arXiv:2305.01513v1 [cs.IR] 1 May 2023

使用集成方法和神经信息检索检索比较论证

CLEF 2021 Touche 实验室论据检索笔记本

维多利亚切卡利娜[1,2], 亚历山大 潘琴科[1]

[1]斯科尔科沃科学技术学院，莫斯科，俄罗斯联邦[2]*Philips Innovation Lab Rus*，莫斯科，俄罗斯联邦

摘要

在本文中，我们介绍了针对 Touché 实验室任务 2（比较问题的论据检索）的提交成果 [1, 2]。我们的团队 Katana 提出了多种基于决策树集成算法的方法，根据比较文档的相关性和论证支持对其进行排名。我们使用 PyTerrier [3] 库，将集成模型应用于排序问题，并考察文本统计特征以及基于比较结构的特征。我们还采用了 BERT [4], 等大型上下文化语言建模技术来解决所提出的排序任务。为了将该技术与排序建模结合起来，我们使用了神经排序库 OpenNIR [5]。

我们的系统显著优于提出的基线，并在竞赛官方评测指标中，相关性排名第一、质量排名第二（评测指标为 NDCG@5 分数）。所提出的模型有助于提升信息检索和对话系统处理比较查询的性能。

关键词

比较论据检索，自然语言处理，神经信息检索

1. 引言

人们每天都会面临在两个实体之间做出选择的问题：哪款手机更可靠，哪种果汁含糖量更低，哪家酒店更适合度假。领域专用的比较系统，如 WolframAlpha 或 Diffen，只能部分解决这一问题，而且依赖结构化数据，因此其适用范围有限。

另一方面，万维网包含大量观点和客观论证，有助于比较决策过程。这促使人们需要开发一种能够处理此类信息的开放领域通用系统。问题在于如何从一组文档中检索相关、支持性且可信的论证。本研究旨在从 ClueWeb12 [1] 语料库中检索文档并对其重新排名，同时考虑支持或反对某一选项的论证。

我们工作的贡献如下：我们首次将基于统计特征与比较特征混合的集成方法用于文档排序；我们首次使用
用于论据检索任务的神经信息检索方法；我们提出了一种模型

CLEF 2021 – 评测论坛会议与实验室，*2021*年*9*月*21*日至*24*日，布加勒斯特，罗马尼亚© 2021 年权所有。根据知识共享署名 4.0 国际许可协议（CC BY 4.0），允许使用。CEUR研讨会论文集（CEUR-WS.org）

[1] http://lemurproject.org/clueweb12

其表现优于基线，并根据相关性和质量指标分别取得了第一名和第二名的结果。

2. 相关工作

与本研究最相关的是此前的共享任务 Touche 2020 [6]。共有 17 名参与者参加了竞赛，并提交了 41 次运行结果。参与者测试了多种方法，包括提取与主张和前提相对应的结构、评估论证质量、使用语言模型表示文档，以及通过相似词扩展查询。搜索引擎 ChatNoir [7]基于 BM25F [8] 方法的排序函数被用作基线。

提交的解决方案中，只有少数能够略微改善基线。此前竞赛中总体表现最佳的方法是先扩展查询，再根据相关性、可信度和支持性质量对文档进行重新排序 [9]。

本研究基于我们在上一届 Touche 共享任务中提交的运行结果 [10]。在本研究中，我们使用预训练语言模型来确定查询与文档之间的相关性。提取比较结构并统计文档中比较句的数量，有助于我们评估相关论证的质量。

因此，论据检索问题也出现在其他场景中。比较论证机器 CAM [11] 根据认可对象和比较方面检索比较句。论文 [12] 探讨了上下文对论证检测系统的影响，并证明了上下文能够提升系统的性能。

3. 数据集与实验设计

3.1. 数据集

组织者提供了50个比较问题（主题），对于这些问题，我们应获取包含支持或反对某一选项的有力论据的文档。竞赛主题可在线获取。 [2]

此外，还提供了上一年度竞赛的 50 个主题及其对应的相关性标注 [13]，用于监督学习。这些文档同样从 ChatNoir 中检索，并由人工进行排名，得分分别为 0（不相关）、1（相关）和 2（高度相关）。我们使用这些数据，基于决策树训练和构建模型，并对 BERT 排序器进行微调。此外，上一年度各参赛团队提交的结果也可供使用。

