附表1-2

南京邮电大学毕业设计(论文)开题报告

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 题　　目 | | 基于AI大语言模型的对话式搜索引擎 | | | | |
| 学生姓名 | | 李峰 | 班级学号 | B21060414 | 专业 | 网络工程 |
| 1. **对指导教师下达的课题任务的学习与理解；**   现阶段人工智能技术的快速发展，信息检索领域开始探索更加智能化的解决方案。传统搜索引擎虽能满足大部分基础检索需求，但在动态交互和深层次语义理解上的能力仍显不足。特别是对于需要多轮上下文关联和实时反馈的复杂场景，现有技术难以高效捕捉用户的真实需求。这种不足催生了基于AI大语言模型的新型搜索引擎的构想，旨在通过自然语言对话和语境推理，为用户提供更精准、更贴合需求的智能化搜索服务。  因此，如何构建一种能够通过自然语言进行对话并理解上下文的智能搜索引擎，成为信息检索领域亟待解决的核心问题。尤其在多轮对话中，如何准确理解并维持上下文信息，从而提供与用户需求更契合的答案，是对传统搜索引擎系统的一大挑战。传统搜索引擎在处理此类情境时，往往难以捕捉用户的真实意图，导致搜索结果的相关性和准确性不足，限制了其在高需求应用场景中的广泛应用。  在此背景下，本课题旨在基于AI大语言模型，设计并实现一款对话式搜索引擎。该引擎不仅是一个信息检索工具，更应具备通过自然语言与用户智能交互的能力。在用户查询的过程中，能够实时分析并理解查询意图，从而提供个性化且精准的搜索结果。通过AI大语言模型在上下文理解和对话生成上的优势，本课题旨在显著提升传统搜索引擎在处理复杂查询和多轮对话时的表现，弥补其在深度理解和个性化反馈方面的不足。  AI大语言模型，尤其是GPT系列（Generative Pre-trained Transformer），在多个领域展示了强大的自然语言处理能力，包括文本生成、问答系统和机器翻译等。这些模型不仅能理解单一问题，还能通过上下文推理生成符合语境的内容。例如，GPT- 4在多轮对话中的表现尤为突出，能够根据对话前文推断用户的真实意图并提供相应的回答。其卓越的适应性和灵活性使得该模型能够根据不同领域的数据进行训练与调整，从而提供定制化、高质量的搜索反馈，弥补传统搜索引擎在语义理解和上下文处理方面的不足。  除了GPT系列，国内也涌现出一批具有竞争力的大语言模型，如华为的盘古大模型、阿里巴巴的M6、百度的文心大模型以及腾讯的混元大模型等。这些国产模型在中文语境下的自然语言理解和生成能力日益成熟，尤其在多轮对话和深度语义理解方面取得了显著进展，逐渐成为搜索引擎和智能对话系统等领域的重要技术基础，展现出巨大的应用潜力。本课题的目标是通过结合传统搜索引擎的强大检索功能与AI大语言模型的语义理解和生成能力，突破传统搜索引擎的局限，开发一个更加智能、用户友好的对话式搜索系统。在该系统中，用户通过与搜索引擎的自然语言对话，不仅能够获得精准的搜索结果，还能够享受更加流畅和直观的交互体验。系统的智能化不仅体现在对用户查询的理解上，还体现在能够根据对话中的反馈进行动态调整与优化，从而不断提高搜索结果的相关性与个性化。例如，在多轮对话中，系统能够根据前文的交流动态调整搜索策略，提供更加贴合用户需求的答案，提升整体使用体验。  本课题旨在探索结合AI大语言模型与传统信息检索技术的方法，设计并实现一款初步具备对话功能的搜索系统，重点提升特定场景下的语义理解能力和用户交互体验。本系统通过改进搜索结果的相关性与准确性，尝试优化用户体验和检索效率，并在一定范围内实现多轮对话的交互功能，以更好地满足用户对复杂查询的需求。研究过程中，系统将结合AI大语言模型的语义处理和上下文分析能力，与传统检索算法相结合，为构建智能化、用户友好的对话式搜索系统提供技术支持。   1. **阅读文献资料进行调研的综述；**   随着互联网信息量的急剧增加用户对从海量信息中快速获取所需内容的需求日益迫切。然而，传统基于关键词匹配的搜索引擎在准确性、相关性以及用户交互方面的局限性日益显现[1][2][3][4]。