兼顾效率与效果！搜索系统中的大规模预排序框架

赵启航*中国科学技术大学，中国安徽 快手，中国北京 zhaoqihang@kuaishou.com

朱睿杰*美国加利福尼亚大学圣克鲁兹分校 ridger@live.cn

杨柳*快手，中国北京 liuyang22@kuaishou.com

何永明 快手，中国北京 heyongming@kuaishou.com

周博 快手，中国北京 yefeng@kuaishou.com

罗成*快手，北京，中国 luocheng06@kuaishou.com

arXiv:2304.02434v2 [cs.IR] 9 May 2023

摘要

在搜索系统领域，多阶段级联架构是一种普遍采用的方法，通常由匹配、粗排和排序等顺序模块组成。人们普遍认为，粗排阶段所使用的模型必须在效果与效率之间取得平衡。因此，最常采用的架构是基于表示的向量乘积模型。然而，这种架构缺乏查询与文档之间的有效交互，导致搜索系统的效果下降。为了解决这一问题，我们提出了一种名为**RankDFM**的新型粗排框架。我们的框架以 DeepFM 为骨干，并采用成对式训练范式来学习查询下视频的排序。RankDFM的特征交叉能力显著提升了离线和在线A/B测试的性能。此外，我们引入了一种可学习的特征选择方案，用于优化模型并减少在线推理所需的时间，其效果等同于树模型。目前，RankDFM 已部署于一个短视频应用的搜索系统中，每日为数亿用户提供服务。

CCS 概念

• **信息系统** →**学习排序**.

关键词

预排序，搜索系统，特征选择，学习排序

ACM 参考文献格式：

赵启航、朱睿杰、杨柳、何永明、周博、罗成。2023 年。效率与效果兼备！搜索系统中的大规模预排序框架。发表于第*45*届国际*ACM* 信息检索研究与发展会议论文集（*SIGIR '23*），*2022*年*7*月*11*日至*15*日，西班牙马德里。ACM，纽约，纽约州，美国，5页。 https://doi.org/xx.xxxx/xxxxxx.xxxxxxxx

*通讯作者

特此许可，无需支付费用，即可为个人或课堂使用而制作本作品的全部或部分数字版或纸质副本，但前提是不得为营利或商业利益而制作或分发副本，并且这些副本必须在首页附有本声明及完整引文。对于本作品中由非 ACM 所有的组件，必须尊重其版权。注明出处后允许进行摘要。否则，如需复制、再版、发布到服务器或重新分发至列表，必须事先获得特定许可和/或支付费用。请通过 permissions@acm.org 申请许可。*SIGIR '23*，*2022*年*7*月*11*日至*15*日，西班牙马德里© 2023 计算机协会。ACMISBN xxx-x-xxxx-xxxx-x/xx/xx...$ 15.00 https://doi.org/xx.xxxx/xxxxxx.xxxxxxxx

1 引言

近年来，互联网服务呈指数级激增，用户不得不应对海量信息。为了缓解这一难题，各大互联网平台纷纷采用搜索系统，帮助数十亿用户找到所需信息。排序系统是搜索系统的关键组成部分，它会从庞大的候选池中审慎地选出有限数量的条目。遗憾的是，复杂精密的排序模型常常受限于系统延迟的严格要求，无法完成对整个候选集的排序过程。[1–3, 7–10, 13] 为克服这一限制，绝大多数排序系统采用多阶段级联架构，该架构由三个线性顺序排列的模块组成：匹配、预排序和排序。图 1 展示了这一架构。

排序系统级联架构中的粗排阶段负责过滤匹配阶段选出的条目，并将其转发至排序模块。与只需对数量级为十的条目进行排序的排序模块不同，该阶段通常需要处理数百个条目。为确保系统效率，粗排模块通常采用基于表示的模型 [6, 12] 等简单模型，尽管由于表达能力有限，这些模型会降低系统的整体有效性。另一方面，基于交互的模型允许特征之间进行交互，与基于表示的 [11]模型相比具有更强的表达能力。然而，这些模型通常需要搜索系统无法承担的高额计算资源，因此会采用浅层 MLP 模型 [3, 11], 等更简单的模型，从而牺牲基于交互的粗排模块的有效性。因此，本文旨在研究一种兼顾效率和有效性的基于交互的粗排模型。