遗憾的是，这些数据不足以用于拟合大型监督式排序模型，例如基于 BERT 技术的模型。在这种情况下，我们使用相邻的问答数据集 ANTIQUE [14]。该数据集由 Yahoo! Webscope L6 中的问题和答案组成，包含 2,626 个开放领域非事实型问题和 34,011 条人工相关性标注。

查询及其排序结果示例见表6、表7 （见附录A）。值得注意的是，Antique 数据集使用了不同的排序得分集合：0、1、2，而非 1、2、3、4，因此我们按照以下映射重新编排 Antique 的排名 1→0、 2→1、 3→1、 4→2。

[2] https://webis.de/events/touche-21/shared-task-2.html

3.2. 评估设置

我们将每个主题作为查询输入 ChatNoir [7] 搜索引擎，并从 ClueWeb12 语料库中提取最多 100 篇不重复的文档。我们清理文档正文中的 HTML 标签和标记，并使用下文介绍的一种已开发方法对其进行排序。

作为辅助数据，组织者提供了上一年竞赛的主题。对于每个拟定主题，我们从 ChatNoir 检索了一组文档，并按照上述方法进行了标注。我们使用这些数据训练待开发的模型并验证组合方法。在验证阶段，我们将排序后的数据划分为训练集和验证集，分别包含 40 个和 10 个主题。

在运行阶段，我们在万维网评估平台 Tira [15]上执行生成的解决方案。在此阶段，为了拟合模型，我们使用上一年的全部排序数据，并预测当前拟定主题的排名。运行结果根据提交运行的人工判断，使用 NDCG 指标进行评估。检索到的文档依据两个标准进行判断：(i) 文档相关性；(ii) 是否提供了充分的论证支持 [16]。

4. 使用树集成进行文档排序

本节中，我们使用树集成这一监督机器学习技术来解决排序问题。我们选择逐点回归树算法（如随机森林），或提升树算法（如 XGBoost 和 LightGBM）。对于 LightGBM 模型，我们采用 LambdaMART [17] 目标函数。它结合了两种目标：一种是通过最小化排序中的逆序数得出的代价函数（LambdaRank [18]），另一种是用于构建梯度提升决策树（MART [19]）的目标函数。我们使用 PyTerrier 平台进行信息检索。[3] 它简化了文本特征的提取，并支持描述检索实验 [20]。

4.1. 特征提取

对于我们的排序机器学习方法，我们使用了来自以下三种来源的特征：(i) PyTerrier 提取的排序特征，(ii) 特定的比较特征，(iii) ChatNoir 基于自定义 BM25 评分函数计算的分数。[4]

4.1.1. PyTerrier 提取的特征

PyTerrier 提供多种模型来衡量查询-文档文本的匹配程度。这些模型包括统计度量（TF-IDF）、基于语言模型的度量（Heimstra、Diriclet），以及根据词项出现字段衡量文档的度量（BM25F、PLF）。所有可用模型的列表见 cite [5]。在这些模型中，我们选择了 BM25、Heimstra、DFIC、DPH、TF-IDF、DiricletLM 和 PL2 进行探索。

[3] https://pyterrier.readthedocs.io/en/latest/index.html [4] https://www.elastic.co/guide/en/elasticsearch/reference/current/index-modules-similarity.html [5] http://terrier.org/docs/current/javadoc/org/terrier/matching/models/package-summary.html

表1：PyTerrier模型中文本特征在验证集上的结果。

方法 - 分数	NDCG@5
BM25	**0.3637**
Heimstra	0.3616
DFIC	**0.3642**
DPH	0.3110
TF-IDF	**0.3637**
DiricletLM	0.3307
PL2	0.3703

我们将选定的每种方法依次独立应用于训练集，根据得到的得分对文档进行排名，并在验证集上评估排序结果。这些测试的结果见表 1。我们选择了结果最有前景的 3 种方法，这 3 种方法结合了 3 个特征。

