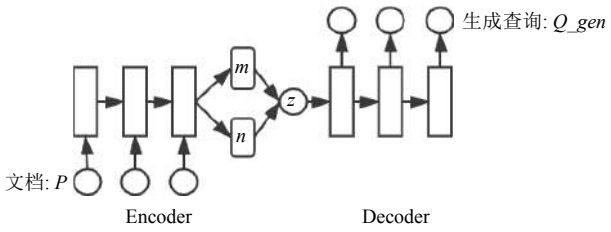

图 2 VED 模型结构

对于每一个 query, $Q_i=\{q_i\}$, 都存在一个对应 passage 集合$D_i=\left\{p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-\right\}$. 其中, p_i^+代表与 query 相关的段落, $p_{i,j}^-$代表与 query 不相关的段落. 基于词的权重方法获得不相关文档集中与查询得分最高的文档, 我们定义查询文档对的相关性得分:

$$score\left(q_i, p_{i,j}^-\right)=\sum_t^k w_t R\left(q_{it}, p_{i.j}^-\right) \tag{1}$$

其中, w_t 代表单词的权重, q_{it} 代表 q_i 中的单词, $p_{i,j}^-$表示某个搜索文档, R 代表每个单词和搜索文档的相关性高低, 此处我们采用传统的 *IDF* 来定义 w_t:

$$w_t=\log\frac{N-n(q_{it})+0.5}{n(q_{it})+0.5} \tag{2}$$

其中, N 为索引中的全部文档数, 我们用 $n(q_{it})$ 指代包含 q_{it} 的文档数. 接下来是单词 q_{it} 与段落$p_{i,j}^-$的相关性得分衡量:

$$R\left(q_{it}, p_{i.j}^-\right)=\frac{f_t\cdot(k_1+1)}{f_t+K}\cdot\frac{qf_t\cdot(k_2+1)}{qf_t+k_2} \tag{3}$$

$$K=k_1\left(1-b+b\frac{pl}{avgpl}\right) \tag{4}$$

其中, k_1, k_2, b 为调节因子, b 对文档长度因子进行调整, k_2 对查询词频因子进行调整, k_1 对文档词频因子进行调整. k_1+1, k_2+1 保证查询词频及文档词频大于 1. 从 K 的定义可得, b 越大, K 值越小, 文档长度对相关性得分的影响越大, 反之越小; 而文档的相对长度越长, K 值将越大, 则相关性得分会越小. k_1 按经验设置为 2, k_2 通常为 0–1000, b 设置为 0.75[24]. f_t 为 q_{it} 在$p_{i,j}^-$中出现的频率, qf_t 为 q_{it} 在 q_i 中出现的频率, pl 为段落$p_{i,j}^-$的长度. $avgpl$ 为所有段落的平均长度.

通过以上方法可以得到查询和不相关文档的得分, 选择其中得分最高的样本对 $(q,p-)$, 作为训练样本, 利用 VED 模型进行训练, 模型结构如图 2. 其中编码器 encoder 输入文档 P, 并输出平均向量 m 和偏差向量 n, m 和 n 作为 z 的后验正态分布的参数; 解码器根据从

后验分布提取的样本 z 生成查询 Q_gen. 本文通过模型生成的查询, 得到高质量的负样本, 构造了增强的数据集: $\{Q+(1-y)Q\_gen, P, y\}$, 其中 Q 代表原始查询.

3.2 深度排序模型

通过深度排序模型得到查询文档的相关性得分. 本任务建立基于 BERT 的分类模型. 训练样本为三元组的格式:

$$C = (q_i, p_i, R_i) \tag{5}$$

其中, R_i 代表 passage 是否是 query 的正确回答, 取值为 0 或 1, 我们通过 Pointwise 的训练方式建立 query 和 passage 的关系. 具体的, 我们将 query q_i 和 passage p_i 拼接成一个序列输入, 如式 (6):

$$X_i = [< CLS >, q_i, < SEP >, p_i] \tag{6}$$

其中, <*SEP*>表示分隔符, <*CLS*>的位置对应的编码表示 query 和 passage 的关系.