在本文中，我们提出了一种新的预排序算法，旨在为实际的搜索系统平衡效率和效果。我们提出的方法名为 RankDFM，采用了一个鲁棒的特征交互信息提取模型，具体而言是以 DeepFM [4],as作为其主干模型。训练阶段采用成对式训练策略，并基于所生成的正样本和负样本

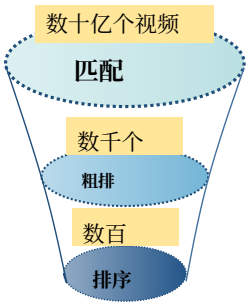

图 1：工业信息检索系统级联架构示意图。

排序模型的输出。因此，RankDFM的目标是学习排序模型的物品列表排序。在推理过程中，RankDFM 直接根据物品的特征计算其排序分数。然而，由于 DeepFM的模型复杂度极高，将其作为粗排阶段的主导模型将对整个系统造成灾难性影响。为了解决这一问题，我们在训练过程中引入了国内生产总值（Gates with Differentiable Polarization）机制 [5]，该机制能够自动学习模型剪枝所需的特征交互，从而有效降低模型复杂度。我们的实验结果包括在线A/B测试和离线评估，证明了在两个真实世界数据集上，RankDFM 优于其他基线方法。

本文的贡献如下：

- 我们提出了一种新颖的预排序框架RankDFM，该框架以 DeepFM 为骨干，通过优化特征交互来增强预训练过程的排序能力。
- 为确保 RankDFM 的模型复杂度与粗排阶段的需求相称，我们引入了国内生产总值（差分极化门控）剪枝机制，从而在保持 RankDFM 性能的同时显著减少计算资源。
- 我们在真实世界数据集上的大量实验以及在线A/B测试，为所提出的 RankDFM 的优越性提供了有力证据。

目前，RankDFM 框架已成功部署在短视频应用搜索系统中，每日为数亿用户提供服务。

2 提出的方法

粗排阶段位于整个搜索级联架构的中间模块。其作用是对匹配层输出的数千个项目进行排序，并将排名靠前的数百个项目发送给精排层，以进行更准确、更精细的排序。换句话说，粗排层的功能是对匹配层的结果进行筛选，以减轻精排层的压力。形式化地，给定一个查询 q，假设匹配层返回的项目集合为 $S_{\mathcal{P}} = \{I_1, I_2, \cdots, I_n\}$，粗排层的作用是学习一个函数：$f = prerank(I \times q, I \times q)$ ($I \in S_{\mathcal{P}}$，$I \times q$ 表示项目与查询），对 $S_{\mathcal{P}}$中的每个项目 I 进行评分，并从中选出排名前 100 的项目，交付给精排层。

本文提出了一种兼具效率与效果的新型预排序框架 RankDFM。以 DeepFM 为骨干网络提升了模型的效果，而引入 GDP 剪枝策略则优化了其计算效率。接下来将全面介绍 RankDFM 的训练和推理细节。RankDFM 框架的可视化示意图如图 2 所示。

2.1 RankDFM 的训练

如上所述，粗排阶段的学习目标是排序模块对候选项的排序结果。通常，学习排序有三种范式：点式、成对式和列表式。为了降低训练难度并提高训练效率，我们选择了成对式这一更经济的训练方法。接下来，我们将描述训练的具体细节。

2.1.1 构造训练样本。成对式范式旨在构造部分有序的样本对，并通过比较学习排序。因此，创建训练所需的样本对是第一步。使用来自粗排层的数百个候选样本来学习精排层的排序结果是一种直接的方法，但该方法存在 SSB（样本选择偏差）问题，使整个样本空间无法被模型访问。因此，我们的方法是调用排序模型，对匹配层候选集 $S_{\mathcal{M}}$ 中的每个项目进行评分，以获得集合 $S^R_{\mathcal{M}}$。然后在集合 $S^R_{\mathcal{M}}$上构造部分样本对。

具体而言，以集合 $S^R_{\mathcal{M}}$ 中的一个项目 I 为例，我们尝试了以下三种方案：

- 层间。我们根据精排层的分数区间将所有物品划分为四个层级。假设 I属于第三层级，则从第一、第二和第四层级的物品集合中选择一定数量的物品，而不从其所在的第三层级中选择样本，以构造偏序样本对。
- 全等级。 与前一种方法类似，所有项目被分为四个等级。假设 I 属于第三级，I 分别从第一、第二、第三和第四级的项目集合中选择一定数量的项目，以构建偏序样本对。这种样本构建方法考虑了同一等级的困难样本。
- 随机。 尽管前一种方法能够很好地考虑困难样本，但它仍存在一个缺陷：该方法需要仔细调节在不同层级采样的锚点样本比例，这一过程耗时且费力。在此背景下，我们采用随机选择负样本的方法。给定一个物品 I和一个阈值 ϵ，从物品集合中随机选择一定数量的样本，其中样本间的排序层分数差异大于阈值 ϵ，以构造偏序样本对。