4.1.2. 比较特征

我们的重点不仅是查找高度相关的文档，还包括查找对一个对象相对于另一个对象进行比较的文档。该研究 [21]假设，比较问题可以由比较结构表示，即比较对象、比较方面和谓词。我们采用了
引文中提出的序列标注模型，并将其应用于查询。该模型有助于我们为每个主题定义比较对象。随后，我们将该模型应用于文档，得到一个比较特征集。

特征 `is_retrieved` 描述文档中是否存在任何比较结构。特征 `objs_score` 定义了在文档中找到查询中的多少个对象（0、1 或 2）。特征 `asp_pred_score` 按以下方式计分：如果查询中的至少一个对象出现在文档中，则文档中每个被标注为方面或谓词的词都会使得分增加 0.5。最后，我们将定义的特征与 ChatNoir 系统获得的得分相结合，得到查询-文档对的特征向量 {`score_pl2`, `score_tf`,`score_bm`, `score_dfic`, `baseline_scores`, `is_retrieved`, `ap_score`, `objs_score`}。

4.2. 模型

4.2.1. 随机森林

我们使用从 Sklearn 导入并封装在 PyTerrier 流水线中的随机森林模型。为找到最佳配置，我们将估计器数量从 10 调整到 150，其中 20 个估计器取得了最佳验证分数 NDCG@5 0.408。

4.2.2. XGBoost

我们还将 Sklearn 中的梯度提升库封装为 PyTerrier 类，并通过将学习率从 $1e{-4}$ 调整到 0.1、将最大深度从 4 调整到 16 来调优超参数。最佳配置为学习率 0.01、最大深度为 6，取得 NDCG@5 0.547。

4.2.3. LightGBM

对于 LightGBM，我们将叶节点数量从 8 调整到 20，并将学习率从 0.001 调整到 0.1。最佳配置的 num_leaves $= 15$ 和学习率为 $= 0.1$，得分为 0.579。

最终模型的特征重要性见表 2。可以看出，最重要的特征是从 ChatNoir 检索到的得分，其次是基于卡方的独立性偏离 [22]以及文档中是否存在比较对象。

5. 基于 BERT 的神经信息检索文档排序

BERT 等上下文化语言模型在排序任务中效率更高，因为它们包含语言单元之间的大量关系。在所提出的工作中，我们使用基于“原始”Transformer 架构的 OpenNIR [5] 重排序模型。

5.1. 文本表示

BERT 接收查询和文档，并对二者进行联合处理。BERT 重排序器的一个显著特征是每一层都使用注入令牌相似度矩阵，这显著提升了性能 [23]。

表 2所提出的 LightGBM 模型中的特征重要性

特征	PI2	TF-IDF	BM25	Dfic	ChatNoir	包含比较对象	对象_得分	方面_谓词
重要性	1.76	1.19	1.51	2.3	20.8	0	1.66	1.51

5.2. 训练过程

首先，我们在 Antique 数据集上预训练该重排序器。我们清理了数据集中的错误符号和化妆品。我们还从数据集中保留了长度超过 300 个字符的文档，因为 ChatNoir 检索到的内容通常不超过 300 个字符。训练过程持续了 500 个训练轮次，学习率为 0.001，每个批次包含 56 个对象。最终，我们的模型在验证集上的 NDCG@5 达到 0.3362。随后，我们使用上一年的 40 个训练主题，按照相同的配置对模型进行 50 个训练轮次的微调。微调后，模型在验证集上的得分提升至 0.412。

6. 结果

6.1. 验证集上的结果

表 3：验证集上的结果。

方法	NDCG@5	时间，毫秒
随机森林	0.408	127.168
XGBoost	0.547	128.848
LightGBM	**0.572**	131.244
BERT 排序器	0.412	1560.947
Baseline’20	0.534	-

表格展示了每种提出的方法在上一年度竞赛数据的验证部分上取得的结果：3。我们还评估了上一年度的基线在验证集上的表现。LightGBM 模型取得了最高得分，也优于基线。XGBoost 的得分较低；作为一种简单算法，随机森林的得分最低。BERT 略胜于随机森林。

