·专题:ChatGPT 与人工智能技术应用·

大语言模型时代下的信息检索研究发展趋势

赵 鑫 窦志成 文继荣*

中国人民大学 高瓴人工智能学院,北京 100872

[摘 要] 以 ChatGPT 为代表的大语言模型带来了人工智能技术的新一轮发展浪潮,获得了广泛的社会关注。大语言模型通过大规模无标注数据预训练、指令微调、人类对齐等关键技术途径,学习到了丰富的世界知识,具有较好的文本理解与生成能力,能够有效求解各种复杂任务。这一重要技术进展对于信息检索领域的发展带来了新的机遇。本文从大语言模型对于已有信息检索架构的改进以及现有检索技术如何改进大语言模型两个方面进行阐述,针对相关科学问题的可行技术方法进行了梳理与展望,探讨大语言模型时代下的信息检索发展趋势,旨在推动信息检索领域的科研进步。

[关键词] 大语言模型;信息检索;生成式检索

语言是人类特有的能力,对于人类社会进步与个体发展具有重要意义。维特根斯坦在《逻辑哲学论》提到:“我的语言的界限意味着我的世界的界限”。有效理解并且使用语言对于人工智能技术的发展极为重要。早在 20 世纪 50 年代,图灵测试就定义了基于自然语言对话的人工智能能力测试范式,旨在评估人工智能是否达到了人类智能的水平。在随后几十年时间里,相关研究者和实践者都在努力推进这一方向的研究。

最近,以 ChatGPT 为代表的大语言模型在自然语言理解与生成方面取得了重要研究进展[1-3]。在 2022 年 11 月底推出以来,ChatGPT 得到了整个社会的高度关注,随后发布的 GPT-4 甚至被相关学者认为是通用人工智能的早期体现[4]。一般来说,大语言模型是指尺寸超过一定量级的预训练语言模型(例如“百亿参数”)[1],它的训练与学习一般包括两个关键阶段。首先,基于海量无标注文本数据进行预训练,通过大规模预训练任务编码语料中所包含的语义信息,进而掌握大量的世界知识[2,3]。其次,通过专门设计的指令微调策略(Instruction Tuning)[5],使得语言模型对于自然指令具有更好的遵循能力,进一步需要引入基于人类偏好与反馈的强化学习方法(Reinforcement Learning from Human Feedback, RLHF),加强与人类价值观的对齐[6]。与传统语言模型相比,大语言模型具有更强的语言理解与生成能力,拥有丰富全面的世界知识,能够进行复杂任务推理与求解[3,4,7,8]。此外,大语言模型展现出了较强的指令遵循能力,能够有效理解自然语言提示(Prompt),进而有效完成相关任务,并且具有一定的新任务泛化能力。