最后，第三种方法摒弃了层级这一概念，无需微调样本数量，简单易行。

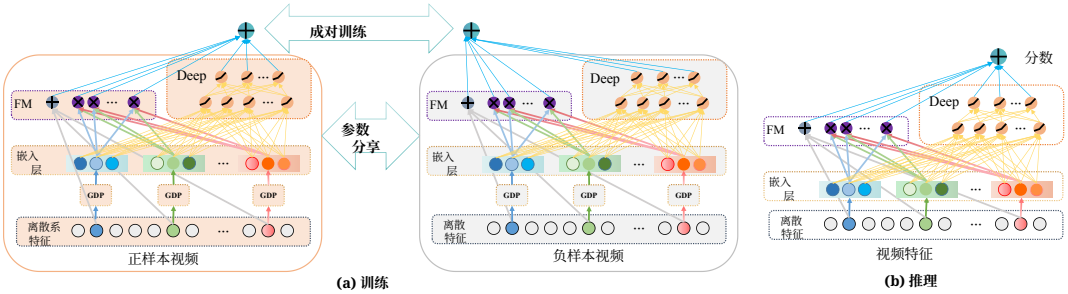

图 2：RankDFM 的整体训练与推理框架。

这些样本对由处于不同层级的项目选出。我们选择这种方法来构造偏序样本对。总之，给定一个查询q，我们根据第三种方法，为其对应的集合 $S^R_{\mathcal{M}}$ 构造偏序样本对。

2.1.2 特征处理。 众所周知，现实世界中的搜索数据流包含大量缺失数据和离群值。我们尝试了许多流行技术来缓解这一问题，例如填充 0 值、填充平均值、删除等，但由于样本特征空间的复杂性，这些方法都被证明效果不佳。视频的不同特征可能服从不同的分布，且数据区间各不相同，这一事实使任务变得更加复杂。因此，要以一致的方式处理众多方面颇具挑战性。为了减轻这些负面影响，有必要对特征进行精细处理。具体而言，首先，为了解决特征分布不一致的问题，我们针对不同分布的特征采用不同的归一化方法，例如对均匀分布特征采用最大值和最小值归一化，对长尾分布特征采用基于对数的归一化等，将所有特征值标准化到 0-1 区间，以便后续进一步处理。接下来，我们实现了一种特征离散化机制，以减少离群值和缺失值在模型训练过程中的干扰。此外，该机制还可以增强模型的泛化能力。简而言之，我们分别将离群值和缺失值放入不同的桶中，然后使用等宽 w 对正常特征值进行离散化，其中 width w 是一个可调节的超参数。

2.1.3 模型训练。 为了减少参数并适配在线推理过程，在训练阶段，我们采用 DeepFM 作为骨干网络，构建参数共享的孪生网络。假设存在许多偏序样本对 P $=<\mathrm{I}_p,\mathrm{I}_n>$，在训练过程中，正样本 DeepFM 塔将对偏序样本对 P 中的正样本 I_p 进行打分，而负样本 DeepFM 塔将对偏序样本对 P 中的负样本 I_n 进行打分：

$$S_p = Pos_DeepFM(\mathcal{I}_p, q, \mathcal{I}_p \times q) \tag{1}$$

$$S_n = Neg_DeepFM(\mathcal{I}_n, q, \mathcal{I}_n \times q) \tag{2}$$

其中，正样本 DeepFM 塔 Pos_DeepFM 与负样本 DeepFM 塔 Neg_DeepFM 的参数是共享的。

对于特定的偏序样本对，我们的训练目标是让正样本的深度 DeepFM塔得分高于正样本负的DeepFM塔对负样本进行评分。此外，模型可以学习查询 q中所有视频的偏序关系。形式化地：

$$\mathcal{L}_{pair} = min(max(S_n - S_p + \sigma), 0) \tag{3}$$

其中 L^{pair} 是基于成对式训练范式的训练目标。同时，为了缓解模型排序分数的过度波动，并将打分固定在有意义的范围内，我们还引入了点式损失，使 RankDFM 学习精排层的分数。最终，我们的训练目标如下：