利用 BERT 对其进行编码,训练一个二分类网络:

$$y_i = softmax(h_\theta(q_i, p_i)) \tag{7}$$

经过 BERT 编码后, 我们取最后一层的<*CLS*>位置的隐向量$h_\theta(q_i, p_i)$作为 query 和 passage 的关系表示. 然后通过 *softmax* 计算最终得分 y_i. 后续我们通过改进的交叉熵损失函数来优化我们的模型:

$$L = -[\beta_1 R_i \cdot \log(y_i) + (1 - R_i) \cdot \log(1 - y_i)] \tag{8}$$

其中, β_1为调节因子, 取值大于 1. 通过调节因子的设置, 使得模型对正样本的错误预测给予更多的关注.

4 实验设置

4.1 数据集

实验采用了两个基准数据集, 分别为 MSMARCO[25] 和 TrecQA[26]. 数据集的统计信息如表 1 所示.

MSMARCO 是由微软提出的基于大规模真实场景数据的数据集, 该数据集基于 Bing 搜索引擎和 Cortana 智能助手中的真实搜索查询产生. MSMARCO 数据集包括约 880 k 的不重复 passage, 约 101 k 的 query. Query 平均长度为 5.97, passage 平均长度为 56.58. 我们采用的测试集为 2019 trec 比赛释放的人工标注好的 9260 条数据.

表 1 数据集统计信息

Datasets	Train (k)	Dev (k)	Test (k)
MAMARCO	880	—	9.3
TrecQA	53	1.1	1.51

TrecQA 是由 Wang 等人提供的基准数据集, 是从 TrecQA 的第 8–13 轨道收集, 由真实的问题组成, 主要回答“谁”、“什么”、“哪里”和“为什么”等类型的问题.

4.2 评价指标

信息检索排序问题常用评价指标有 *MRR*、*MAP* 和 *NDCG*. 在 *MAP* 中, 文档和查询要么相关, 要么不相关, 也就是相关度非 0 即 1. *NDCG* 中做出改进, 相关度分成从 0 到 r 的 r+1 个等级 (r 可设定). 根据实验数据集的特性, MSMARCO 测试数据集的相关度分为 0 到 3, TrecQA 测试数据集的相关度非 0 即 1. 因此, 对于 MSMARCO 数据集我们采用 *NDCG* 和 *MRR* 指标, 对于 TrecQA 数据集我们采用 *MAP* 和 *MRR* 指标.

MRR (Mean Reciprocal Rank) 平均倒数排序, 公式如下:

$$MRR = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{p_i} \tag{9}$$

其中, Q 是问题的个数; p_i 为第 i 个问题中的第一个正确答案的排名位置. 即把第一个正确答案在排序给出结果中的位置取倒数作为它的准确度, 再对所有的问题求平均, 这个评价指标只关心第一个正确答案.

MAP (Mean Average Precision): 单个查询的平均准确率是每篇相关文档检索出后的准确率的平均值. *MAP* 是每个主题平均准确率的平均值. *MAP* 是反映系统在全部相关文档上性能的单值指标. 系统检索出来的相关文档越靠前 (*rank* 越高), *MAP* 就可能越高. 如果系统没有返回相关文档, 则准确率默认为 0.

NDCG (Normalized Discounted Cumulative Gain) 归一化折损累计增益, 计算公式如下:

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{10}$$

$$DCG@k = \sum_{i=1}^{k}\frac{2^{rel_i}-1}{\log_2(i+1)} \tag{11}$$

$$IDCG@k = \sum_{i=1}^{|REL|}\frac{2^{rel_i}-1}{\log_2(i+1)} \tag{12}$$

其中, k 表示 k 个文档组成的集合, *rel* 表示第 i 个文档的相关度. $|REL|$表示文档按照相关度从大到小排序, 取前 k 个文档组成的集合.