大语言模型所带来的能力跃升对于信息技术的发展正产生着重要影响,有望深刻改变人们获取信息的方式,促进信息获取技术的更新升级。实际上,

文继荣 教授,博士生导师,中国人民大学信息学院院长,高瓴人工智能学院执行院长,新一代智能搜索与推荐教育部工程研究中心主任。研究领域为信息检索、人工智能。

赵鑫 中国人民大学高瓴人工智能学院教授,博士生导师,获得国家自然科学基金优秀青年科学基金项目资助。研究领域为信息检索、自然语言处理。

收稿日期:2023-06-28;修回日期:2023-08-17

* 通信作者,Email: jrwen@ruc.edu.cn

解决信息过载问题，提升人类获取信息的能力和效率，一直是学术界和工业界共同关注的研究方向。从 20 世纪 90 年代开始，以“搜索引擎”为代表应用的信息检索技术得到了快速发展[9-14]，极大地提升了人类从海量互联网数据中获取信息的效率。近年来，搜索引擎提供支持的功能逐步丰富，但是仍然沿用经典的检索范式：给定基于关键词的用户查询，搜索引擎高效地从海量的文档中检索到和该查询需求相关的文档，并按照相关性排序后返回给用户。通常来说，检索系统分为离线和在线两个阶段。在离线阶段，对文档进行预处理并构建索引（包括早期的倒排索引[9, 10]以及近年来的向量索引[13, 14]）。在在线阶段，检索系统接收到用户查询后，首先进行用户查询理解，并将理解处理后的查询送入索引中，通过检索模型（如经典的 BM25 等概率检索模型[9]或者基于神经网络的检索模型[11, 12]）计算文档的相关性，召回最相关的 TopK 候选文档，然后再采用较为复杂、精细的精排模型对候选文档进行排序后输出。这种以索引为核心的“索引—召回—精排”检索架构被广泛应用在各种信息检索系统中。

以 ChatGPT 为代表的生成式大模型和以搜索引擎为代表的检索模型是两种不同的信息获取方式。传统的检索模型侧重于“检索”，可以从海量的互联网（或其他信息源）中获取准确的信息，但是对于检索结果通常不做深入分析，当用户信息需求比较复杂时，需要用户浏览多个结果才能获取所需要的信息。而生成式大模型则是将大量知识存储在参数化的模型中，可以直接根据用户的问题生成答案，能够更便捷地满足用户的信息需求，但是由于返回信息是生成的，可能会存在虚假、陈旧或错误信息。将两种检索范式的优势进行融合与互补，打造更为高效、准确的信息获取技术，具有重要的科学价值与应用意义。

针对上述问题，本文将重点围绕以下两个方面展开讨论（如图 1 所示）。首先，大模型在自然语言理解和生成上的强大能力能够提升信息检索系统的综合性能并且有望实现新的检索范式，这个方面将分为两个部分进行介绍，包括对于传统信息检索架构中核心技术的性能提升（第 1 节）以及生成式检索范式（第 2 节）；其次，传统信息检索技术也可以有效地改善生成模型所面临的技术问题，如幻象、时效性、私有化等问题（第 3 节）。

1　大语言模型赋能传统信息检索范式

大语言模型研究的快速发展，对于传统信息检索技术将会带来重要的性能改进与提升。本部分将探讨如何将大语言模型技术用于改进现有信息检索技术（表 1）。

1.1　对于复杂用户信息需求理解的加强

在现代搜索引擎中，用户查询不再局限于简单的关键字检索，对于复杂任务的查询需求（如“某国 2022 年 GDP 比世界平均 GDP 的增幅比率”）日益增多。大语言模型初步展现出了一定自主智能能力（如 AutoGPT），可对于复杂用户信息需求进行有效解析，从而生成完整的查询解决方案（如生成多步的查询规划[16, 17]）。进一步，ReAct 方法在问答任务中通过提示大语言模型生成与任务相关的推理文本，并根据需求生成搜索动作[15]。此外，还可以借助大语言模型仿真用户的检索行为（称之为“生成式智能体”[26]），进而丰富用户的检索行为或者用于大规模用户实验（User Study）的推广。例如，可以收集用户的历史查询偏好，借助这些查询偏好建立基于大语言模型的仿真搜索智能体，可用于研究用户搜索行为、搜索引擎评测等问题。此外，由于类 ChatGPT 模型对于上下文对话的支持能力较好[1]，可以用来加强交互式检索、会话式检索等面向用户需求理解的检索任务。

1.2　对于复杂文档信息的语义理解能力提升

随着信息技术的不断发展，互联网数据的形式日趋复杂，体现为内容呈现形式多样（如包含表格、文档等）、文档内容长度显著增加、多语言数据并存等。在已有的索引结构中，基于关键字的表示方法和稠密表示方法对于复杂文档的语义建模能力相对较弱，不能够有效捕捉复杂文档的语义信息。与传