$$\mathcal{L} = \alpha\mathcal{L}_{pair} + (1-\alpha)\mathcal{L}_{point} \tag{4}$$

$$\mathcal{L}_{point} = min((S_n - r_n)^2 + (S_p - r_p)^2) \tag{5}$$

其中， r_n 和r_p 分别是精排层对负样本和正样本的评分。

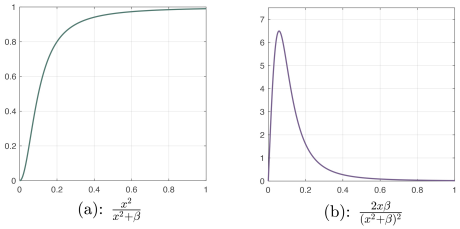

图 3：g_β（x）和 g'_β（x）的图。 $\beta = 0.01$

2.1.4 模型剪枝。 粗排阶段不仅要求有效性，还要求效率。我们采用 DeepFM 作为骨干网络来构建粗排模型，以确保模型的有效性。然而，DeepFM 的高复杂度也使粗排阶段不堪重负，因此模型剪枝极其必要。我们采用 GDP 机制进行剪枝。具体而言，GDP 使用 $g_\beta(x)$ 作为极化函数，这是因为 $g_\beta(x)$ 具有良好的特性：函数值是精确的 0、1 值，并且根据 $g_\beta(x)$ 关于 x 的导数 $g'_\beta(x)$，函数值只能稳定在 0、1 点（如图 3 所示）。我们将 $g_\beta(x)$ 作为门控函数插入模型的二阶及更高阶交互模块中，经过一个训练周期后，即可获得 $g_\beta(x)$ 值为 0 的特征交互，即不重要的交互。

$$g_\beta(x) = \frac{x^2}{x^2+\beta}, g'_\beta(x) = \frac{2x\beta}{(x^2+\beta)^2} \tag{6}$$

其中， β 是一个足够小的数。我们从模型中移除不重要的交互后，模型的效率显著提升，而模型的有效性并未显著降低。

2.2 RankDFM 的推理

在在线推理阶段，我们不需要构造偏序样本对，只需将样本特征输入模型进行评分，以便在粗排阶段进行排序和过滤。

3 实验

3.1 数据集

我们引入了两个数据集来评估 RankDFM 的性能：Kse0820 和 Kse0720。为了评估 RankDFM 在长尾查询和随机查询上的整体性能，每个数据集都被划分为两个子数据集：随机和长尾。每个子数据集包含5000 个不同的查询，以及与每个查询对应的匹配层召回结果。每个数据集包含约 1000 万个视频。

3.2 实验设置

3.2.1 基线。

- ***RankNet.*** 一种传统的基于概率的成对式排序算法。
- ***Gbrank.***一种基于 XGBoost 的成对式排序算法。
- ***COLD***。阿里巴巴集团提出的一种基于 DNN 的预排序框架。

3.2.2 指标。 ***Recall@N***。 预排序模型排序后的前 600 个项目与排序模型给出的前 N 个项目的重合率。 $N = 10$，50，600。

*3.2.3*参数配置。 下面给出一些详细的参数配置。深度侧的 DeepFM 结构为 $input \times 128 \times 64 \times 32 \times 16 \times 1$，特征的嵌入大小为 3，桶的宽度 w 为 0。02，用于平衡不同损失的参数 α 值为0。9，学习率为 0。0001。

3.3 实验结果

如表 1所示，我们提出的RankDFM 在两个数据集上均取得了最先进的性能。此外，我们还对 RankDFM 的若干变体进行了对比实验，结果如表 2 所示。值得注意的是，我们提出的点式损失能够约束模型评分并提升模型的有效性，而模型经过国内生产总值机制剪枝后的功效并未实质性降低。