在右侧列中，我们还加入了训练每个模型所需的时间。可以看出，基于集成的模型具有大致相同的时间复杂度，而 BERT 需要更多训练时间。

6.2. 测试集上的结果

在最终测试中，检索文档由人工标注，得分范围为 0 到 3。评估依据两项独立标准进行：文档与给定主题的相关性，以及文本质量。质量标准包括良好的语言风格、易读性、恰当的句子结构，以及没有拼写错误和头韵。

每项标准都有一个单独的文件，其中包含评估员的得分。两项评估的结果见表4 和表5。就相关性而言，我们团队 Katana 的各次提交在所有团队中表现最佳；就文本质量而言，则排名第二。

与验证集上的情况一样，XGBoost 和 LightGBM 的表现最佳。这一点可以很好地解释，因为这些模型的损失基于排序质量函数：XGBoost 使用 NDCG，LightGBM 使用 LambdaMART。第一个模型对相关性的描述略好（0.489），在所有参赛者中排名第一。相反，LightGBM 在质量方面表现更好，得分为 0.684，在质量榜单中排名第二，略逊于 Top 1。随机森林方法在这两项上的得分都略低于基线。这可以用其集成构建算法较为简单来解释。BERT 在质量方面表现相当不错，但相关性方面较弱。训练时使用了相邻任务（事实型问答）的数据，这或许是解决方案不够准确的原因。

表 4
Katana 的相关性运行结果的 NDCG@5 得分
团队、基线和 Top-2 方法的相关性运行
NDCG@5 得分

方法	NDCG@5
随机森林	0.393
XGBoost（第一名）	**0.489**
LightGBM	0.460
BERT 排序器	0.091
ChatNoir基线	0.422
Thor团队（前二）	0.478

表 5 Katana团队、基线和第一名方法的运行质量 NDCG@5 得分

方法	NDCG@5
随机森林	0.630
XGBoost	0.675
LightGBM（前二）	**0.684**
BERT 排序器	0.466
ChatNoir基线	0.636
Rayla团队（Top 1）	0.688

7. 结论

本文介绍了我们针对论证检索共享任务提出的解决方案。我们重点关注集成方法，并采用统计方法、语言建模和比较结构提取来检索相关特征。我们还使用了基于 BERT 技术的神经重排序器，以便在任务中利用上下文化模型的信息。

最佳结果由采用排序代价函数进行训练的梯度提升方法取得：XGBoost 和 LightGBM。所提出的方法优于基线，分别在相关性和质量排序中位列第一和第二。BERT 上下文化模型表明，大规模训练数据是必需的。

致谢

本研究部分由优先项目“RATIO：稳健论证机器”（SPP 1999）中的项目“ACQuA：用论证回答比较问题”（资助编号 BI 1544/7-1 和 HA 5851/2-1）提供支持。我们感谢 Maik Frobe 在 TIRA系统中为软件运行提供支持。

附录A：训练数据示例

表6Touche任务数据集中相关性不同的查询和文档示例

查询	文档	Rank
什么更好 对环境而言， 真的还是假的 圣诞树?	疾病与病症内容由我们的医学审查委员会审阅 真的还是人造的？关于……有很多令人困惑的信息 哪种对你的健康和环境更有益。	2
	你可能以为自己是在保护树木，但塑料替代品也有问 题。那么，人造圣诞树和真圣诞树，哪种更“环保”?	1
	本文是环保星期五系列共103篇中的第25篇：11月 28日的小贴士——圣诞树：在选择真圣诞树时陷入两难 ，还是选一棵假树?	0

表7Antique 数据集中相关性不同的查询和文档示例

查询	文档	Rank
为什么我们要在 knife、knob、knee 这些词的 开头加上字母 k?	它们是撒克逊词汇。Knife 原本会读作 ker-niff。	4
	我猜，历史上“kn”的发音应该是 不同于“n”，而随着时间推移，单词的发音方式已经发生了变- 化 .	3
	因为英语是一种很奇妙的语言。	2
	我也不太（k）知道！	1