4.3 训练细节

AQGM 模型的 encoder 和 decoder 部分, 采用

LSTM 网络[27] 进行编码解码, 隐单元设置为 300. 通过该模型获得的训练数据: $\{Q+(1-y)Q\_gen, P, y\}$. 后续采用 BERT-base 得到查询文档对的相关性得分, 使用谷歌预训练的 BERT-base-uncased 作为 BERT 模型的初始化参数, 在下游分类任务上进行微调. 通过对语料的分析, 在模型中设置的参数如下: 输入模型的句子最大长度为 *max_sentence_length*=384; *batch_size*=64; 学习率为 2e–5、3e–5 和 4e–5; 优化函数采用 Adam[28]; 调节因子β_1设置为 1.1; 训练的 *epoch* 设置为 5.

本文采用 *MAP*、*MRR* 和 *NDCG*@10 指标作为评测模型性能的度量. 在测试集上评测 AQGM+BERT-base 模型和基线模型的得分, 并进行对比. 对于 MSMARCO 数据集, 设置 BM25, BERT-base, Doc2query+BERT-base 作为基准模型, 进行实验对比. 为进一步证实模型的有效性, 在 TrecQA 数据集上设置对照试验, 分别为 K-NRM 模型[29] 与 AQGM+K-NRM、BERT-base 模型与 AQGM+BERT-base 和 AQGM+BERT-base 模型与 Doc2query+BERT-base.

5 实验结果与分析

为得到 AQGM+BERT-base 模型的最优性能, 本文采用不同的初始学习率在 MSMARCO 和 TrecQA 数据集上进行实验, *MRR* 指标如表 2. 结果显示, 当学习率为 3e–5, 该模型在 MSMARCO 数据集上性能最优; 当学习率为 2e–5, 该模型在 TrecQA 数据集上性能最优.

表 2 不同学习率下的性能分析 (%)

学习率	MSMARCO (*MRR*@10)	TrecQA (*MRR*)
2e–5	34.7	93.4
3e–5	34.9	93.1
4e–5	34.4	91.5

表 3 和表 4 展示了不同数据集上, AQGM+BERT-base 模型和基线模型的 *MRR*、*MAP* 和 *NDCG*@10 评价指标得分.

分析表 3 可得, 在 MSMARCO 数据集上, 与 BERT-base 模型对比, AQGM+BERT-base 模型融入生成的对抗查询, 使 *MRR*@10 指标提升 1.2%, 证明对抗式数据增强方式能一定程度上提高神经网络排序模型的性能. 与基于文档扩充的 Doc2query+BERT-base 模型对比, 本文模型在 *MRR*@10, *NDCG*@10 指标分别提升 0.3% 和 1.5%, 证明对抗式数据增强相比文档扩充, 具有一定的优势.

表 3 MSMARCO 评测结果 (%)

方法	*MRR*@10	*NDCG*@10
BM25	18.6	17.3
Bert-base	33.7	58.0
Doc2query+Bert-base	34.6	60.2
AQGM+Bert-base(ours)	34.9	61.7

表 4 TrecQA 评测结果 (%)

方法	*MRR*	*MAP*
K-NRM	83.2	79.8
AQGM+K-NRM	85.7	83.1
Bert-base	90.2	85.9
Doc2query+Bert-base	90.4	86.5
AQGM+Bert-base(ours)	93.4	87.2

为进一步证实算法的有效性, 本文在 TrecQA 数据集上设置实验进行验证, 如表 4. K-NRM 模型加入对抗式数据增强方法 AQGM, 使 *MRR* 和 *MAP* 指标上升 2.5%、3.3%; AQGM+BERT-base 模型与 BERT-base 模型相比, 指标分别提升 3.2%、1.3%; AQGM+BERT-base 模型与 Doc2query+BERT-base 模型相比, 指标分别提升 3.0%、0.7%. 以上分析可得, 对抗式数据增强方式的有效性.

综上, 本文模型相比基线模型, 在 MSMARCO 和 TrecQA 数据集上性能均有一定提升, 在排序学习中融入对抗查询, 提高模型的稳健性. 在实际检索文档过程中, 返回与查询相关度高的文档, 能提高用户的搜索兴趣.

