工程与管理科学
Engineering and Management Science
第 6 卷◆第 07 期◆版本 1.0◆2024 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

网络信息检索中的人工智能信息处理技术研究

赵霞
上海理工大学图书馆
DOI：10.12238/ems.v6i7.8188

[摘　要] 本文详细讨论了人工智能（AI）技术在文献检索中的应用，并分析了它如何通过提高检索效率和精确度来改变传统的文献检索方法。文中首先介绍了 AI 在文献分类和自动摘要生成等方面的关键作用。随后讨论了这些技术实施中遇到的主要挑战，比如数据质量、用户隐私保护和技术透明度等问题，并提出了相应的解决策略。
[关键词] 人工智能；文献检索；自动化；智能化功能；神经网络

Research on Artificial Intelligence Information Processing Technology in Network Information Retrieval

Zhao Xia
Shanghai University of Technology Library
[Abstract] This article discusses in detail the application of artificial intelligence (AI) technology in literature retrieval, and analyzes how it can change traditional literature retrieval methods by improving retrieval efficiency and accuracy. The article first introduces the key role of AI in literature classification and automatic abstract generation. Subsequently, the main challenges encountered in the implementation of these technologies were discussed, such as data quality, user privacy protection, and technical transparency, and corresponding solutions were proposed.
[Keywords] artificial intelligence; bibliography retrieval; Automation; Intelligent functionality; neural network

1 引言

随着科研活动的增多和数字资源的丰富，研究人员在获取和管理文献资料方面面临着越来越大的挑战。传统的文献检索方法往往依赖于人工处理和关键词搜索，这种方法存在效率低下和信息过载的问题，使得研究人员很难快速准确地获取所需信息，随着人工智能技术的不断发展，特别是机器学习和自然语言处理等领域的进步，文献检索也迎来了新的机遇。

人工智能的引入为文献检索带来了革命性的改进，通过机器学习算法的应用，文献检索系统能够自动学习和优化搜索结果，提高了检索效率和准确度。例如，基于用户行为和反馈的个性化推荐系统可以根据用户的偏好和历史搜索记录，为其提供相关的文献资源。此外，自然语言技术的发展使得系统能够理解和处理自然语言查询，进一步提升了检索体验。另一方面，神经网络在文献检索中扮演着重要角色，它们能够学习文本之间的语义相似性，通过计算文献之间的相似度，为用户提供与其查询相关的文献资源，这种语义匹配能够更好地理解用户的检索意图，从而提供更精确的检索结果。

然而，人工智能在文献检索中的应用也面临着一些挑战。首先是数据质量问题。尽管现代文献资源日益丰富，但其中可能存在大量的错误或不准确的信息。这些错误和不准确性会直接影响到检索结果的准确性和可信度，从而降低了用户对检索系统的信任度和满意度。另一个挑战来自于用户隐私保护问题。在个性化推荐和智能化检索过程中，系统需要收集和分析用户的搜索历史、点击行为等数据来提供更精准的检索结果。然而，这些个人数据的使用可能涉及到用户隐私的泄露和滥用。因此，如何在提高检索效率的同时确保用户数据的安全性和隐私性成为了一个需要解决的技术和伦理问题。

2 人工智能在信息检索中的作用

2.1 文献分类

在文献检索中，文献分类是一个基本且重要的功能，尤其在面对海量的科研资料时。采用人工智能技术，尤其是深

度学习模型，系统能够自动学习并识别文献中的关键信息，如研究主题、方法论以及研究结果等。这些模型通过分析文献的摘要、引言、结论等部分，提取出文献的核心概念，并据此对文献进行有效分类。

2.2 自动摘要生成

在当今快速发展的学术环境中，自动摘要生成技术已成为人工智能在文献检索领域的一项关键应用。利用自然语言处理（NLP）和机器学习技术，特别是深度学习方法如序列到序列模型（Seq2Seq）和注意力机制，AI 能够从繁杂的学术文献中提取核心信息，并自动生成精炼的文献摘要。这些技术通过学习大量的文献数据，训练模型来理解文本的结构和语义，进而高效地识别出文献中的关键观点和论据。

2.3 引用神经网络

文献检索应用的简单知识表示由神经网络提供，其中节点表示 IR 的对象（例如关键字，引文，作者），链接或突触表示它们的加权关联。反向传播网络的学习特性和 Hopfield 网络的搜索特性为我们识别数据库中的相关信息提供了准确的方法。Hopfield 网络作为神经网络引入，使用一种特殊类型的存储器，可以直接寻址内容。信息和知识存储在互连的神经元（节点）和连接在单层中的加权突触（链路）中。这些信息被检索和遍历，直到网络达到稳定状态。ID3 和 IDSR 算法主要用于符号学习中实现信息检索，ID3 算法用于构造决策树，由 Quinlan 提出。它根据分而治之的原则对对象进行分类。熵根据以下函数计算：

$$entrophy = -p_{pos}\log p_{pos} - p_{neg} \quad (1)$$

其中，p_{pos} 和 p_{neg} 分别表示负文档和正文档的比率。

为了根据遗传和进化的原理制定系统并解决问题，我们使用遗传算法。第（t + 1）次迭代中的新种群通过选择更好和更适合的个体来形成。在这里，一些成员转换采用遗传算子，制定一个新的解决方案。通过应用 GA 的方法，我们的目标是通过一组最佳文档来搜索研究人员需求的最佳匹配（图 1）。