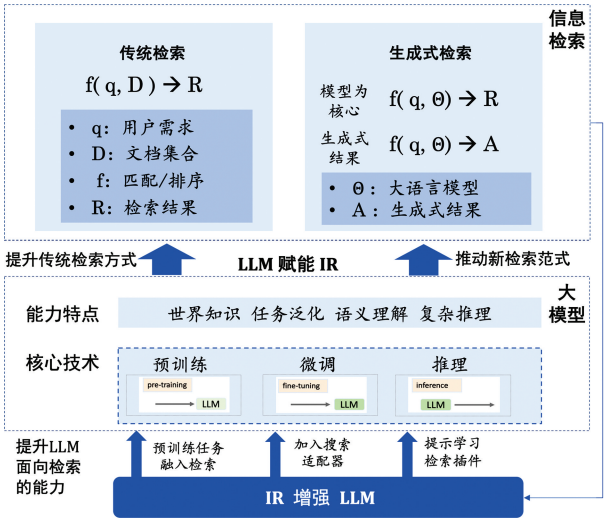

图 1　大模型时代下的信息检索研究

表 1 大语言模型赋能传统检索范式的代表性工作

类别	代表性工作	技术特点
对于复杂用户信息需求理解的加强	AutoGPT	生成多步查询规划
	ReAct[15]	生成任务相关的推理
	DSP[16]	将问题拆解为流水线程序
	Toolformer[17]	使用大模型进行行动决策
对于复杂文档信息的语义理解能力提升	InPars[18]	大模型生成文档相关的查询
	LLM-AUGMENTER[19]	基于外部知识生成回复
对于相关度匹配和排序模型的增强	Permutation generation[20]	大模型的相关性排序
	LLMRank[21]	自然语言模式实现用户推荐
	LRL[22]	不使用领域特定数据的精排
	InPars-v2[23]	使用大模型进行数据增强
面向检索结果的答案抽取与摘要生成	Dynamic reference[24]	大模型动态地参考检索结果
	Self-Prompting[25]	利用参考文档生成伪问答对

统模型相比，大语言模型的语义建模能力有了显著提升，使用了多种形式（如表格、代码等）和多个语种的文本数据进行预训练，同时增加了输入窗口的长度限制，能够有效提升对于复杂文档语义信息的理解与表示能力[3]。为了更好地理解用户的信息需求，还可以借助大语言模型针对文档信息生成可能相关的查询[18]，进而提升面向用户信息需求的文档语义表示能力[19]。

1.3 对于相关度匹配和排序模型的增强

在建立用户信息需求表示和文档语义表示后，需要设计对应的匹配函数（或重排序函数），实现候选文档的相关度打分。在这一过程中，对于个性化的用户偏好建模非常重要。传统方法往往通过设置额外的个性化模型辅助匹配函数。大语言模型提供了一种更为通用的解决途径，可以统一建模个性化历史和候选文档重排序两个步骤：将两种信息都表述成自然语言，构建基于统一提示语句的重排序模型[20, 21]，并可以通过列表形式（Listwise）直接对候选文档进行排序[22]。进一步，现有研究表明[5]，指令微调能够有效激活大语言模型的任务处理能力，并且带来一定的指令泛化能力，可以借助大规模的检索特设指令，对于检索任务的相关能力进行诱导提升。此外，匹配和重排序函数需要依靠大量的真实用户标注数据进行训练。然而在冷启动场景或者垂直应用领域，难以收集充足的标注数据。可以使用大语言模型生成伪标注训练数据（如为索引文档生成可能的相关查询[23]），缓解标注数据稀疏问题[14]。

1.4 面向检索结果的答案抽取与摘要生成

传统搜索引擎主要关注文档层面的检索结果，只要返回相关文档或段落就认为其解决了用户的信息需求。对于一些特定查询（如开放性查询），如果能够直接返回相关答案，将会减少用户的查询和阅读时间，提升检索过程中的用户效率。为方便讨论，这里假设相关文档已经被搜索引擎召回，核心问题是如何从该文档中进行有效的答案抽取与摘要。大语言模型在自然语言处理任务中具有优异的表现，可以将这一任务转化为阅读理解或者文档摘要问题，从而使用大语言模型抽取相关答案或对于文档内容进行摘要。在答案提取过程中，可以由大模型自我判断决定何时使用检索进行回答，从而动态地利用检索结果[24]，也可以使用上下文学习的方式让模型更好地抽取相关答案[25]。此外，针对面向查询的文档摘要，还可以将其应用于改进检索结果中页面摘要（Snippet）的生成质量。