3.4 敏感性分析

我们在 Kse0820 数据集上针对参数 w进行了敏感性实验。如图 4 所示，实验结果表明，模型对参数 w具有稳定性，并且在 $w = 0.02$时性能最佳。

表 1：不同方法在预排序上的性能比较。

基线	指标	Kse0820		Kse0720	
		随机	长尾	随机	长尾
Ranknet	Recall@10	0.9217	0.9116	0.9122	0.8991
	Recall@50	0.8976	0.8528	0.8790	0.8488
	Recall@600	0.6246	0.5981	0.5998	0.5726
Gbrank	Recall@10	0.9566	0.9316	0.9457	0.9229
	Recall@50	0.9217	0.8847	0.9068	0.8765
	Recall@600	0.6521	0.6202	0.6231	0.6078
COLD	Recall@10	0.9662	0.9442	0.9615	0.9396
	Recall@50	0.9381	0.8925	0.9227	0.8916
	Recall@600	0.6684	0.6351	0.6420	0.6213
RankDFM	Recall@10	**0.9738**	**0.9507**	**0.9679**	**0.9443**
	Recall@50	**0.9444**	**0.9080**	**0.9292**	**0.8991**
	Recall@600	**0.6753**	**0.6412**	**0.6493**	**0.6288**

表 2：RankDFM 不同变体的性能比较。

方法	指标	Kse0820		Kse0720	
		随机	长尾	随机	长尾
RankDFM（没有意义）	Recall@10	0.9721	0.9495	**0.9679**	0.9431
	Recall@50	**0.9444**	0.9058	0.9270	0.8969
	Recall@600	0.6737	0.6388	0.6468	**0.6288**
RankDFM（国内生产总值）	Recall@10	0.9729	0.9501	0.9660	0.9439
	Recall@50	0.9420	0.9069	0.9278	0.8984
	Recall@600	0.6745	0.6403	0.6481	0.6255
RankDFM	Recall@10	**0.9738**	**0.9507**	0.9664	**0.9443**
	Recall@50	0.9425	**0.9080**	**0.9292**	**0.8991**
	Recall@600	**0.6753**	**0.6412**	**0.6493**	0.6261

表 3：所提出的预排序模型和基础预排序模型的在线有效性和效率。

模型	HPCR	观看次数	CPU(%)	延迟(ms)
base	+0%	+0%	60	3
RankDFM	/	/	80	15
RankDFM（国内生产总值）	+0.127%	+0.398%	65	5

3.5 在线A/B测试

我们进行了为期 7 天的在线A/B测试，以验证 RankDFM 的有效性和效率。从播放次数和 HPCR（首页点击率）这两项关键指标可以看出，我们的模型为整个系统带来了显著的正向收益，而 CPU 使用率和延迟这两项效率指标与基线模型相同。

4 结论

本文提出了一种独特且高效有效的预排序框架 RankDFM。RankDFM 以 DeepFM 模型为骨干，旨在增强模型的特征交叉能力、提高其效能，并基于成对式训练范式构建用于部分排序学习的孪生 DeepFM 塔。

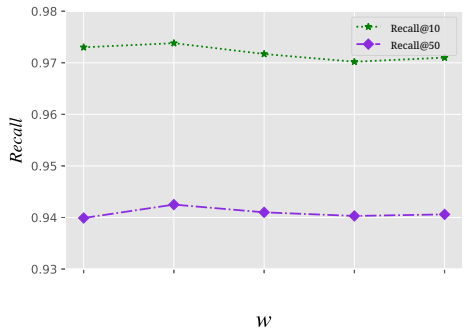

图 4：工业信息检索系统级联架构示意图。

与此同时，为了克服 DeepFM 模型效率低下的问题，我们采用 GDP 剪枝机制对模型进行剪枝，从而缩短模型的推理时间。通过充分的离线和在线实验，证明了我们提出的 RankDFM 的有效性。

参考文献

[1] Qiwei Chen、Huan Zhao、WeiLi、Pipei Huang 和 Wenwu Ou。2019。阿里巴巴电子商务推荐的行为序列 Transformer。*CoRR* abs/ 1905 06874 2019 1905 06874 1905 06874 。（）。arXiv：。[2] 2016 http://arxiv.org/abs/ 。Paul Covington、Jay Adams和 Emre Sargin。。用于 YouTube 推荐的深度神经网络。载于第*10*届*ACM* 推荐系统会议论文集，美国马萨诸塞州波士顿，*2016* 年*9*月*15*日至*19*日。ACM，191–198。https://doi.org/ 10 1145 2959100 2959190 。/ 。[3] Miao Fan、Jiacheng Guo、Shuai Zhu、ShuoMiao、MingmingSun 和 Ping Li。2019。MOBIUS：迈向百度赞助搜索中的下一代查询-广告匹配。载于第*25*届*ACMSIGKDD*知识发现与数据挖掘国际会议论文集，*KDD2019*，*2019*年*8*月*4*日2509 2517 1011453292500美国阿拉斯加州安克雷奇，*8*月—。ACM, 3330551 –。https://doi.org/ 。/。[4]

