计算机与自主智能研究进展
Advances in Computer and Autonomous Intelligence Research
第 2 卷◆第 3 期◆版本 1.0◆2024 年
文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

大语言模型在垂直搜索引擎中的应用

黄颖　潘仁前　焦森
中国电子科技集团公司第二十八研究所
DOI:10.12238/acair.v2i3.8633

[摘　要] 现阶段,大语言模型具有高质量完成自然语言的理解、分析、生成等特性,对垂直搜索引擎中的问题理解、增强搜索等方面产生积极影响,同时与知识图谱相结合有助于进一步提升搜索结果的相关性和深度。
[关键词] 大语言模型；垂直搜索引擎；知识图谱；自然语言处理；数据挖掘
中图分类号：H0　文献标识码：A

Large Language Modeling in Vertical Search Engines

Ying Huang　Renqian Pan　Miao Jiao
Twenty-eighth Research Institute of China Electronics Technology Group Corporation
[Abstract] At this stage, big language models have the property of completing the understanding, analysis, and generation of natural language with high quality, which positively affects the problem understanding and enhanced search in vertical search engines, and at the same time, the combination with knowledge graph helps to further improve the relevance and depth of search results.
[Key words] big language modeling; vertical search engine; knowledge graph; natural language processing; data mining

引言

随着信息量的迅猛增长,垂直搜索引擎因其具备的专业化特质与查询结果的高度相关性而备受用户喜爱。然而,随着用户对海量文档的搜索深度与结果精确度需求的不断提升,对传统搜索技术提出了严峻的挑战[1]。近年来,大语言模型的兴起为破解该难题开辟了新途径,以ChatGPT为代表的生成式大语言模型,在自然语言理解、自然语言处理等方面性能出色[2],仅使用少量示例数据,即可高质量地完成自然语言理解和生成类任务。在面对海量文本搜索需求时,将大语言模型应用于垂直搜索引擎中,不失为一种有益的尝试。

1 大语言模型

2022年11月30日,OpenAI发布了生成式大语言模型ChatGPT[3],相较于常规模型,ChatGPT具有千亿参数,具备流畅的文本生成能力,能够撰写新闻稿,模仿人类叙事,创作诗歌,初步验证了通过海量数据和大量参数训练出来的大语言模型能够迁移到其他类型的任务,展示了生成式模型的强大功能。随后发布的GPT-4模型,在提高语言处理能力基础上,增加了图像识别、提取能力,可回答数学、编程、视觉、药物、法律和心理等通识问题,被认为是一个早期的通用人工智能系统[4]。

在大语言模型时代,自然语言处理能力大幅提升,除了能高质量完成自然语言生成类任务,还具备以生成式框架完成各种开放域自然语言理解任务的能力。只需要将模型输出转换为任务特定的输出格式,无需针对特定任务标注大量的训练数据,ChatGPT即可在少量样本的情况下达到令人满意的性能。

大语言模型由于具备上下文学习能力、可观的通用知识容量、较好的泛化性和较强的复杂推理能力,在应用端落地的定位理解,便从“替代”转化为“赋能”。因此,将大语言模型与搜索相结合,在理解问题、回答准确度和个性化等方面都存在明显优化,可以大幅降低处理信息复杂度、提升用户体验。

2 传统垂直领域信息检索的弱点

信息检索旨在根据用户的查询意图,从现有的资源库(通常为数据库、自然语言文本等)中寻找相关的信息资源返回给用户。作为克服信息过载最重要的技术之一,信息检索系统已成为人类获取知识信息的重要途径[5],但传统垂直领域的信息检索仍存在一些技术问题：

传统的垂直领域搜索引擎一般基于领域内专业知识库进行搜索,通常分为数据库类和自然语言文本类,一般采用数据库查询机制、关键词匹配及排序算法等搜索技术体系,通过对用户查询关键词进行语义解析,融合预设的索引架构与排序准则,从海量数据资源中筛选提取关联信息[6]。然而,这种技术策略的局限性在于难以深刻领悟用户查询意图,特别是当面

