arXiv:2304.04590v1 [cs.IR] 10 Apr 2023

LADER：用于生物医学文献检索的日志增强密集检索

Qiao Jin 美国马里兰州贝塞斯达市 美国国立卫生研究院
qiao.jin@nih.gov

Ashley Shin 美国马里兰州贝塞斯达市 美国国立卫生研究院
ashley.shin@nih.gov

Zhiyong Lu 美国马里兰州贝塞斯达市 美国国立卫生研究院
zhiyong.lu@nih.gov

摘要

具有相似信息需求的查询往往会产生相似的文档点击，尤其是在生物医学文献搜索引擎中，因为这类查询通常较短，且热门文档占总点击量的大部分。受此启发，我们提出了一种用于生物医学文献搜索的新型架构，即日志增强密集检索（LADER）。这是一个简单的插件模块，可利用从相似训练查询中检索到的点击日志来增强稠密检索器。具体而言，LADER 使用稠密检索器为给定查询查找相似文档和相似查询。随后，LADER 根据相似查询与输入查询的相似度，对这些相似查询中相关（被点击）文档进行加权评分。LADER 得出的最终文档分数，是以下两者的平均值：(1) 稠密检索器给出的文档相似度分数；(2) 相似查询点击日志汇总得到的文档分数。尽管方法简单，LADER 在近期发布的生物医学文献检索基准数据集 TripClick 上取得了新的最先进水平（SOTA）。在高频（“HEAD”）查询上，LADER 的表现远超最佳检索模型，相对 NDCG@10 提升 39%（0.338 对 0.243）。在频率较低的（“TORSO”）查询上，LADER 同样表现更佳，相较此前最先进水平的 NDCG@10 相对提升 11%（0.303 对 0.272）。对于相似查询稀少的罕见（“TAIL”）查询，LADER 仍优于此前的最先进方法（NDCG@10：0.310 对 0.295）。在所有查询上，LADER 无需额外训练即可将稠密检索器的表现提升 24%–37%（相对 NDCG@10），并且随着日志增多，预期性能还会进一步提升。我们的回归分析表明，频率更高、查询相似度熵更高且文档相似度熵更低的查询，往往能从日志增强中获益更多。

CCS 概念

• **信息系统** → **语言模型**; • **计算方法** → **自然语言处理**; • **应用计算** →**生命与医学科学**.

关键词

TripClick；生物医学文献检索；稠密检索

特此授予许可，允许为个人或课堂使用而免费制作本作品的全部或部分数字或纸质副本，前提是不得出于营利或商业利益而制作或分发副本，并且副本首页须附有本声明及完整引文。本作品中由作者以外的其他方拥有的组成部分的版权必须予以尊重。经注明出处后进行摘要是允许的。如需以其他方式复制、再版、发布到服务器或重新分发至列表，须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。*SIGIR '23*，*2023*年*7*月*23*日至*27*日，中国台湾台北。© 2023版权归所有者/作者所有。出版权授权予 ACM。ACMISBN 978-1-4503-9408-6/23/07…$ 15.00 https://doi.org/10.1145/XXXXXX.XXXXXX

ACM参考文献格式：

Qiao Jin、Ashley Shin 和 Zhiyong Lu。2023。LADER：用于生物医学文献检索的日志增强密集检索。载于 第*46*届*ACM*国际信息检索研究与发展会议论文集（*SIGIR '23*），*2023*年*7*月*23*日至*27*日，中国台湾台北。 ACM，美国纽约州纽约，6页。 https://doi.org/10.1145/XXXXXX.XXXXXX

1 引言

生物医学文献检索是知识发现和临床决策支持的关键步骤 [8, 12, 17]。它具有区别于其他信息检索（IR）任务的若干显著特性：(1) 大多数查询都很短。在广泛使用的生物医学文献搜索引擎PubMed1 中，查询的平均长度约为3.5 个词元 [7, 9, 10]；(2) 大量相关的查询-文档对可以轻松地从用户点击日志中收集。TripClick [29],是近期发布的生物医学文献检索基准数据集，包含从 Trip数据库[2] 日志中收集的 130 万个查询-文档相关性信号；(3) 用户大多浏览第一页上的文档 [7], , 而发布的 Trip数据库日志中，点击次数最多的前 31% 文档占据了 80%的点击。这些独特特征促使人们通过直接从相似查询的点击日志中进行检索，来增强生物医学文献检索，因为具有相似信息需求的查询往往具有相似的文档点击 [1, 38, 40]。这种方法本质上类似于近期的检索增强方法 [11, 20, 39]。图 1展示了TripClick 训练集中的一个示例查询，以及来自其相似查询的被点击文档。本工作的目标是利用此类文档增强生物医学文献检索。