2 基于大模型的生成式检索方式

除了改进已有检索框架外，大模型技术还有望突破传统框架的束缚，进而实现新的信息检索范式。下面将围绕两个新的信息检索范式展开讨论。

2.1 以模型为核心的检索

前文介绍了以索引为核心的“索引—召回—精排”的分阶段信息检索架构。这一经典框架涉及多个阶段，需要联合多个模块进行工作，无法实现端到端的性能优化，特别是索引的构建过程无法和检索模型进行联合优化。为了解决这一问题，近年来，研

究人员提出了以模型为核心的新式检索架构(Model Based IR)[27-30]。这种新架构抛弃了传统的文档索引和检索方式，转而训练一个基于 Seq2Seq 编码—解码结构的生成模型。给定输入查询，模型对查询进行编码后，直接调用解码器生成相关文档的标识符。可以认为，模型隐式地记忆了从查询到文档标识符的映射关系，不需要单独的索引，从查询到文档的检索过程直接通过该生成模型完成，因此这种以模型为核心的方法也一般称为生成式检索。以模型为中心的检索架构的好处是可以直接与下游任务进行联合优化。相关前沿工作初步实现了这种以模型为核心的检索架构，取得了不错的检索性能[27-30]，其中最有代表性的是 Google 提出的 DSI。DSI 尝试使用了整数、数字符号序列和基于层次化聚类的树状编码三种标识符体系，设计了多种序列生成方式来训练基于 T5 的生成模型并在 Natural Question 数据集上验证了这种生成式检索的效果。因为这一研究方向才刚刚起步，在文档标识符体系设计、模型的高效更新、面向排序的效果优化等关键问题上仍然需要深入探索。

2.2　基于生成式结果的信息检索

传统的信息检索技术重点聚焦“检索”，但是基于列表的搜索结果形式并不能直接返回查询答案。在用户查询某一话题时，如果能给为用户返回一篇与查询相关的知识性文章(类似维基百科页面)，将能够有效提升用户获取信息的效率，改善用户满意度。此类以生成式内容作为查询结果的工作得到了研究人员们的关注，代表性工作包括 WebGPT[31] 和 WebBrain[32]。WebGPT 模仿人使用搜索引擎完成复杂信息获取任务的过程，基于用户在这个过程中产生的行为数据，微调了 GPT-3 模型；而 WebBrain 则采用了海量的维基百科的文章数据，分别训练相关文献的检索模型，以及基于文献生成最终内容的生成模型。

3　信息检索技术赋能大语言模型

尽管大语言模型具有很强的能力，它们仍然存在着很多缺陷[4]，包括幻觉问题(生成虚假内容)、时效性问题(受到预训练语料限制)等。这些问题很难通过改善大模型的内在机制来完全消除。为了应对这一挑战[40-46]，众多研究工作都采用了检索来强化大模型：通过检索让大模型访问到相关知识，并通过检索到的知识来引导大模型生成更为准确的内容，进而在一定程度上缓解幻觉和时效性不足的问题。根据检索使用阶段的不同，现有该方向的研究可以分为三种：(1) 在大模型训练阶段融合检索，让大模型具备原生检索能力；(2) 通过轻量化微调来让大模型具备更好的使用检索结果的能力；(3) 基于大模型的零样本能力，直接在生成阶段设计提示指令融合检索结果。