临复杂或含糊的查询请求时,所呈现结果的相关度与精确性存在较低的情况。

此外,传统的垂直领域搜索引擎在处理自然语言文本时,通常采用信息抽取的方式建立检索知识库,依据预设的知识结构,抽取目标知识(例如实体识别、关系抽取、事件抽取等)。面对不断涌现的应用场景,该方式只能采用“看到一类,定义一类,构建一类”的模式构建知识库,效率较低。由于知识表示存在限制,只能对已定义的知识类型进行抽取,无法对难以定义结构的知识类型进行抽取；已有知识抽取过程包含多个中间子任务(如实体/事件识别、实体/事件消歧、关系抽取、属性抽取等),这种序列化且多步处理的抽取范式往往会引起错误传递和积累问题,抽取准确率会随着知识结构复杂度的升高而急剧下降[7]。

3 大语言模型在垂直领域信息检索中的应用

3.1问题理解

传统的信息搜索引擎是一种以关键词匹配为核心的知识调用方式,因此传统的搜索入口一般是关键词,而对于自然语言表达的问句搜索也多采用实体抽取等语义分析方法抓取查询语句中的关键词。然而,将大语言模型引入垂直搜索引擎,可以更深层次地进行问题理解,并生成更符合用户期望的搜索提议及反馈。

问题理解的核心是将表达多样、具有歧义的自然语言映射为无歧义的目标语义结构。大语言模型借助“预训练大语言模型+任务微调”的典型范式,可以通过上下文学习、用户反馈和提示等语义分析模型的能力对垂直领域内的专业知识进行学习和微调。面对用户提出的含糊或开放式问题时,通过提示词数据的小样本学习,在语义分析的编码解码过程中增加语义知识和领域知识的引导[7],可以深入理解查询背后的意图及情境信息,转换为对现有搜索服务、数据服务、SQL表达式等调用模式。

因此,将大语言模型应用于垂直领域信息检索,有助于打破用户表达需求的限制,使任何人都能以更自然的方式进行信息检索。

3.2搜索结果精准定位

由于垂直领域搜索既可以对各类领域内专业文本进行全文搜索,也可以对领域数据库信息进行全库搜索,甚至将两者信息结合作为搜索目标。在使用传统搜索时,用户通常会提出相对简洁的查询问题,搜索引擎会将与搜索相关的网页链接和摘要的列表作为查询结果,如果是针对数据库的搜索,会将与关键词相关的搜索结果纳入到列表中。用户还可以通过访问超链接的形式直接访问源材料,并且通过页面展示的结果溯源内容,便于让用户看到不同信息源间信息的一致性或分歧。但是,这种传统搜索方式需要用户自行对搜索结果进行信息整合,既困难又耗时。

如果将传统搜索结果输入给大语言模型,利用大语言模型的即时信息和知识的汇总能力,可以使用户能够获得相对精准的答案,而不是深入浏览多个页面或数据库表格,自行分析总结答案；不仅如此,还可以将数据库搜索、页面搜索与文本搜索的结构结合起来。一般情况下,将排序前十的搜索结果作为语料提供给大语言模型,由大语言模型结合问题进行知识汇总,以生成式方式返回问答结果,同时对结果附有引用链接,在保证可靠性的同时便于用户溯源或深入研究。

3.3个性化用户体验

大语言模型可以通过分析用户的搜索历史、偏好和行为模式提供个性化推荐,包括内容推荐、搜索界面和交互方式的个性化设置等方面,从而提升用户满意度和忠诚度。大语言模型能够较细致地解析用户过往的搜索历史和相关情境信息,进而在提供搜索服务时实现更高层次的个性化与连贯性的对话体验。该技术的显著优势在于能精准捕捉用户的查询习惯与偏好,基于此定制化生成既满足用户独特需求又促进有效互动的检索结果与内容反馈。