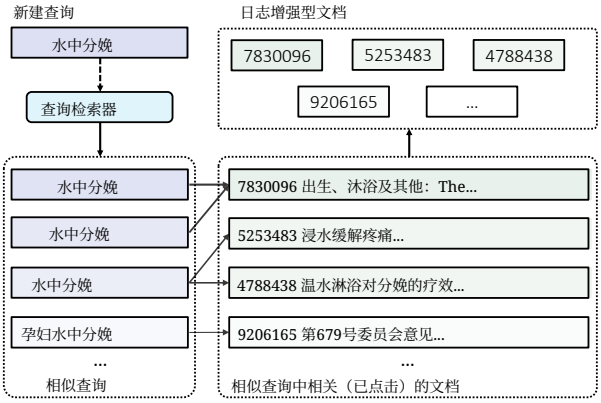


图1：使用检索到的相似查询的点击日志来增强生物医学文献检索的示例。

[1]https://pubmed.ncbi.nlm.nih.gov/[2] https://www.tripdatabase.com/

一种模型可以直接将相似查询的日志中被点击过的文档作为检索结果返回，这类似于推荐系统中的基于记忆的协同过滤 [33]。然而，这可能会遗漏此前未被点击过的相关文档。因此，日志中被点击过的文档更适合用于增强现有的检索器，而且从理论上说与检索器无关。我们选择使用稠密检索器，在这类检索器中，查询和文档会被编码并在低维嵌入空间中进行匹配，因为近期研究表明，基于 BERT [6] 等预训练语言模型的稠密检索器在各种任务上的表现优于传统的稀疏检索器 [19, 21, 37]。

在本文中，我们介绍了日志增强稠密检索（LADER），这是一种新颖且简单的插件模块，通过对从相似查询日志中聚合的文档分数与稠密检索器产生的原始分数进行插值，来增强稠密检索器。我们在最近引入的 TripClick 基准数据集 [29] 上开展了实验，并表明：（1）在所有频率组的查询上，LADER的表现均优于此前最先进水平（SOTA）模型；在高频（“HEAD”）查询上，NDCG@10的相对提升接近 40%；（2）LADER 无需额外训练即可将骨干稠密检索器的 NDCG@10相对提升24%-37%，并且随着日志增多，预计还能取得进一步的性能提升。我们的回归分析表明，频率更高、相似查询分数的熵更低以及相似文档分数的熵更高的查询，往往能从对数增强中获益更多。

2 方法

我们在 §2.1 中介绍骨干稠密检索器的训练，并在 §2.2 中介绍如何使用 LADER 模块进行推理。

2.1 稠密检索器训练

我们训练了一个稠密检索器，其中包含一个查询编码器（QEnc）和一个文档编码器（DEnc），二者均为 12 层 Transformer（Trm）编码器 [34]。编码器以 PubMedBERT-base[13], 初始化，后者是一种面向生物医学领域的 BERT 模型。在训练过程中，每个实例都包含一个查询 q、一个与该查询相关的文档 d^+ 以及一个与该查询不相关的文档 d^-。它们分别输入对应的编码器，并使用最后一个[CLS] 隐藏状态作为嵌入：

$$E(q) = \text{QEnc}(q) = \text{Trm}([\text{CLS}]\ q\ [\text{SEP}])_{[\text{CLS}]}$$

$$E(d) = \text{DEnc}(d) = \text{Trm}([\text{CLS}]\ d_{\text{title}}\ [\text{SEP}]\ d_{\text{abstract}}\ [\text{SEP}])_{[\text{CLS}]}$$

其中 [CLS] 和 [SEP] 是 BERT 中使用的特殊词元。

我们使用两种损失函数的组合来优化模型：批内负对数似然损失 L^i [19] 和三元组对比损失 L^t [32]。批内负对数似然损失有助于密集检索器区分正例文档与小批次中的所有其他文档：

$$\mathcal{L}^i = -\log \frac{\exp(E(q)^T E(d^+))}{\sum_{i \in B} \exp(E(q)^T E(d_i))}$$

其中 B 表示小批次中的所有文档。三元组损失进一步对比正文档和困难负文档：

$$\mathcal{L}^t = \max(0, \text{Dist}(E(q), E(d^+)) - \text{Dist}(E(q), E(d^-) + \alpha)$$