3.1　融合原生检索能力的大模型训练方法

融合检索的方式之一是在大模型训练阶段就引入记忆检索能力[40-43]。这能够让大模型更好的理解与使用检索返回的内容，将通用世界知识内置在模型中，并通过检索的方式从外部动态引入事实知识。假设给定一个大规模文档库，可以通过预处理的方式，为训练语料中的每句话提前检索出文档库中的相似文档，然后将这些文档使用合适的方式与当前上下文融合，进而预测出后续内容。这种方式的好处是可以有效控制大模型的参数量，让大模型通过原生的检索能力从外部文档库和知识库引入事实知识。缺点是模型的训练过程相比于普通大模型的训练更为复杂。

3.2　基于搜索适配器(Adapter)微调来增强大模型的检索能力

在预训练阶段融合原生检索能力的学习代价很高，即使是微调大语言模型的成本也需要较大的算力开销。因此，在微调阶段，一般采用即插即用的适配器(小型神经网络组件)的方式，以少量参数、高效且经济地微调大语言模型[44]。这种微调方式不更新大模型的原始参数，也确保了多个任务之间的独立性。适配器可以通过插入额外的模型层的方式来

表 2　生成式检索的代表性工作

类别	代表性工作	技术特点
以模型为核心的检索	DSI[28]	基于 T5 实现可微索引
	DSI-QG[33]	设计查询生成任务增强生成模型的训练
	NCI[34]	采用了前缀感知的动态可适配解码器
	Ultron[35]	采用基于 PQ 和 URL 的文档标识符
	SEAL[36]	生成子字符串作为文件标识符
	DSI++[37]	探讨了生成式检索中文档动态添加的问题
	AutoTSG[38]	采用词项集作为文档标识符
基于生成式结果的信息检索	WebGPT[31]	微调 GPT 来模仿人类使用搜索的行为，生成整合后的答案
	WebBrain[32]	基于海量维基数据训练的带有引用信息的短文档生成
	WebCPM[39]	基于交互式网页搜索的问答模型框架

实现，也可以将一些与任务相关的特定的锚标记插入到输入序列中来实现。使用搜索适配器的方式，一方面降低了资源要求，同时通过面向特定任务的优化尽可能提升了模型在使用检索组件时的效果。

3.3 使用提示学习在生成过程中融合检索结果

在大模型使用阶段，可以通过提示（Prompting）的方式将检索结果融合到上下文中[45, 46]。这种方式不需要在本地训练或者微调大模型，直接调用黑盒大模型的接口即可。近期，OpenAI 也发布了检索插件（Retrieval Plugin），支持语义检索来帮助用户访问个人或者企业文档。然而，目前的工作仍然停留在简单使用提示信息的阶段，在检索结果的有效性、面向生成的思维链抽取、检索和生成中的虚假信息规避等众多问题上，仍需要深入研究。此外，这种方式可能会受到上下文长度限制的影响，当检索内容的长度过长时，可能需要对文档内容进行截断。为了缓解这一问题，近期多个大模型也推出了上下文长度拓展的加强版本。

4 开放性研究问题

（1）面向信息检索的可信内容生成：尽管大语言模型展现出了强大的语言生成能力，然而它的可信生成问题仍然具有非常大的研究挑战，所产生的生成幻象等问题难以被有效解决，这一问题在信息获取场景下将产生非常严重的危害。此外，大语言模型还缺乏对于能力边界的自我感知能力，在搜索应用中存在着较大的风险与限制。目前解决幻象的主要途径包括基于人类反馈的对齐微调以及检索增强的推理生成：前者通过基于人类反馈的强化学习算法减少幻象的发生，但实现较为复杂、需要大量人工参与；后者基于检索到的相关信息源，在推理阶段减少幻象的发生，但需要依赖外部信息源，受限于模型利用外部信息的能力。

（2）高效的信息检索大模型架构：大语言模型的训练和推理目前都存在着严重的性能问题，在真实部署中将会面临着较大挑战，需要设计专门的优化与适配算法来提升部署效率。目前提升大模型训练和推理的主流技术分别为轻量化微调和模型量化：轻量化微调与全参数微调的差距仍然比较明显，模型量化与任务目标的协同优化还有待深入探究。此外，传统搜索引擎架构在性能上的优化已经趋于成熟，如何利用与整合已有架构来减少新型大模型信息获取技术的部署成本还需要进一步研讨，此部分还缺乏相关研究工作。