附录 B：排名结果示例

在本附录中，您可以找到根据 LightGBM 和基线方法排名的排名前三的文档示例。

表 8与查询“斯坦福大学的录取率是否高于麻省理工学院？”相关性不同的文档示例

斯坦福大学的录取率是否高于麻省理工学院?	
LightGBM 前 3 名	基线前 3 名
1. 斯坦福大学和哈佛大学的录取率相近，约为 7%。 麻省理工学院的录取成功率略高，录取率略低于10%，即2015届的1742人。哈佛大学 ，斯坦福大学和麻省理工学院是文化、商业和政府政策领域的全球领导者。	1. 斯坦福大学和哈佛大学的录取率相近，约为7%。 麻省理工学院的录取率则略高，成功录取率略低于10%，即2015届录取了1742人。哈佛大学 , 斯坦福大学和麻省理工学院在文化、商业和政府政策
2. 十多年来 , 我一直担任招生官，负责麻省理工学院的招生工作。在这段时间里 , 我读过更多超过 10,000 个应用程序，并且已经观看了数千部汇-数千名新生进入麻省理工学院。这是一种荣幸，可以在最具活力、最令人兴奋的大学工作在世界上 .	2. 十多年来 , 我曾担任广告-麻省理工学院的招生官。在那段时间里 , 我已经阅读了更多内容超过 10,000 份申请，并且见证了千—数千名新生进入麻省理工学院。这是一种荣幸。在最具活力、最令人兴奋的大学工作在世界上 .
3. 我们的主要改进针对家-庭收入低于 $75,000 的家庭——使麻省理工学院学费免费并取消	3. 所有这些事实信息，加上许多其他细节，都可以在麻省理工学院的招生资-料中找到。事实上，今年，麻省理工学院将发放 7,400 万美元的本科生助学金。

表 9针对查询“哪款智能手机的电池续航时间更长：Xperia 还是 iPhone？”的不同相关性示例文档

哪款智能手机的电池续航时间更长：Xperia 还是 iPhone?	
LightGBM 前 3 名	基线 前 3 名
1. 1. 将调低设置的省电应用-电池续航时间较低时的设置，以尽可能延长使用时间尽可能延长电池的续航时间。索尼已将基准设定为-mark 配备 Xperia 内的 1200万像素摄像头S.	1. iPhone 4 是苹果迄今最薄的智能手机，但它提供了更好的屏幕和更快的处理器，视频通话，以及许多其他改进。
2. 如何延长苹果 iPhone 的电池续航时间 4s 许多拥有 iPhone 4s 的用户都有投诉关于电池续航时间。 苹果已承认这些问题，并正在着手解决。	2. 索尼 Xperia 的评价：高于平均水平的智能-电话 Gizmotraker”，就电池续航时间而言，】【。至于续航时间，通话时间约为 7 小时 30 分钟，待机时间为 450 小时。处于待机状态。
3. 索尼爱立信随售品附带一张 8GB 存储卡软件包中的索尼爱立信 Xperia arc s 低于平均电池续航时间。大多数用户的续航时间约为 24 Xperia 的续航时间。X27 的 1600 毫安时电池在需要充电之前，但重度用户可能需要此前的一次电力注入。	3. 如何延长苹果 iPhone 的电池续航时间 4s 许多拥有 iPhone 4s 的用户都有投诉关于电池续航时间。 苹果已承认这些问题，并正在努力解决。