其中Dist 表示距离度量，α 是正样本对与负样本对之间的目标间隔。

最终损失是两个损失函数的加权和：

$$\mathcal{L} = \beta \mathcal{L}^i + (1 - \beta)\mathcal{L}^t$$

其中，β 是用于损失加权的超参数。我们通过基于梯度的优化器端到端地训练 QEnc 和 DEnc 中的参数。

2.2 LADER 推理

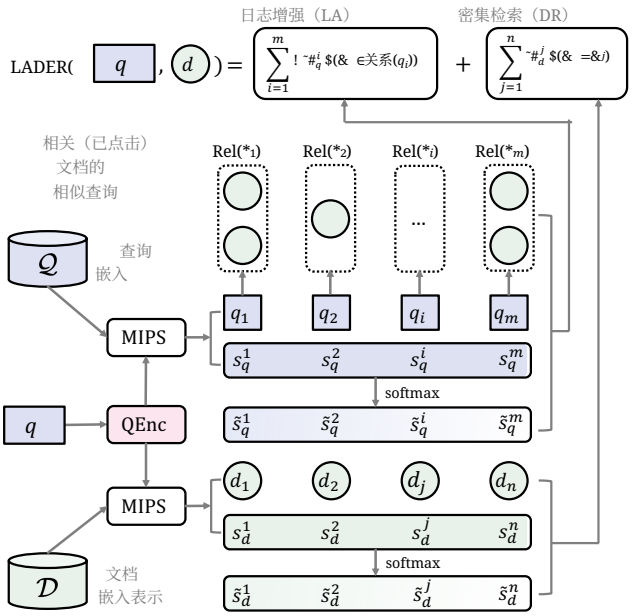


图 2：LADER 模型。

LADER 的整体架构如图 2 所示。我们使用训练好的 QEnc 对训练数据集中的所有查询进行编码，并使用训练好的 DEnc 对集合中的所有文档进行编码，分别得到 $\text{Q} \in \mathbb{R}^{N_q \times 768}$ 和 $\text{D} \in \mathbb{R}^{N_d \times 768}$。$N_q$ 表示训练查询的数量，N_d 表示集合中文档的数量。

在推理过程中，我们首先通过 QEnc 对新查询 q 进行编码。然后，我们执行最大内积搜索（MIPS），从训练集获取排名前m的相似查询（$q_1, q_2, \ldots, q_m$）及其内积相似度（$s_q^1, s_q^2, \ldots, s_q^m$），同时从文档集合中获取排名前$n$的相似文档（$d_1, d_2, \ldots, d_n$）及其相似度（$s_d^1, s_d^2, \ldots, s_d^n$）：

$$(q_1, s_q^1), (q_2, s_q^2), ..., (q_m, s_q^m) = \text{MIPS}(\text{QEnc}(q), \mathcal{Q})$$

$$(d_1, s_d^1), (d_2, s_d^2), ..., (d_n, s_d^n) = \text{MIPS}(\text{QEnc}(q), \mathcal{D})$$

相似度分数进一步通过 softmax 函数进行归一化：

$$\tilde{s}_q^1, \tilde{s}_q^2, ..., \tilde{s}_q^m = \text{softmax}(s_q^1, s_q^2, ..., s_q^m)$$

$$\tilde{s}_d^1, \tilde{s}_d^2, ..., \tilde{s}_d^n = \text{softmax}(s_d^1, s_d^2, ..., s_d^n)$$

我们将查询到其在训练集中的相关文档的映射记为 $\text{Rel}: q \mapsto \{d\}$。文档的最终得分 d

表 1：模型在 TripClick 测试集上的性能。基线结果来自 [29]。LA：对数增强；DR：稠密检索。

模型	标题 (DCTR相关性)			正文 (RAW相关性)			尾部(原始 相关性)		
	NDCG@10	MRR	Recall@10	NDCG@10	MRR	Recall@10	NDCG@10	MRR	Recall@10
BM25 [30]	0.140	0.290	0.138	0.206	0.283	0.262	0.267	0.258	0.409
RM3 伪相关反馈 [23, 26]	0.141	0.300	0.136	0.194	0.261	0.254	0.242	0.227	0.384
PACRR [16]	0.175	0.356	0.162	0.212	0.302	0.262	0.267	0.261	0.409
MP [27]	0.183	0.372	0.173	0.243	0.347	0.297	0.281	0.280	0.409
KNRM [36]	0.191	0.393	0.173	0.235	0.338	0.283	0.272	0.265	0.409
卷积KNRM [5]	0.198	0.420	0.178	0.243	0.358	0.288	0.271	0.265	0.409
TK [15]	0.208	0.434	0.189	0.272	0.381	0.321	0.295	0.279	**0.459**
LADER（我们的方法）	**0.338**	**0.664**	**0.304**	**0.303**	**0.427**	**0.353**	**0.310**	**0.306**	0.449
LADER（不含 LA）	0.247	0.532	0.237	0.241	0.350	0.293	0.260	0.257	0.394
不使用 DR 的 LADER	0.324	0.649	0.284	0.266	0.396	0.298	0.232	0.236	0.330

候选集 $\{d_1, d_2, ..., d_n\} \cup \{\text{Rel}(q_1), \text{Rel}(q_2) ..., \text{Rel}(q_m)\}$ 中的得分是按如下方式计算的加权和：

$$\text{Score}(d) = \sum_{j \in [1,...,n]} \tilde{s}_d^j \, \mathbf{1}(d = d_j) + \sum_{i \in [1,...,m]} \lambda \tilde{s}_q^i \, \mathbf{1}(d \in \text{Rel}(q_i))$$

其中，前一部分是稠密检索得分，后一部分是来自相似查询的对数增强得分，$\mathbf{1}$ 是指示函数，λ 是用于控制对数增强程度的超参数。我们根据最终得分对候选文档进行排序，并返回给输入查询。

3 实验

3.1 设置

数据集。 我们在 TripClick 基准数据集 [29], 上评估 LADER 方法，该数据集包含 69.2 万个唯一查询和 150 万篇文档（PubMed 摘要）。根据查询频率，这些查询被划分为头部（>44）、中部（6–44）和尾部（<6）子集。验证集和测试集中的头部、中部和尾部子集均各包含 1,175 个查询。按照 [14, 29], , 我们使用两组相关性分数：“RAW”相关性用于评估中部和尾部查询，其中被点击的文档得分为 1，其他文档得分为 0。分级文档点击率 [3, 4] （“DCTR”）相关性用于评估头部查询，其中文档得分定义为点击次数除以曝光次数。训练稠密检索器（§2.1）时，我们使用 [14], 发布的 1,000 万组三元组，每组三元组包含一个 TripClick 训练查询、一篇被点击的文档，以及一篇从 BM25 负例中采样的非相关文档。