尽管传统搜索引擎能够快速返回大量信息，但在复杂查询或多轮对话场景中，往往难以精准理解用户意图，导致用户需反复修改查询以获得满意答案。这种信息过载不仅降低了搜索效率，也极大影响了用户体验。尤其在面对模糊查询、语义理解不清或特定领域的专业问题时，传统搜索引擎通常只能返回不够精准或无关的结果，使得用户在解决复杂问题时感到困扰[1][2]。  近年来，基于深度学习的AI大语言模型，如GPT系列（Generative Pre-trained Transformer）和BERT（Bidirectional Encoder Representations from Transformers）等，取得了显著进展，成为自然语言处理（NLP）领域的主流技术[5][6][13]。这些模型在机器翻译、文本生成、情感分析、对话系统等多个领域展现出强大能力。在对话式搜索引擎中，AI大语言模型通过卓越的语言理解和生成能力，为用户提供更加智能、个性化的服务。以文森等人（2024）为例，基于大语言模型的问答技术研究进展综述指出，GPT- 4模型具有出色的上下文理解和文本生成能力[5]。GPT-3在多轮对话中的表现尤为突出，能够理解复杂的查询意图，并生成连贯、符合语境的回答。其多任务学习和大规模预训练赋予了强大的泛化能力，能够适应不同领域的对话和搜索需求。此外，赵鑫等人（2023）探讨了大语言模型时代下的信息检索研究发展趋势，指出这些模型在语义理解和上下文处理方面的优势，使搜索引擎能够更精准地理解用户需求[6][13]。  国内亦涌现出一批具有竞争力的大语言模型，如华为的盘古大模型、阿里巴巴的M6、百度的文心大模型及腾讯的混元大模型等。这些国产模型在中文语境下的自然语言理解和生成能力日益成熟，尤其在多轮对话和深度语义理解方面取得显著进展，逐渐成为搜索引擎和智能对话系统等领域的重要技术基础，展现出巨大的应用潜力[5][6]。  在对话式搜索引擎的设计与实现方面，姜琨等人（2020）对倒排索引压缩算法进行了研究综述，指出了倒排索引在搜索引擎中的重要性以及优化其压缩算法以提升搜索效率的方法[1][7]。胡涛（2021）基于ElasticSearch设计并实现了智能搜索引擎，展示了ElasticSearch在分布式搜索引擎中的应用与优势[2][3][8]。董元和等人（2023）进一步研究了基于ElasticSearch分布式搜索引擎的信息检索方法，提出了优化信息检索算法以提升搜索准确性和效率的策略[3][8]。罗晗骞和石月鑫（2020）探讨了人工智能（AI）技术在搜索引擎中的应用，指出AI技术在提升搜索引擎智能化水平中的重要作用[4]。范朗（2020）研究了ElasticSearch在海量数据存储与查询优化方面的应用，进一步证明了ElasticSearch在大规模信息检索中的有效性[9]。柳帆（2021）基于ElasticSearch的科技资源检索系统的研究与实现，展示了其在科技资源管理与检索中的实际应用案例[8]。  在对话式搜索引擎中，结合AI大语言模型与传统信息检索算法成为提升搜索引擎性能的关键。Kishore等人（2022）提出使用遗传算法增强元搜索引擎连贯性的研究，展示了智能优化算法在搜索引擎中的应用[11]。Poply等人（2024）通过AI驱动的音频转录技术，推进了搜索引擎功能的发展，尤其在处理多媒体内容检索方面的创新[11]。Zhu等人（2023）在其关于大语言模型在信息检索中的综述中，详细探讨了大语言模型如何通过语义理解提升信息检索效果[12]。Mao等人（2023）提出基于提示框架的对话式搜索，展示了大语言模型在捕捉上下文搜索意图方面的应用潜力[13]。  综述相关文献可见，AI大语言模型在对话式搜索引擎中的应用逐渐成为研究热点。大语言模型，如GPT-3和BERT，凭借出色的语言理解与生成能力，为对话式搜索引擎提供强大技术支持，尤其在多轮对话和长尾问题处理上展现巨大潜力。然而，现有研究也暴露了对话式搜索引擎面临的一些技术难点。首先，高效处理多轮对话中的上下文信息，确保模型能够动态调整搜索结果，是当前技术研究的重要挑战。现有大语言模型虽能理解上下文，但在多轮对话中，随着对话内容增加，模型可能出现对上下文理解不准确或结果不一致的情况。