参考文献

[1]A. Bondarenko、L. Gienapp、M. Fröbe、M. Beloucif、Y. Ajjour、A. Panchenko、C. Biemann、B. Stein、H. Wachsmuth、M. Potthast、M. Hagen，《Touché 2021概览：论据检索》，载于：D. Hiemstra、M.-F. Moens、J. Mothe、R. Perego、M. Potthast、F. Sebastiani（编），《信息检索进展》。第43届欧洲信息检索研究会议（ECIR 2021），《计算机科学讲义》第12036卷，施普林格，柏林、海德堡、纽约，2021年，第574–582页。网址：https://link.springer.com/ chapter/10.1007/978-3-030-72240-1_67。doi: 10.1007/978-3-030-72240-1_67.[2]A. Bondarenko、L. Gienapp、M. Fröbe、M. Beloucif、Y. Ajjour、A. Panchenko、C. Biemann、B. Stein、H. Wachsmuth、M. Potthast、M. Hagen，《Touché 2021概览：论据检索》，载于：G. Faggioli、N. Ferro、A. Joly、M. Maistro、F. Piroi（编），《CLEF 2021评测实验室工作札记论文》，CEUR研讨会论文集，2021年。[3] S. MacAvaney、C. Macdonald、N. Tonellotto，《从词袋到BERT及更进一步的信息检索：使用PyTerrier和OpenNIR的ECIR 2021实践实验教程》，载于《第43届欧洲信息检索研究会议论文集》，2021年，第728–730页。[4] J. Devlin、M.-W. Chang、K. Lee、K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，载于《2019年计算语言学协会北美分会会议：人类语言技术论文集》，第1卷（长短篇论文），计算语言学协会，明尼阿波利斯，明尼苏达州，2019年，第4171–4186页。网址：https://www.aclweb.org/anthology/N19-1423。doi:10.18653/v1/N19-1423.[5] S. MacAvaney，《OpenNIR：完整的神经信息检索排序流程》，载于：J. Caverlee、X. B. Hu、M. Lalmas、W. Wang（编），《WSDM '20：第十三届ACM国际网络搜索与数据挖掘会议》，美国德克萨斯州休斯敦，2020年2月3日至7日，美国计算机协会，2020年，第845–848页。网址：https://doi.org/10.1145/3336191.3371864。doi:10.1145/3336191.3371864.[6] A. Bondarenko、M. Fröbe、M. Beloucif、L. Gienapp、Y. Ajjour、A. Panchenko、C. Biemann、B. Stein、H. Wachsmuth、M. Potthast、M. Hagen，《Touché 2020概览：论据检索》，2020年，第384–395页。doi:10.1007/978-3-030-58219-7_26.[7] M. Potthast、M. Hagen、B. Stein、J. Graßegger、M. Michel、M. Tippmann、C. Welsch，《Chat-Noir：面向ClueWeb09语料库的搜索引擎》，载于：B. Hersh、J. Callan、Y. Maarek、M. Sanderson（编），《第35届ACM信息检索研究与发展国际会议（SIGIR 2012）》，美国计算机协会，2012年，第1004页。doi:10.1145/2348283.2348429.[8] S. E. Robertson、H. Zaragoza、M. J. Taylor，《多加权字段的简单BM25扩展》，载于：D. A. Grossman、L. Gravano、C. Zhai、O. Herzog、D. A. Evans（编），《2004年ACM CIKM国际信息与知识管理会议论文集》，美国华盛顿特区，2004年11月8日至13日，美国计算机协会，2004年，第42–49页。网址：https://doi.org/10.1145/1031171.1031181。doi:10.1145/1031171.1031181.[9] T. Abye、T. Sager、A. J. Triebel，《用于回答比较问题的开放领域网络搜索引擎》，载于：L. Cappellato、C. Eickhoff、N. Ferro、A. Névéol（编），《CLEF 2020评测论坛会议与实验室工作札记》，希腊塞萨洛尼基，2020年9月22日至25日，《CEUR研讨会论文集》第2696卷，CEUR-WS.org，2020年。