配置。 我们使用PyTorch [28] 和HuggingFace 的库 [35]实现LADER。训练密集检索器时，我们使用AdamW优化器 [22, 25] 训练20k 步，学习率为 2e-5，权重衰减为1e-2，批量大小为256，预热步数为10k，采用余弦学习率衰减调度、 $\alpha = 5.0$、 $\beta = 0.9$以及三元组损失的欧氏距离。对于推理，我们使用 FAISS 的 FlatIP 索引 [18]、 $m = n = 1000$、 $N_q = 685\ 649$ $N_d = 1\ 523\ 871$ ，以及 , 、实现 MIPS。对于 HEAD 和中部查询，我们使用 $\lambda = 0.5$，对于尾部查询，我们使用 $\lambda = 0.2$。我们还进行了两项消融实验：无对数增强的 LADER（LA）其中 LA 分数设为 0，以及不含密集检索（DR）的 LADER，其中 DR 分数设为 0。

用于比较的基准方法。我们在 [29]的所有查询上将 LADER 与多种基准方法进行比较，包括 BM25 [30], RM3 伪相关反馈（PRF） [23, 26], 位置感知卷积循环相关性匹配（PACRR） [16], 匹配金字塔（MP） [27], 基于核的神经排序模型（KNRM） [36], 卷积KNRM（卷积KNRM） [5], 以及 Transformer-Kernel（TK）[15]。对于头部查询，我们还与当前基于预训练语言模型的最先进的方法进行比较 [14] ，包括由不同 BERT 模型初始化的双编码器 [2, 13, 31]。按照各自采用的指标，我们使用 NDCG@10、MRR 和 Recall@10 与基准方法进行比较 [29], 并使用 NDCG@10、MRR@10 以及 Recall@1k 与头部查询上的最先进水平进行比较 [14] 。

3.2 主要结果

与基准方法的比较。 表 1 展示了 LADER 与多种基准方法在所有查询上的全面比较 [29] 。LADER 在每个查询子集和所有指标上均优于所有基准方法, 唯独在尾部查询的 Recall@10 指标上例外。对于频率最高的头部查询, LADER 大幅超过表现最佳的基准方法（TK），在 NDCG@10、MRR 和 Recall@10 上的相对增幅分别为 63%（0.338 对 0.208）、53%（0.664 对 0.434）和 61%（0.304 对 0.189）。对于频率较低的中部查询，相较于此前的最先进水平，性能提升相当可观，所有指标的相对增幅均为 10% 至 12%。对于罕见的尾部查询，LADER 在 NDCG@10（0.310 对 0.295）和 MRR（0.306 对 0.280）上仍优于此前的最先进水平，但 Recall@10（0.449 对 0.459）略低。

与 ***BERTDOT*** 的比较. 在表 2 中，我们将 LADER 与较新的基于 BERT 的检索器（BERT DOT）进行比较 [14], , 其结果仅在头部查询上可用。与 BERT DOT 稠密检索器相比，LADER 在 NDCG@10、MRR@10 和 Recall@1k 上分别提升了 39%（0.338 v.s. 0.243）、24%（0.659 v.s. 0.530）和 7%（0.893 v.s. 0.828）。这表明了对数增强的有效性，因为我们的骨干稠密检索器（LADER w/o LA）的表现与其对应模型相近。

表 2：LADER 在头部查询上的结果，与 BERT DOT [14] 对比；LA：对数增强；DR：稠密检索。

模型	NDCG@10	MRR@10	R@1k
BM25 [30]	0.140	0.276	0.834
BERT DOT [14]			
使用 DistillBERT [31]	0.236	0.512	0.813
与 SciBERT [2]搭配	0.243	0.530	0.793
使用 PubMedBERT [13]	0.235	0.509	0.828
LADER（我们的方法）	**0.338**	**0.659**	**0.893**
LADER（不含 LA）	0.247	0.526	0.878
不含 DR 的 LADER	0.324	0.644	0.670
日志增强型（LA-）			
稀疏检索器（BM25）	0.312	0.598	0.889
原始 PubMedBERT [13]	0.137	0.288	0.453

相较于骨干稠密检索器的改进。 在所有查询上，LADER 将骨干稠密检索器的性能相对提升了 24%-37% 的 NDCG@10（LADER 与不含 LA 的 LADER 对比），且无需额外训练。与躯干查询或尾部查询相比，头部查询上的性能提升更为显著，下一节将对此进行进一步分析。