（3）信息检索技术的可靠评测方法：从早期的 TREC 发布的数据集[47]到最近的 MSMARCO 数据集[48]，大部分信息检索评测集合都是采用人工标注生成，然后通过对于标准答案的匹配进行方法性能评测，进而在后续实验中可以复用大规模标注数据。然而，这一方法并不适合开放性的查询任务评测，特别是当大语言模型在预训练语料中已经学习相关文本数据，不能有效评估基于生成技术的检索性能。此外，大模型对于查询的回答通常以句子形式给出，当查询的标准答案较为复杂时，很难通过自动化的途径精确解析和评判大模型所给出答案文本。因此，针对复杂查询检索能力的评测仍然是一个开放性研究问题。

5 总 结

大语言模型对于人工智能领域产生了重要影响，对于科学研究提供了一种非常有效的技术路径，有望推动信息技术的升级与变革。本文聚焦信息检索这一研究领域，分别从大语言模型对于现有信息检索范式与技术的提升以及现有检索技术如何改进大语言模型两个方面展开了相关阐述，总结介绍了相关技术途径以及发展趋势，并对于存在的研究问题进行了相关讨论。目前，大语言模型在信息检索上的研究方兴未艾，未来可探索的空间还比较宽广。中国是信息检索技术发展的一支重要科技力量，相信通过我国学者的持续努力和深入合作，能够在这一领域内不断做出创新性的研究工作。

参 考 文 献

[1] OpenAI. Introducing ChatGPT. (2022-11-30)/[2023-06-27]. https://openai.com/blog/chatgpt.

[2] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. (2020-05-28)/[2023-06-27]. https://arxiv.org/pdf/2005.14165.pdf.

[3] Zhao WX, Zhou K, Li JY, et al. A survey of large language models. (2023-03-31)/[2023-06-27]. https://arxiv.org/pdf/2303.18223.pdf.

[4] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. (2023-03-22)/[2023-06-27]. https://arxiv.org/pdf/2303.12712.pdf.

[5] Chuang HW, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. (2022-10-20)/[2023-06-27]. https://arxiv.org/pdf/2210.11416.pdf.

[6] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. (2022-03-04)/[2023-06-27]. https://arxiv.org/pdf/2203.02155.pdf.

[7] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. (2022-06-15)/[2023-06-27]. https://arxiv.org/pdf/2206.07682.pdf.

[8] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 2022: 24824—24837.

[9] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613—620.

[10] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333—389.

[11] Huang PS, He XD, Gao JF, et al. Learning deep structured semantic models for web search using clickthrough data// Proceedings of the 22nd ACM international conference on Information & Knowledge Management. New York: Association for Computing Machinery, 2013: 2333—2338.

[12] Guo JF, Fan YX, Pang L, et al. A deep look into neural ranking models for information retrieval. Information Processing & Management, 2020, 57(6): 102067.

[13] Qu YQ, Ding YC, Liu J, et al. RocketQA: an optimized training approach to dense passage retrieval for open-domain question answering. (2020-10-16)/[2023-06-27]. https://arxiv.org/pdf/2010.08191.pdf.

[14] Zhao WX, Liu J, Ren RY, et al. Dense text retrieval based on pretrained language models: a survey. (2022-11-27)/[2023-06-27]. https://arxiv.org/pdf/2211.14876.pdf.

[15] Yao SY, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models. (2022-10-06)/[2023-06-27]. https://arxiv.org/pdf/2210.03629.pdf.

[16] Khattab O, Santhanam K, Li XL, et al. Demonstrate-Search-Predict: composing retrieval and language models for knowledge-intensive NLP. (2022-12-28)/[2023-06-27]. https://arxiv.org/pdf/2212.14024.pdf.