因此，保证对话连贯性和一致性仍需解决[3]。其次，将搜索引擎传统信息检索算法与大语言模型结合，以实现更高效和更精确的搜索结果，是另一个值得关注的研究方向。传统搜索引擎多依赖关键词匹配和排名算法，而大语言模型通过语义理解提供更智能的检索能力。如何结合这两种技术，在保证响应速度的同时提升查询精度，仍是亟待解决的问题[4]。最后，尽管大语言模型在语义理解和生成方面表现出色，但其在实际应用中的效率和稳定性仍是挑战。例如，模型的计算资源需求较高，可能影响实时搜索引擎的响应速度和用户体验。因此，未来研究可探索更高效的模型压缩和优化算法，以降低计算开销，提升系统可用性[7]。  结合课题理解，构建一种能够通过自然语言进行对话并理解上下文的智能搜索引擎，成为信息检索领域亟待解决的关键问题。尤其在多轮对话中，如何准确理解并保持上下文信息，以提供与用户需求更契合的答案，是传统搜索引擎面临的挑战。传统搜索引擎在处理此类情境时，往往难以准确捕捉用户的真实意图，从而影响搜索结果的相关性和准确性，限制了其在一些复杂应用场景中的表现。  本课题的研究旨在探索人工智能与信息检索技术的结合，以提供更为高效和智能的搜索体验。通过开发对话式搜索引擎，期望能够提高信息检索的准确性与效率，满足用户日益增长的个性化需求。随着AI大语言模型技术的进一步发展，基于对话式搜索引擎的智能化水平可能会逐步提升，更好地适应用户需求，进一步优化完善系统。   1. 姜琨, 朱磊, 宋省身, 等. 倒排索引压缩算法研究综述[J]. 小型微型计算机系统, 2020, 41(4):715-723. 2. 胡涛.基于ElasticSearch的智能搜索引擎的设计与实现[D].西安电子科技大学,2021.DOI:10.27389/d.cnki.gxadu.2021.000419. 3. 董元和, 贾炎, 朱勇, 等. 基于ElasticSearch分布式搜索引擎的信息检索方法研究[J]. 湖北师范大学学报(自然科学版), 2023, 43(04):56-61. 4. 罗晗骞, 石月鑫. 人工智能（AI）技术在搜索引擎中的应用[J]. 电子技术与软件工程, 2020, (21):47-48. 5. 文森, 钱力, 胡懋地, 等. 基于大语言模型的问答技术研究进展综述[J]. 数据分析与知识发现, 2024, 8(06):16-29. 6. 赵鑫, 窦志成, 文继荣. 大语言模型时代下的信息检索研究发展趋势[J]. 中国科学院, 2023, 786. 7. 姜琨, 朱磊, 宋省身, 等. 倒排索引压缩算法研究综述[J]. 小型微型计算机系统, 2020, 41(04):715-723. 8. 柳帆. 基于ElasticSearch的科技资源检索系统的研究与实现[J]. 现代计算机, 2021, 27(26):93-100. 9. 范朗. Elasticsearch海量数据存储查询优化[J]. 工业控制计算机, 2020, 33(10):85-87. 10. Kishore, B. R., Amogh, V. R., Vrushab, S., Kumar, C. A., & Arya, A. (2022). Enhancing the Coherence of a Meta Search Engine using Genetic Algorithm. 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 1562-1568. 11. Poply, S., Sharma, A. K., & Rathore, P. S. (2024). Advancing Search Engine Functionality Through AI-Powered Audio Transcription. 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET), Indore, India, 1-4. doi: 10.1109/ACROSET62108.2024.10743633. 12. Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., & Wen, J. R. (2023). Large language models for information retrieval: A survey. arXiv preprint, arXiv:2308.07107. 13. Mao, K., Mo, Z., Ma, F., Hou, J., Cao, H., & Qian, H. (2023). Large language models know your contextual search intent: A prompting framework for conversational search. Conference on Empirical Methods in Natural Language Processing, 1211-1225.   **3. 根据任务书的任务及文献调研结果，初步拟定的执行（实施）方案（含具体进度计划）。**  初期调研表明，随着互联网信息激增，用户亟需快速获取所需内容，而传统搜索引擎因依赖关键词匹配在处理复杂查询和多轮对话时存在不足，难以准确理解用户意图，影响体验。因此，开发能通过自然语言对话并理解上下文的智能搜索引擎成为信息检索的重要课题。  数据驱动的方法通过学习大量历史数据有效满足用户需求，而AI大语言模型（如GPT系列）在自然语言处理中的出色表现，为提升对话系统智能化提供了技术支持。同时，国产大语言模型如华为盘古、阿里M6、百度文心、腾讯混元等，在中文理解与生成方面取得显著进展，为本课题提供了多样化选择。  本课题的研究主要涉及以下内容：   1. 模型集成与优化：集成主流AI大语言模型（如GPT-4、星火、通义千问）与ElasticSearch，优化语义搜索与倒排索引管理，提升检索效率和结果相关性。 2. 系统开发与多模态交互：采用模块化设计，开发基于Qt的跨平台客户端，实现文本与图像的多模态交互功能，提升用户搜索体验。 3. 数据管理与安全：构建高效的数据存储和检索机制，使用MySQL和Redis进行数据库管理与缓存优化，确保数据隐私与系统安全。   本课题研究的重难点和创新点如下：   1. 重难点：实现AI大语言模型与传统搜索引擎的高效集成是本课题的主要技术挑战。需要确保自然语言理解与搜索结果的精准匹配，同时优化系统的实时响应能力和多模态处理性能。此外，多语言支持尤其是在中文语境下的优化，以及保障数据隐私与系统安全性，都是亟需解决的关键问题。 2. 创新点：创新地融合了AI大语言模型的语义理解能力与传统检索算法，提升了搜索引擎在复杂查询和多轮对话中的表现。通过支持多模态交互（文本与图像），丰富了用户的搜索方式。同时，采用跨平台和容器化的高效架构设计，确保系统的高可用性和易维护性。此外，实施本地化部署以保护用户数据隐私，进一步增强了系统的安全性和用户信任度。   综述所述，绘制如下项目框架图：    图1 AI智能搜索引擎系统架构图  具体进度计划：   1. 毕设开始后第一、二周，完成开题报告，按照开题报告格式撰写开题材料。 2. 毕设开始后第三到七周，收集整理论文相关资料，深入了解课题核心知识要点，熟悉开发环境，完成中期检查。 3. 毕设开始后第八周，解析课题相关模块功能需求探索的问题，解决细节问题。 4. 毕设开始后第九到十周，完成外文翻译, 中英文一对一翻译，排版整齐，语句通顺。 5. 毕设开始后第十一周，撰写毕业设计(论文)报告，交指导教师批阅。 6. 毕设开始后第十二周，根据指导老师和评阅老师意见修改论文。 7. 毕设开始后第十三周，再次修改论文，完成答辩前准备，实现ppt等相关介绍资料的确定。 8. 毕设开始后第十四周，毕业设计(论文)答辩。 | | | | | | |
| 指导教师批阅意见 | 李峰同学认真调研了课题 "基于AI大语言模型的对话式搜索引擎”，详细分析了当前人工智能在信息检索领域的应用。该课题难度适中，具备较高的可实现性，且与学生的专业发展方向高度契合。研究计划明确，时间安排合理，能够有效确保课题的顺利推进。综合考虑各方面因素，同意该课题开题，  指导教师(签名)： 年 月 日 | | | | | |