3.3 分析

训练稠密检索器的重要性。 在表 2 中，我们展示了使用不同骨干稠密检索器进行对数增强的结果。经过对数增强的原始 PubMedBERT 表现远逊于 LADER（0.137 对比 0.338 NDCG@10），这表明骨干稠密检索器需要在检索任务上进行训练（§2.1）。我们还将 LADER 中的稠密检索器替换为 BM25，并将其称为 LABM25，使用 Pyserini [24] 实现。LABM25 将原始 BM25基线的性能提升了 122%（0.312 对比 0.140 NDCG@10），证明了对数增强的有效性与检索器无关。然而，LADER 仍然优于 LABM25（0.338 对比 0.312 NDCG@10），这表明稠密检索器能够更好地发挥对数增强的潜力。

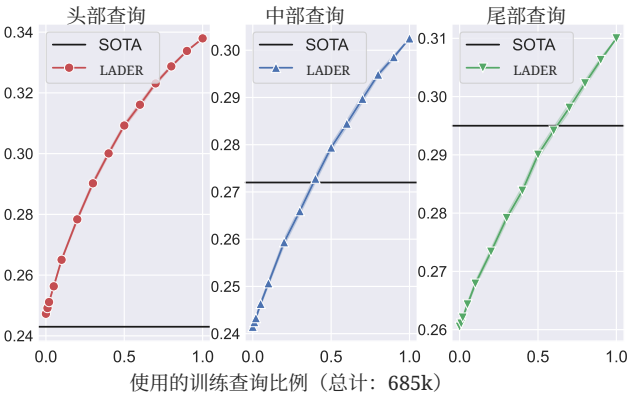


图 3：采用不同比例的训练查询进行对数增强检索时，LADER 的 NDCG@10。

对数规模的影响。 在图 3 中，我们展示了 LADER 使用从训练集采样的不同比例查询进行对数增强时的性能。对于所有查询，随着待检索查询数量的增加，性能都会提升；由于这些曲线尚未达到饱和，预计增加更多日志后还能获得更大的性能提升。我们还发现，对于频率较低的查询，需要更多训练查询用于检索，LADER 才能超越当前最先进的方法。

哪些查询从对数增强中获益更多？我们为每个查询收集了 9 个特征，包括查询长度（QL）、查询组（HEAD、TORSO 或 TAIL）、ENT(Q)：

$[\tilde{s}_q^1, \tilde{s}_q^2, ..., \tilde{s}_q^m]$，ENT(D)：$[\tilde{s}_d^1, \tilde{s}_d^2, ..., \tilde{s}_d^n]$ 的熵，E(REL)：相关文档数量的期望值，以及最相似的 1 个和 5 个查询中的相关文档平均数量（REL1 和 REL5）。我们使用经过最小-最大归一化的查询特征，拟合线性回归模型，以预测对数增强带来的 NDCG@10 增益（LADER 与不使用对数增强的 LADER 相比）。图 4 所示的特征系数表明：（1）查询频率是最重要的特征：频率较高的查询（HEAD 组）比平均水平获益更多，而罕见查询（TAIL 组）获益较少；（2）没有非常相似文档（ENT(D) 较高）的查询，比存在非常相似文档的查询获益更多；（3）日志中存在非常相似查询（ENT(Q) 较低）的查询，比不存在此类查询的查询更能从对数增强中获益。

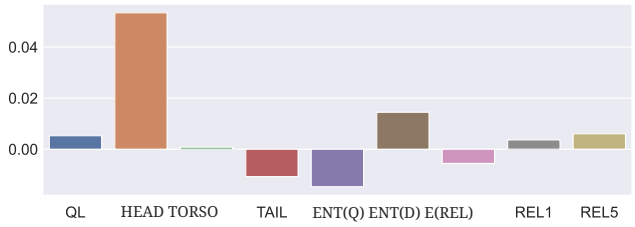


图4：回归分析中的特征系数。正值表示对数增强带来的收益更大。

4 结论与局限

我们提出了 LADER，这是一种简单新颖的插件模块，利用搜索日志增强稠密检索器。结果表明，LADER 在 TripClick 上达到了新的最先进水平，并且无需额外训练即可显著提升其骨干检索器的性能。我们还对其特性进行了全面分析。

LADER 的一个局限是，由于增加了查询间检索步骤，搜索延迟会增加约 N_q/N_d（在 TripClick 上为 45%）。本研究的另一个不足之处是，我们仅使用了一个搜索引擎的数据。未来仍值得测试 LADER 在其他搜索引擎上的泛化能力，这将有助于新文献搜索项目的冷启动。

致谢

我们感谢 TripClick 基准数据集的组织者分享数据，也感谢 SIGIR 评审提出的建设性意见。本研究由美国国立卫生研究院（NIH）院内研究计划、美国国家医学图书馆资助。