Huifeng Guo、Ruiming Tang、YunmingYe、ZhenguoLi和Xiuqiang He。2017。DeepFM：一种基于因子分解机的用于点击率预测的神经网络。载于第*26*届国际人工智能联合会议论文集。AAAI Press，1725–1731。[5] Yi Guo、Huan Yuan、Jianchao Tan、ZhangyangWang、SenYang 和Ji Liu。2021。GDP：通过具有可微极化的门控实现稳定的神经网络剪枝。载于 *2021 IEEE/CVF*国际计算机视觉会议，*ICCV2021*，加拿大魁北克省蒙特利尔，*2021*年*10*月*10*日至*17*日。IEEE，5219–5230。https: //doi.org/ 10 1109 48922 2021 00519 。/ICCV 。 。[6] Po-Sen Huang、Xiaodong He、Jianfeng Gao、LiDeng、Alex Acero 和 Larry P.Heck。2013。利用点击数据学习用于网络搜索的深层结构化语义模型。载于 第*22*届*ACM*信息与知识管理国际会议，*CIKM'13*，美国加利福尼亚州旧金山，*2013*年*10*月*27*日至*11*月*1*日。ACM，2333–2338。https://doi.org/10.1145/2505515.2505665[7] Xiang Li、Xiaojiang Zhou、Yao Xiao、Peihao Huang、Dayao Chen、ShengChen 和 Yunsen Xian。2022。AutoFAS：预排序系统的自动特征与架构选择。载于 *KDD '22*：第*28*届 *ACMSIGKDD* 知识发现与数据挖掘会议，美国华盛顿特区，*2022*年*8*月*14*日至*18*日。ACM, 3241 3249 10 1145 3534678 3539083 –。https://doi.org/ 。/ 。[8] Zequn Lyu、Yu Dong、Chengfu Huo和 Weijun Ren。2020。用于个性化点击率预测的深度匹配排序模型。载于 第三十四届 *AAAI* 人工智能会议，*AAAI 2020*，第三十二届创新人工智能应用会议，*IAAI 2020*，第十届*AAAI* 人工智能教育进展研讨会，*EAAI 2020*，美国纽约州纽约市，*2020*年*2*月*7*日至*12*日。AAAI Press，156–163。https://doi.org/index。 5346 php/AAAI/article/view/[9] Xu Ma、Pengjie Wang、Hui Zhao、Shaoguo Liu、Chuhan Zhao、Wei Lin、Kuang-Chih Lee、Jian Xu和Bo Zheng。2021。迈向预排序中效果与效率之间更佳的权衡：一种基于可学习特征选择的方法。载于 *SIGIR '21*：第*44*届 *ACM SIGIR* 信息检索研究与发展国际会议，线上会议，加拿大，*2021*年*7*月*11*日至*15*日。ACM, 2036 2040 10 1145 3404835 3462979 –。https://doi.org/ 。/。[10] Qi Pi、Weijie Bian、Guorui Zhou、Xiaoqiang Zhu 和 Kun Gai。2019。点击率预测中长序列用户行为建模实践。

在第*25*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集, *KDD 2019*, 美国阿拉斯加州安克雷奇,*2019*年*8*月*4*日至*8*日 中。ACM, 2671 2679 10 1145 3292500 3330666 – .https://doi.org/ . / . Zhe Wang, [11] Liqin Zhao, Biye Jiang, Guorui Zhou,Xiaoqiang Zhu, andKun Gai. 2020. COLD：迈向下一代预排序系统。*CoRR* abs/2007.16122 (2020)。arXiv:2007.16122[12] Wenjin Wu, Guojun Liu, Hui Ye, Chenshuang Zhang, Tianshu Wu, Daorui Xiao, Wei Lin, andXiaoyuZhu.2018. EENMF：面向电子商务赞助搜索的端到端神经匹配框架。*CoRR* abs/1812.01190 (2018)。arXiv:1812.01190http://arxiv.org/abs/1812.01190[13] Lixin Zou, Shengqiang Zhang, Hengyi Cai, DehongMa, Suqi Cheng, Shuaiqiang Wang, Daiting Shi, Zhicong Cheng, and Dawei Yin. 2021. 基于预训练语言模型的百度搜索排序。在 *KDD '21*：第*27*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议，虚拟会议，新加坡，*2021*年*8*月*14*日至*18*日 中。ACM, 4014–4022。https://doi.org/10.1145/3447548.3467147