首要地,大语言模型通过深入剖析用户过往的查询记录,能有效辨识出用户的兴趣焦点及查询习性。此分析超越了单一查询的直观信息,进而揭示查询背后隐藏的深层意图与用户行为模式。举例来说,模型能够察觉用户在特定时间范围内对某一议题的持续关注度,或关注用户在多样查询中的偏好倾向。鉴于上述洞见,搜索引擎得以校准其算法机制,从而提供更加个性化的搜索反馈,贴合用户的特异化需求[8]。

3.4搜索源插件化接入

在信息获取与知识服务学科范围内,大语言模型的应用正不断拓宽其功能范畴,尤其在通过插件化服务与实时信息源对接的方面。这一对接机制赋予了大语言模型接入并融合最新数据资源的能力,进而向用户提供时效性突出、精确度高的信息服务。以ChatGPT为例,其引入的插件功能不仅增强了模型在信息处理方面的效能,也极大提高了用户体验的质量。

ChatGPT的插件机制支持与各类资料库、数据源的对接。在技术实施层面,插件服务往往涵盖API(应用程序编程接口)调用及数据流的即时处理,借助此类接口与外界数据资源进行互动,及时获取信息,并运用其卓越的自然语言处理技术对这些信息进行解析与整合。这一系列操作不仅对模型提出了高效数据处理能力的要求,还强调了在处理个人或机密机构数据时,必须确保数据的安全性及隐私保护措施的可靠性。

4 未来发展方向

4.1可解释性与可控性

随着模型参数和深度的增长,模型的决策过程变得越来越难以解释,提升模型的可解释性是增强其透明度的关键,通过采用如特征重要性分析和决策树可视化这样的可解释性技术,可以深入洞察模型的预测机制；另外,提高模型的控制能力意味着可以根据不同的场合和用户的要求来调整模型的操作,以保证模型输出的准确性和灵活性。

4.2多模态融合

大语言模型的成功不断推动着多模态大模型的研究和发

展。多模态融合技术结合了文本、图像、语音等多种信息源，以提供更为全面和丰富的搜索体验。这种信息融合不但强化了信息表达的能力，还增加了搜索结果的关联度和准确度。未来，多种模态（甚至任意模态或信号）输入，能更深入地洞察不同模态数据间的固有联系，进而在数据搜索阶段给出更全面和深度的分析方案。

4.3鲁棒性和安全性

提高模型的鲁棒性涉及增强其对输入数据变化的抵抗力，确保在面对噪声或异常值时仍能稳定工作。在确保模型安全的同时，防止其产生有害的内容的关键在于实施内容的过滤和采用安全措施。

5 结论

基于深度学习与自然语言处理技术的进步，大语言模型极大地优化了垂直搜索引擎的精确度与效能，增强了检索结果的关联度与深度内涵，为用户提供更富有多样性和个性化的搜索体验。展望未来，随着技术的不断演进，大语言模型在搜索引擎领域能够产生更为广阔且深远的影响。

[参考文献]

[1]姜华.面向计算机领域的垂直搜索引擎设计[J].软件,2023,44(09):113-115.

[2]李松山.ChatGPT背景下科技期刊面临的挑战及应对策略[J].传播与版权,2023,134(19):58-61.

[3]周中元,刘小毅,李清伟,等.ChatGPT技术及其对军事安全影响[J].指挥信息系统与技术,2023,14(2):7-16.

[4]Bubeck S,Chandrasekaran V,Eldan R,etal.Sparks of artificial general intelligence:early experiments with GPT-4.2023.

[5]Zhao W X,Liu J, Ren R Y,etal.Dense text retrieval based on pretrained language models: a survey.2022.

[6]蒋慧琳.工业品垂直搜索引擎的设计与实现[D].华东师范大学,2022.

[7]车万翔,窦志成.大语言模型时代的自然语言处理:挑战、机遇与发展[J].中国科学:信息科学,2023,53(9):1645-1687.

[8]杨杰,徐越,余建桥,等.基于搜索引擎日志的用户查询意图分类[J].指挥信息系统与技术,2019,10(2):74-79.