参考文献

[1] Ricardo A. Baeza-Yates、Carlos A. Hurtado 和 Marcelo Mendoza。2007。通过查询聚类改进搜索引擎。信息科学与技术协会期刊58, 12 (2007), 1793 1804 10 1002 20627 – 。https://doi.org/ 。/asi。 Iz Beltagy、Kyle Lo 和 Arman Cohan。。SciBERT：用于科学文本的预训练语言模型。载于 *2019* 年自然语言处理实证方法会议暨第9届自然语言处理国际联合会议论文集，*EMNLP-IJCNLP 2019*，中国香港，*2019*年*11*月*3*日至*7*日，Kentaro Inui、Jing Jiang、Vincent Ng 和 Xiaojun Wan（编）。计算语言学协会，3613 3618 10 18653 1 19 1371 – 。https://doi.org/ 。/ [2] 2019

v /D - Aleksandr Chuklin、Ilya Markov 和 Maarten de Rijke。。网络搜索点击模型。Morgan & Claypool Publishers。https://doi.org/10.2200/S00654ED1V01Y201507ICR043[3] Nick Craswell、Onno Zoeter、Michael J. Taylor 和 Bill Ramsey。2008。点击位置偏差模型的实验比较。载于网络搜索与数据挖掘国际会议论文集，*WSDM 2008*，美国加利福尼亚州帕洛阿尔托，*2008*年*2*月*11*日至*12*日，Marc Najork、Andrei Z. Broder 和 Soumen Chakrabarti（编）。ACM，87 94 10 1145 1341531 1341545 – 。 2015

https://doi.org/ 。/ 。Zhuyun Dai、Chenyan Xiong、Jamie Callan 和 Zhiyuan Liu。。用于临时搜索中 N 元语法软匹配的卷积神经网络。载于第十一届 *ACM* 网络搜索与数据挖掘国际会议论文集，*WSDM 2018*，美国加利福尼亚州玛丽娜德尔雷伊，*2018*年*2*月*5*日至*9*日，Yi Chang、Chengxiang Zhai、Yan Liu 和 Yoelle Maarek（编）。ACM，126–134。[5] 2018

https://doi.org/10.1145/3159652.3159659[6] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向 Transformer 预训练。载于*2019*年北美计算语言学协会分会会议：人类语言技术论文集，*NAACL-HLT 2019*，美国明尼苏达州明尼阿波利斯，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文），Jill Burstein、Christy Doran 和 Thamar Solorio（编）。计算语言学协会，4171 4186 10 18653 1 19 1423 – 。https://doi.org/ 。/v /n - [7] 2019

Rezarta Islamaj Dogan、G. Craig Murray、Aurélie Névéol 和 Zhiyong Lu。2009。通过日志分析了解 PubMed® 用户的搜索行为。数据库2009 (2009年11月)。https://doi.org/10.1093/database/bap018

a r X i v: h t t p s:// academic.oup.com/database/article-pdf/doi/10.1093/database/bap018/1022874/bap018.pdf bap018。[8] John W. Ely、Jerome A. Osheroff、M. Lee Chambliss、Mark H. Ebell 和 Marcy E. Rosenbaum。2005。研究论文：回答医生的临床问题：障碍与潜在解决方案。美国医学信息学协会期刊12, 2 (2005), 217–224。https://doi.org/10.1197/jamia.M1608[9] Nicolas Fiorini、Kathi Canese、Grisha Starchenko、Evgeny Kireev、Won Kim、Vadim Miller、Maxim Osipov、Michael Kholodov、Rafis Ismagilov、Sunil Mohan 等。2018。最佳匹配：PubMed 的全新相关性搜索。*PLoS Biology*16, 8 (2018), e2005343。[10] Nicolas Fiorini、Robert Leaman、David J Lipman 和 Zhiyong Lu。2018。用户智能如何改善 PubMed。自然生物技术36, 10 (2018), 937–945。[11] Giacomo Frisoni、Miki Mizutani、Gianluca Moro 和 Lorenzo Valgimigli。2022。BioReader：一种用于生物医学文献的检索增强型文本到文本 Transformer。载于*2022*年自然语言处理实证方法会议论文集。计算语言学协会，阿联酋阿布扎比，5770 5793 2022 390 – 。 [12]