[17] Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: language models can teach themselves to use tools. (2023-02-09)/[2023-06-27]. https://arxiv.org/pdf/2302.04761.pdf.

[18] Bonifacio L, Abonizio H, Fadaee M, et al. InPars: unsupervised dataset generation for information retrieval// Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Association for Computing Machinery, 2022: 2387—2392.

[19] Peng BL, Galley M, He PC, et al. Check your facts and try again: improving large language models with external knowledge and automated feedback. (2023-02-24)/[2023-06-27]. https://arxiv.org/pdf/2302.12813.pdf.

[20] Sun WW, Yan LY, Ma XY, et al. Is ChatGPT good at search? investigating large language models as re-ranking agent. (2023-04-19)/[2023-06-27]. https://arxiv.org/pdf/2304.09542.pdf.

[21] Hou YP, Zhang JJ, Lin ZH, et al. Large language models are zero-shot rankers for recommender systems. (2023-05-15)/[2023-06-27]. https://arxiv.org/pdf/2305.08845.pdf.

[22] Ma XG, Zhang XC, Pradeep R, et al. Zero-shot listwise document reranking with a large language model. (2023-05-03)/[2023-06-27]. https://arxiv.org/pdf/2305.02156.pdf.

[23] Jeronymo V, Bonifacio L, Abonizio H, et al. InPars-v2: large language models as efficient dataset generators for information retrieval. (2023-01-04)/[2023-06-27]. https://arxiv.org/pdf/2301.01820/pdf.

[24] Ren RY, Wang YH, Qu YQ, et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation. (2023-07-20)/[2023-08-16]. https://arxiv.org/pdf/2307.11019.pdf.

[25] Li JL, Zhang ZS, Zhao H. Self-prompting large language models for zero-shot open-domain QA. (2022-12-16)/[2023-06-27]. https://arxiv.org/pdf/2212.08635.pdf.

[26] Park J, O'Brien JC, Cai CJ, et al. Generative agents: interactive simulacra of human behavior. (2023-04-07)/[2023-06-27]. https://arxiv.org/pdf/2304.03442.pdf.

[27] Zhou YJ, Yao J, Dou ZC, et al. DynamicRetriever: a pre-trained model-based IR system without an explicit index. Machine Intelligence Research, 2023, 20(2): 276—288.

[28] Tay Y, Tran VQ, Dehghani M, et al. Transformer memory as a differentiable search index. Advances in Neural Information Processing Systems. 2022: 21831—21843.

[29] Metzler D, Tay Y, Bahri D, et al. Rethinking search. ACM SIGIR Forum, 2021, 55(1): 1—27.

[30] Ren RY, Zhao WX, Liu J, et al. TOME: a two-stage approach for model-based retrieval. (2023-05-18)/[2023-06-27]. https://arxiv.org/pdf/2305.11161.pdf.

[31] Nakano R, Hilton J, Balaji S, et al. WebGPT: browser-assisted question-answering with human feedback. (2021-12-17)/[2023-06-27]. https://arxiv.org/pdf/2112.09332.pdf.

[32] Qian HJ, Zhu YT, Dou ZC, et al. WebBrain: learning to generate factually correct articles for queries by grounding on large web corpus. (2023-04-10)/[2023-06-27]. https://arxiv.org/pdf/2304.04358.pdf.

[33] Zhao SY, Ren HX, Shou LJ, et al. Bridging the gap between indexing and retrieval for differentiable search index with query generation. (2022-06-21)/[2023-06-27]. https://arxiv.org/pdf/2206.10128.pdf.

[34] Wang YJ, Hou YY, Wang HN, et al. A neural corpus indexer for document retrieval. (2022-06-06)/[2023-06-27]. https://arxiv.org/pdf/2206.02743.pdf.

[35] Zhou YJ, Yao J, Dou ZC, et al. Ultron: an ultimate retriever on corpus with a model-based indexer. (2022-08-19)/[2023-06-27]. https://arxiv.org/pdf/2208.09257.pdf.