网址：http://ceur-ws.org/Vol-2696/paper_130.pdf。[10] V. Chekalina、A. Panchenko，《使用深度预训练语言模型和自然语言理解检索比较论据》，载于：L. Cappellato、C. Eickhoff、N. Ferro、A. Névéol（编），《CLEF 2020评测论坛会议与实验室工作札记》，希腊塞萨洛尼基，2020年9月22日至25日，《CEUR研讨会论文集》第2696卷，CEUR-WS.org，2020年。网址：http://ceur-ws.org/Vol-2696/paper_210.pdf。[11] M. Schildwächter、A. Bondarenko、J. Zenker、M. Hagen、C. Biemann、A. Panchenko，《回答比较问题：能否胜过十条蓝色链接？》，载于：L. Azzopardi、M. Halvey、I. Ruthven、H. Joho、V. Murdock、P. Qvarfordt（编），《2019年人类信息交互与检索会议论文集》，CHIIR 2019，英国苏格兰格拉斯哥，2019年3月10日至14日，美国计算机协会，2019年，第361–365页。网址：https://doi.org/10.1145/3295750.3298916。doi:10.1145/3295750.3298916.[12] M. Fromm、E. Faerman、T. Seidl，《TACAM：主题与上下文感知论证挖掘》（2019），第99–106页。网址：https://doi.org/10.1145/3350546.3352506。doi:10.1145/3350546.3352506.[13] A. Bondarenko、M. Fröbe、M. Beloucif、L. Gienapp、Y. Ajjour、A. Panchenko、C. Biemann、B. Stein、H. Wachsmuth、M. Potthast、M. Hagen，《Touché 2020概览：论据检索》，载于：L. Cappellato、C. Eickhoff、N. Ferro、A. Névéol（编），《CLEF 2020评测实验室工作札记论文》，《CEUR研讨会论文集》第2696卷，2020年。网址：http://ceur-ws.org/Vol-2696/。[14] H. Hashemi、M. Aliannejadi、H. Zamani、W. B. Croft，《ANTIQUE：非事实型问答基准》，12036（2020），第166–173页。网址：https://doi.org/10.1007/978-3-030-45442-5_21。doi:10.1007/978-3-030-45442-5_21.[15] M. Potthast、T. Gollub、M. Wiegmann、B. Stein，《TIRA集成研究架构》，载于：N. Ferro、C. Peters（编），《变化世界中的信息检索评估》，《信息检索丛书》，施普林格，柏林、海德堡、纽约，2019年。doi:10.1007/978-3-030-22948-1_5.[16] L. Braunstain、O. Kurland、D. Carmel、I. Szpektor、A. Shtok，《在社区问答网站中支持寻求建议问题的人工回答》，载于：N. Ferro、F. Crestani、M. Moens、J. Mothe、F. Silvestri、G. M. D. Nunzio、C. Hauff、G. Silvello（编），《信息检索进展：第38届欧洲信息检索研究会议（ECIR 2016）》，意大利帕多瓦，2016年3月20日至23日。论文集，《计算机科学讲义》第9626卷，施普林格，2016年，第129–141页。网址：https://doi.org/10.1007/978-3-319-30671-1_10。doi:10.1007/978-3-319-30671-1_10.[17] Q. Wu、C. J. Burges、K. M. Svore、J. Gao，《针对信息检索指标调整bboosting》，《信息检索》13（2010），第254–270页。网址：https://www.microsoft.com/en-us/research/publication/adapting-boosting-for-information-retrieval-measures/。[18] C. Burges、R. Ragno、Q. Le，《使用非光滑代价函数进行学习排序》，载于：B. Schölkopf、J. Platt、T. Hoffman（编），《神经信息处理系统进展》第19卷，麻省理工学院出版社，2007年。网址：https://proceedings.neurips.cc/paper/2006/file/af44c4c56f385c43f2529f9b1b018f6a-Paper.pdf。[19] J. Friedman，《随机梯度提升》，《计算统计与数据分析》38（2002），第367–378页。doi:10.1016/S0167-9473(01)00065-2.[20] C. Macdonald、N. Tonellotto，《使用PyTerrier进行信息检索的声明式实验》，载于：K. Balog、V. Setty、C. Lioma、Y. Liu、M. Zhang、K. Berberich（编），《ICTIR '20：

2020 年 ACM SIGIR 信息检索理论国际会议，线上活动，挪威，2020 年 9 月 14–17 日，美国计算机协会，2020 年，第 161–168 页。网址：https://dl.acm.org/doi/10.1145/3409256.3409829。[21] V. Chekalina、A. Bondarenko、C. Biemann、M. Beloucif、V. Logacheva、A. Panchenko，《深度学习中 Python 和 MATLAB 哪个更好？自然语言中的比较问题作答》，载于：欧洲计算语言学协会第十六届会议论文集：系统演示，计算语言学协会，线上，2021 年，第 302–311 页。网址：https://www.aclweb.org/anthology/2021.eacl-demos.36。[22] I. Kocabas、B. Dincer、B. Karaoğlan，《基于独立性偏离度量的信息检索非参数项加权方法》，《信息检索》（2013 年），第 1–24 页。doi：10.1007/s10791-013-9225-4。[23] S. MacAvaney、A. Yates、A. Cohan、N. Goharian，CEDR：用于文档排序的上下文化嵌入（2019 年），第 1101–1104 页。网址：https://doi.org/10.1145/3331184.3331317。doi：10.1145/3331184.3331317。