https://aclanthology.org/ 。emnlp-main。Vishrawas Gopalakrishnan、Kishlay Jha、Wei Jin 和 Aidong Zhang。2019。生物医学领域基于文献的发现方法综述。生物医学信息学杂志93 (2019)。https://doi.org/10.1016/j.jbi.2019.103141[13] Yu Gu、Robert Tinn、Hao Cheng、Michael Lucas、Naoto Usuyama、Xiaodong Liu、Tristan Naumann、Jianfeng Gao 和 Hoifung Poon。2021。面向生物医学自然语言处理的领域专用语言模型预训练。*ACM* 医疗保健计算学报3, 1, 文章 2 (2021年10月)，23页。https://doi.org/10.1145/3458754[14] Sebastian Hofstätter、Sophia Althammer、Mete Sertkan 和 Allan Hanbury。2022。为 TripClick 健康检索建立强基线。载于信息检索进展——第*44*届欧洲信息检索研究会议，*ECIR 2022*，挪威斯塔万格，*2022*年*4*月*10*日至*14*日，论文集，第二部分（计算机科学讲义，第*13186*卷），Matthias Hagen、Suzan Verberne、Craig Macdonald、Christin Seifert、Krisztian Balog、Kjetil Nørvåg 和 Vinay Setty（编）。Springer，144–152。https://doi.org/10.1007/978-3-030-99739-7_17[15] Sebastian Hofstätter、Hamed Zamani、Bhaskar Mitra、Nick Craswell 和 Allan Hanbury。2020。面向高效文档检索的长文本局部自注意力。载于第*43*届 *ACM SIGIR* 信息检索研究与发展国际会议论文集，*SIGIR 2020*，线上会议，中国，*2020*年*7*月*25*日至*30*日，Jimmy X. Huang、Yi Chang、Xueqi Cheng、Jaap Kamps、Vanessa Murdock、Ji-Rong Wen 和 Yiqun Liu（编）。ACM，2021–2024。https://doi.org/10.1145/3397271.3401224[16] Kai Hui、Andrew Yates、Klaus Berberich 和 Gerard de Melo。2017。PACRR：一种位置感知神经信息检索相关性匹配模型。载于*2017*年自然语言处理实证方法会议论文集，*EMNLP 2017*，丹麦哥本哈根，*2017*年*9*月*9*日至*11*日，Martha Palmer、Rebecca Hwa、以及 Sebastian Riedel（编）。计算语言学协会，1049–1058。https://doi.org/10.18653/v1/d17-1110[17] Qiao Jin、Chuanqi Tan、Mosha Chen、Ming Yan、Ningyu Zhang、Songfang Huang、Xiaozhong Liu 等。2022。《精准医疗的最先进证据检索器：算法开发与验证》。*JMIR* 医学信息学 10，12（2022），40743（），e。[18] Jeff Johnson、Matthijs Douze 和 Hervé Jégou。2019。《使用 GPU 进行十亿级相似性搜索》。*IEEE* 大数据汇刊 7，3（2019），535–547。[19] Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick S. H. Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020。《用于开放域问答的密集段落检索》。载于*2020*年自然语言处理经验方法会议论文集（*EMNLP 2020*），线上会议，*2020*年*11*月*16*日至*20*日，Bonnie Webber、Trevor Cohn、Yulan He 和 Yang Liu（编）。计算语言学协会，6769–6781。https://doi.org/10.18653/v1/2020.emnlp-main.550[20] Urvashi Khandelwal、Omer Levy、Dan Jurafsky、Luke Zettlemoyer 和 Mike Lewis。2020。《通过记忆实现泛化：最近邻语言模型》。载于第八届国际学习表征会议（*ICLR 2020*），埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。https://openreview.net/forum?id=HklBjCEKvH [21] Omar Khattab 和 Matei Zaharia。2020。《ColBERT：通过 BERT 上下文晚期交互实现高效且有效的段落搜索》。载于第*43*届 *ACM* 国际信息检索研究与开发会议论文集（*SIGIR 2020*），线上会议，中国，*2020*年*7*月*25*日至*30*日，Jimmy X. Huang、Yi Chang、Xueqi Cheng、Jaap Kamps、Vanessa Murdock、Ji-Rong Wen 和 Yiqun Liu（编）。ACM，39 48 10 1145 3397271 3401075 – 。https://doi.org/。/。 [22] 2015

Diederik P. Kingma 和 Jimmy Ba。。Adam：随机优化方法。载于第三届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，Yoshua Bengio 和 Yann LeCun（编）。http://arxiv.org/abs/1412 6980 。Victor Lavrenko 和 W. Bruce Croft。。[23] 2001

基于相关性的语言模型。载于*SIGIR 2001*：第*24*届 *ACM* 国际信息检索研究与开发年会论文集，*2001*年*9*月*9*日至*13*日，美国路易斯安那州新奥尔良，W. Bruce Croft、David J. Harper、Donald H. Kraft 和 Justin Zobel（编）。ACM，120 127 10 1145 383952 383972 –。https://doi.org/。/。 [24]

Jimmy Lin、Xueguang Ma、Sheng-Chieh Lin、Jheng-Hong Yang、Ronak Pradeep 和 Rodrigo Nogueira。2021。Pyserini：用于稀疏和稠密表示的可复现信息检索研究 Python 工具包。载于第*44*届 *ACM* 国际信息检索研究与开发年会论文集（*SIGIR 2021*）。2356–2362。[25] Ilya Loshchilov 和 Frank Hutter。2019。解耦权重衰减正则化。载于第七届国际学习表征会议（*ICLR 2019*），美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net。https://openreview.net/forum?id = 6 7 Bkg RiCqY Yuanhua Lv 和 ChengXiang Zhai。2009。利用伪反馈估计查询语言模型的方法比较研究。载于第*18*届 *ACM* 信息与知识管理会议论文集（*CIKM 2009*），中国香港，*2009*年*11*月*2*日至*6*日，David Wai-Lok Cheung、Il-Yeol Song、Wesley W. Chu、Xiaohua Hu 和 Jimmy Lin（编）。ACM，1895–1898。https://doi.org/ 10 1145 1645953 1646259 。/。 [26]

Liang Pang、Yanyan Lan、Jiafeng Guo、Jun Xu、Shengxian Wan 和 Xueqi Cheng。2016。《将文本匹配视为图像识别》。载于第三十届 *AAAI* 人工智能会议论文集，*2016*年*2*月*12*日至*17*日，美国亚利桑那州菲尼克斯，Dale Schuurmans 和 Michael P. Wellman（编）。AAAI Press，2793–2799。http://www.aaai.org/ocs/index.php/AAAI/AAAI 16 11895 /paper/view/ [27]