[36] Bevilacqua M, Ottaviano G, Lewis P, et al. Autoregressive search engines: generating substrings as document identifiers. (2022-04-22)/[2023-06-27]. https://arxiv.org/pdf/2204.10628.pdf.

[37] Mehta SV, Gupta J, Tay Y, et al. DSI++: updating transformer memory with new documents. (2022-12-19)/[2023-06-27]. https://arxiv.org/pdf/2212.09744.pdf.

[38] Zhang PT, Liu Z, Zhou YJ, et al. Term-sets can be strong document identifiers for auto-regressive search engines. (2023-05-23)/[2023-06-27]. https://arxiv.org/pdf/2305.13859.pdf.

[39] Qin YJ, Cai ZH, Jin D, et al. WebCPM: interactive web search for Chinese long-form question answering. (2023-05-11)/[2023-06-27]. https://arxiv.org/pdf/2305.06849.pdf.

[40] Izacard G, Lewis P, Lomeli M, et al. Atlas: few-shot learning with retrieval augmented language models. (2022-08-05)/[2023-06-27]. https://arxiv.org/pdf/2208.03299.pdf.

[41] Guu K, Lee K, Tung Z, et al. REALM: retrieval-augmented language model pre-training. (2020-02-10)/[2023-06-27]. https://arxiv.org/pdf/2002.08909.pdf.

[42] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens. (2021-12-08)/[2023-06-27]. https://arxiv.org/pdf/2112.04426.pdf.

[43] Li JY, Tang TY, Zhao WX, et al. The web can be your oyster for improving language models// Findings of the Association for Computational Linguistics: ACL 2023. Toronto: Association for Computational Linguistics, 2023: 728—746.

[44] Hu EJ, Shen YL, Wallis P, et al. LoRA: low-rank adaptation of large language models. (2021-06-17)/[2023-06-27]. https://arxiv.org/pdf/2106.09685.pdf.

[45] Shuster K, Komeili M, Adolphs L, et al. Language models that seek for knowledge: modular search & generation for dialogue and prompt completion. (2022-03-24)/[2023-06-27]. https://arxiv.org/pdf/2203.13224.pdf.

[46] Lazaridou A, Gribovskaya E, Stokowiec W, et al. Internet-augmented language models through few-shot prompting for open-domain question answering. (2022-03-10)/[2023-06-27]. https://arxiv.org/pdf/2203.05115.pdf.

[47] Harman, D. Overview of the first TREC conference// Proceedings of the 16th annual international ACM SIGIR conference on Research and development in Information Retrieval. New York: Association for Computing Machinery, 1993: 36—47.

[48] Bajaj P, Campos D, Craswell N, et al. MS MARCO: a human generated machine reading comprehension dataset. (2016-11-28)/[2023-06-27]. https://arxiv.org/pdf/1611.09268.pdf.

The Development of Information Retrieval in the Era of Large Language Model

Wayne Xin Zhao Zhicheng Dou Ji-Rong Wen*

Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872

Abstract Large language models (LLMs) exemplified by ChatGPT has sparked a new wave of development in artificial intelligence technology and received widespread social attention. By exploring a series of technical approaches (i.e., massive unsupervised pre-training, instruction tuning, and human alignment), LLMs can effectively encode world knowledge and possess excellent language understanding and generation ability, making it possible to solve various complex tasks via a unified model. The technical advance of LLMs has brought new opportunities for the development of the information retrieval (IR) field. This article will elaborate it on two major aspects: how to enhance existing IR techniques by LLMs and how existing IR techniques can benefit LLMs. We systematically review and summarize feasible technical approaches to key research issues, and discuss how IR techniques would be shaped by LLMs, which aims to better explore the development path of IR field in the era of LLMs.

Keywords large language models; information retrieval; generative retrieval

(责任编辑 崔国增 姜钧译)

* Corresponding Author, Email: jrwen@ruc.edu.cn