另一方面，信息检索（IR）需要应用 Logistic 回归，因为它对解决各种现实问题非常有帮助。通过该方法计算特定集合和特定查询之间的关系，并按照集合的概率值递增或递减的顺序对集合进行排序。Logistic 回归估计如下

$$\log O(R|Q,C) \approx c_0 \sum_{i=1}^{s} c_i X_i \quad (2)$$

其中 c_i 是相关的系数集，s 是一组统计数据，X_i 为集合文档的查询和数据库（表 1）。

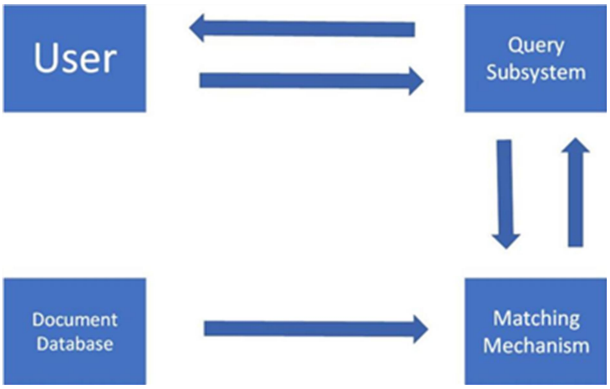

图 1　遗传算法过程

表 1　信息检索模型的简要回顾

序号	方法	特征
1	神经网络	信息和知识存储在相互连接的神经元（节点）和加权突触（链接）连接在一个单一的层
2	霍普菲尔德网络	使用一种特殊类型的存储器可以直接寻址内容
3	符号学习	ID3 和 IDSR 算法在符号学习中用于实现和构建决策树算法
4	遗传算法	为了建立一个系统和解决问题的遗传和进化的原则，我们使用了遗传算法
5	贝叶斯定理	该模型通过使用数据的后验概率分布，提供了一个项目的所有可能类别
6	逻辑回归	用于解决各种现实生活中的问题，其中集合的排名是通过增加或减少其概率值来完成

3 所面临的挑战

3.1 数据质量控制

在文献检索系统中引入人工智能带来了诸多便利，然而，数据质量控制仍是一个突出的挑战。自动处理和分析大量文献数据时，确保所处理信息的准确性和可靠性至关重要。要解决这一问题，首先需要从数据采集的源头把关，确保引入系统的原始数据就具有较高的质量。此外，可以利用先进的数据清洗技术和异常检测算法去除或纠正数据集中的错误和异常值。例如，采用自然语言处理工具可以帮助识别和校正文献中的错别字、语法错误和格式问题，而统计方法和机器学习模型则可以用于识别数据中的异常点和不一致性。进一步地，构建强大的数据验证和更新机制也是确保数据质量的关键。这包括定期对数据库进行维护和更新，以反映最新的研究成果和学术讨论。

3.2 用户隐私保护

在当前的文献检索系统中，随着个性化推荐和数据分析技术的广泛应用，用户隐私保护尤显重要。这些系统往往需要收集和分析用户的搜索历史、阅读习惯以及偏好设置等私人信息，以提供更加定制化的服务。然而，这种数据的收集和使用过程中极易触及用户隐私的边界，因此，开发和实施严格的数据处理和存储协议成为了必要措施，以确保用户信息的安全不被侵犯。首先，所有收集的用户数据都应该经过严格的匿名处理和加密，保证即便数据被非法访问，也难以追溯到具体个人。总之，文献检索系统在设计和实施过程中，必须将用户隐私保护作为核心考虑，采取一系列技术和管理措施来确保个人信息的安全，从而增强用户对系统的信任度，

推动技术的健康发展和广泛应用。

3.3 技术透明度

在人工智能技术，特别是用于文献检索的 AI 系统中，技术透明度和可解释性是极其重要的。用户往往对 AI 的决策过程感到困惑，特别是当系统提供的结果直接影响到他们的研究和学习选择时。因此，提高 AI 系统的可解释性，让用户能够理解并信任系统所提供的检索结果及其背后的逻辑，成为了提升用户体验和系统透明度的关键。首先，可以通过开发和集成可解释的 AI 模型来增强系统的透明度。例如，如果一个文献被推荐，系统应该能够解释为什么这篇文献是相关的，包括其被引用的次数、发表的期刊的影响因子、以及与用户之前搜索或阅读的文献的关联度等信息。总之，通过上述措施提高 AI 系统的透明度和可解释性，不仅能够增强用户的信任和满意度，还能促进用户与系统的有效互动，提升系统的整体性能和用户体验。这在科研领域尤为重要，因为科研工作者依赖于准确且可靠的信息来支持其研究决策。