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Köpf、Edward Z. Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。2019。PyTorch：一种命令式风格、高性能的深度学习库。载于神经信息处理系统进展*32*：*2019*年神经信息处理系统年会（*NeurIPS 2019*），*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，Hanna M. Wallach、Hugo Larochelle、Alina Beygelzimer、Florence d'Alché-Buc、Emily B. Fox 和 Roman Garnett（编）。8024–8035。 [28]

h t t p s:// proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html[29] Navid Rekabsaz、Oleg Lesota、Markus Schedl、Jon Brassey 和 Carsten Eickhoff。2021。TripClick：大型健康网络搜索引擎的日志文件。载于第*44*届 *ACM* 国际信息检索研究与开发会议论文集（线上会议，加拿大）（*SIGIR '21*）。计算机协会，纽约，NY，美国，2507–2513。https://doi.org/10.1145/3404835.3463242[30] Stephen E. Robertson 和 Hugo Zaragoza。2009。概率相关性框架：BM25及其拓展。信息检索基础与趋势 3，4（2009），333–389。https://doi.org/ 10 1561 1500000019 。Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。2019。DistilBERT，BERT 的蒸馏版本：更小、更快、更便宜、更轻量。*CoRR* abs/1910.01108（2019）。arXiv:1910.01108 http://arxiv.org/abs/1910.01108 [31]

[32] Florian Schroff、Dmitry Kalenichenko 和 James Philbin。2015。FaceNet：用于人脸识别与聚类的统一嵌入表示。载于*IEEE* 计算机视觉与模式识别会议，*CVPR 2015*，美国马萨诸塞州波士顿，*2015*年*6*月*7*日至*12*日。IEEE 计算机学会，815 823 10 1109 2015 7298682 –。https://doi.org/。/CVPR。。[33] Xiaoyuan Su 和 Taghi M. Khoshgoftaar。2009。协同过滤技术综述。先进人工智能 2009（2009），421425:1–421425:19。https://doi.org/10.1155/2009/421425[34] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017。注意力就是你所需要的一切。载于神经信息处理系统进展*30*：*2017*年神经信息处理系统年会，*2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩，Isabelle Guyon、Ulrike von Luxburg、Samy Bengio、Hanna M. Wallach、Rob Fergus、S. V. N. Vishwanathan 和 Roman Garnett（编）。5998–6008。
h
t
t
ps://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html[35] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。2020。Transformers：最先进的自然语言处理。载于*2020*年自然语言处理经验方法会议论文集：系统演示。计算语言学协会，线上，38–45。https://www.aclweb.org/anthology/2020.emnlp-demos.6[36] Chenyan Xiong、Zhuyun Dai、Jamie Callan、Zhiyuan Liu 和 Russell Power。2017。通过核池化实现端到端神经即席排序。载于论文集

第*40*届国际*ACM SIGIR*信息检索研究与开发会议，日本东京新宿，*2017*年*8*月*7*日至*11*日，Noriko Kando、Tetsuya Sakai、Hideo Joho、Hang Li、Arjen P. de Vries 和 Ryen W. White（编）。ACM，55 64 10 1145 3077136 3080809 –。[37] https://doi.org/。/。Lee Xiong、Chenyan Xiong、Ye Li、Kwok-Fung Tang、Jialin Liu、Paul N. Bennett、Junaid Ahmed 和 Arnold Overwijk。2021。用于密集文本检索的近似最近邻负对比学习。载于 第*9*届国际学习表征会议，*ICLR 2021*，线上会议，奥地利，*2021*年*5*月*3*日至*7*日。OpenReview.net。https://openreview.net/forum?id=zeFrfgyZln [38] Zhijun Yin、Milad Shokouhi 和 Nick Craswell。2009。使用外部证据进行查询扩展。载于 信息检索进展，第*31*届欧洲信息检索研究会议，*ECIR 2009*，法国图卢兹，*2009*年*4*月*6*日至*9*日。会议录（《计算机科学讲义》，第*5478*卷），Mohand Boughanem、Catherine Berrut、Josiane Mothe 和 Chantal Soulé-Dupuy（编）。Springer，362–374。https://doi。org/10.1007/978-3-642-00958-7_33[39] Zheng Yuan、Qiao Jin、Chuanqi Tan、Zhengyun Zhao、Hongyi Yuan、Fei Huang 和 Songfang Huang。2023。RAMM：通过多模态预训练实现检索增强的生物医学视觉问答。*CoRR* abs/2303.00534（2023）。https://doi.org/10 48550 2303 00534 2303 00534 。/arXiv。。arXiv:。Shengyao[40] Zhuang、Hang Li 和 Guido Zuccon。2022。密集段落检索的隐式反馈：一种反事实方法。载于*SIGIR '22*：第*45*届国际*ACM SIGIR*信息检索研究与开发会议，西班牙马德里，*2022*年*7*月*11*日至*15*日，Enrique Amigó、Pablo Castells、Julio Gonzalo、Ben Carterette、J. Shane Culpepper 和 Gabriella Kazai（编）。ACM，18–28。https://doi.org/10.1145/3477495.3531994