6 结论与展望

本文探索对比得到一种更适应于文本检索重排序的模型. AQGM+BERT-base 模型在 MS MARCO 和 TrecQA 数据集上得到了有效验证. 该方法简单且易理解, 在数据增强方面, 提供了一种新的思路, 从更具有挑战性的生成对抗查询的角度出发, 获得高质量的负样本. 此次尝试获得了有效的验证, 这为之后在这一领域的探索打开了良好的开端.

参考文献

1 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613–620. [doi: 10.1145/361219.361220]

2 Robertson SE, Jones KS. Relevance weighting of search terms. Journal of the American Society for Information Science, 1976, 27(3): 129–146. [doi: 10.1002/asi.4630270302]

3 Liu TY. Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 2009, 3(3): 225–331.
4 Wu HC, Luk RWP, Wong KF, *et al*. Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems, 2008, 26(3): 13.
5 Mitra B, Craswell N. An introduction to neural information retrieval. Hanover: Now Foundations and Trends, 2018.
6 Lee K, Chang MW, Toutanova K. Latent retrieval for weakly supervised open domain question answering. arXiv preprint arXiv: 1906.00300, 2019.
7 Goldberg Y, Levy O. Word2Vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv: 1402.3722, 2014.
8 Kim Y, Jernite Y, Sontag D, *et al*. Character-aware neural language models. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, AZ, USA. 2016.
9 Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 2009, 3(4): 333–389.
10 Plansangket S, Gan JQ. Re-ranking Google search returned web documents using document classification scores. Artificial Intelligence Research, 2017, 6(1): 59–68.
11 Baliński J, Daniłowicz C. Re-ranking method based on inter-document distances. Information Processing & Management, 2005, 41(4): 759–775.
12 Qu YL, Xu GW, Wang J. Rerank method based on individual thesaurus. NTCIR Workshop, 2007.
13 Guo J, Fan Y, Ai Q, *et al.* A deep relevance matching model for Ad-hoc retrieval. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. Indianapolis, IN, USA. 2016. 55–64.
14 Pang L, Lan YY, Guo JF, *et al*. A study of MatchPyramid models on Ad-hoc retrieval. arXiv preprint arXiv: 1606.04648, 2016.
15 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805, 2019.
16 Voorhees EM. Query expansion using lexical-semantic relations. In: Croft BW, van Rijsbergen CJ, eds. SIGIR '94. London: Springer, 1994. 61–69.
17 Berger A, Lafferty J. Information retrieval as statistical translation. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, CA, USA. 1999. 222–229.
18 Zerveas G, Zhang RC, Kim L, *et al*. Brown University at TREC deep learning 2019. Proceedings of the 8th Text Retrieval Conference (TREC) Notebook 2019. Gaithersburg, ML, USA. 2019.
19 Nogueira R, Yang W, Lin J, *et al*. Document expansion by query prediction. arXiv preprint arXiv: 1904.08375, 2019.
20 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. Advances in Neural Information Processing Systems, 2014, 3(2672): 2680.
21 Wang J, Yu LT, Zhang WN, *et al*. IRGAN: A minimax game for unifying generative and discriminative information retrieval models. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Japan. 2017. 515–524.
22 Bahuleyan H, Mou LL, Vechtomova O, *et al*. Variational attention for sequence-to-sequence models. arXiv: 1712.08207, 2018.
23 Nguyen TV, Rao N, Subbian K. Learning robust models for e-commerce product search. arXiv: 2005.03624, 2020.
24 Yang PL, Fang H, Lin J. Anserini: Reproducible ranking baselines using lucene. Journal of Data and Information Quality, 2018, 10(4): 16.
25 Bajaj P, Campos D, Craswell N, *et al*. MS MARCO: A human generated Machine reading Comprehension dataset. arXiv: 1611.09268, 2018.
26 Kingma DP, Ba JL. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA. 2015.
27 Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. Proceedings of the Interspeech, 13th Annual Conference of the International Speech Communication Association. Portland, ON, USA. 2012.
28 Wang MQ, Smith NA, Mitamura T. What is the jeopardy model? A quasi-synchronous grammar for QA. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic. 2007. 22–32.
29 Xiong CY, Dai ZY, Callan J, *et al*. End-to-end neural Ad-hoc ranking with kernel pooling. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Japan. 2017. 55–64.