4 应对措施

4.1 增强数据质量管理

数据质量管理是确保文献检索系统有效性和可靠性的关键组成部分。以下详细阐述了如何通过严格的数据清洗、先进的数据验证技术，以及定期更新数据源来增强数据质量管理：数据清洗和预处理是提升文献数据质量的第一步。这一过程包括识别并去除重复记录、修正明显的错误，如错别字和格式错误，以及标准化数据格式，使之适用于分析和检索。例如，在处理学术文献时，必须确保各文献的引用格式一致，如统一作者名字的书写方式或期刊名称的缩写。使用自然语言处理工具可以自动识别和纠正文献中的语法和语义错误，进一步提升数据的质量。为确保输入数据的准确性和完整性，可以采用各种数据验证技术。这包括实时数据验证，如在数据录入时即进行格式和逻辑检查。例如，一个有效的数据验证规则可能包括检查文献条目中的发表年份是否逻辑合理（如不可能是未来的年份）。通过与主要的学术数据库和期刊出版商建立数据交换协议，可以系统地收集最新的研究成果，并定期将其集成到检索系统中，确保用户始终能够获取到最新和最全面的研究信息。通过这些策略，文献检索系统能够提供更高质量的数据，支持科研人员进行更精准和深入的学术研究。这不仅增强了研究的效率，也提升了研究成果的质量和可信度。

4.2 定期更新数据源

定期更新数据源是确保文献检索系统有效性的核心部分，尤其是在科学和学术研究迅速发展的今天。详细来说，以下是一些关键措施和方法，用于保持数据源的现代性和相关性：使用自动化工具如爬虫技术，定期从预定的学术数据库和期刊网站抓取最新发布的文献。这些工具可以设置为识别并下载新文献，并自动提取关键元数据（如标题、作者、摘要、关键词等），这种方法可以确保数据的即时更新和高准确性。监测和更新现有文献的引用情况。引用次数是评估文献影响力的重要指标，定期更新引用数据可以帮助研究人员了解特定研究的当前影响和相关性。跟踪重要文献的后续研究或相关评论，并将这些信息添加到原有文献条目中。这有助于用户获得关于某一领域研究动态的全面视图。通过实施上述策略，文献检索系统可以持续提供最新、最全面的学术资源。

5 结语

在本文中，我们深入探讨了人工智能信息处理技术在网络信息检索中的广泛应用，并详细分析了其带来的转变和挑战。人工智能技术，特别是深度学习、自然语言处理和机器学习等领域的不断进步，已经极大推动了信息检索的效率和精度，为用户提供了更智能化、个性化的搜索体验。在数据隐私方面，如何在提高检索效率的同时保护用户的个人隐私成为了一个关键问题，同时，模型可解释性的缺乏可能影响用户对系统的信任度和可接受程度。信息质量和信息过载问题也最要通过算法优化和用户教育来解决，以提高检索结果的质量和用户满意度。通过持续的技术创新和改进，我们可以不断优化信息检索系统，解决这些挑战，并进一步提高系统的性能和用户体验，未来，随着人工智能技术的进一步发展，我们可以预见一个更加智能化、个性化和多样化的信息检索领域的到来，它将为用户提供更高效、更准确的检索服务，极大地丰富知识的获取和交流方式。

[参考文献]

[1]冯燕青.大数据时代人工智能在网络信息检索中的应用[J]. 科技创新与应用，2023，13（3）：165-168.

[2]布艳艳. 基于人工智能技术的图书馆信息检索模型[J]. 电子设计工程，2021，29（14）：24-28.

[3]梁丰.大数据时代人工智能在网络信息检索中的应用[J]. 科技创新导报，2020，17（18）：112-113.

[4]殷楠楠.大数据时代人工智能在网络信息检索中的应用分析[J]. 现代信息科技，2019，3（17）：15-16，19.

[5]张少宇. 人工智能在计算机网络技术中的应用探讨[J]. 电脑知识与技术，2019，15（13）：223-224.

[6]聂为之，王岩，杨嵩，等. 基于循环生成对抗网络的跨媒体信息检索算法[J]. 计算机学报，2022（045-007）：1529-1538.

[7]石湘，刘萍. 基于知识元语义描述模型的领域知识抽取与表示研究——以信息检索领域为例[J]. 数据分析与知识发现，2021，5（04）：123-133.

[8]陈乐，刘迎春. 基于用户需求挖掘的交互式信息检索算法设计[J]. 计算机仿真，2022（039-005）：418-422.

基金项目：本文系国家自然科学基金项目（面上）“联合约束下面向V2G技术的智能电网的安全控制研究”（2022/01-2025-12）（项目编号：62173231）的研究成果之一。