基于 AI 的图书馆数据分析与服务优化：自然语言转 SQL 的创新实践

王璇[1]，刘若歆[2]，侯新宇[1]

1. 中国传媒大学图书馆，北京 100024

2. 中国传媒大学媒体融合与传播国家重点实验室，北京 100024

摘　　要： 本文介绍了基于人工智能（AI）的图书馆数据分析方法，该方法能够将图书馆工作人员的自然语言查询转化为 SQL 查询语句，实现对现有图书馆数据的便捷分析。通过该系统，工作人员无需专业的数据库知识即可快速获取所需信息，从而优化图书馆数据服务与管理。实验结果表明，该方法能够显著提升数据分析效率，为智慧图书馆建设提供了一种实用的解决方案，是中国传媒大学图书馆在积极推进人工智能与各学科融合的背景下的一次有益尝试，旨在提升图书馆服务的智能化水平。

关 键 词： 图书馆数据分析；自然语言处理；SQL 生成；Deepseek；AI 应用

Library Data Analysis and Service Optimization Based on AI: An Innovative Practice of Text to SQL Conversion

Wang Xuan[1], Liu Ruoxin[2], Hou Xinyu[1]

1.Library of Communication University of China, Beijing 100024

2.State Key Laboratory of Media Integration and Communication of Communication University of China, Beijing 100024

Abstract: This paper presents an Artificial Intelligence (AI)–based library data analysis method that converts natural language queries from into SQL query statements, enabling convenient analysis of existing library data. Through this system, staff members can quickly access required information without specialized database knowledge, thereby optimizing library data services and management. The experimental results show that this method can significantly improve the efficiency of data analysis, providing a practical solution for the construction of smart libraries. It is a beneficial attempt by the Communication University of China Library in the context of actively promoting the integration of artificial intelligence with various disciplines, aiming to enhance the intelligence level of library services.

Keywords: library data analysis; natural language processing; Text to SQL; Deepseek; AI application

引言

近年来，国家高度重视人工智能发展，将其上升为国家战略，并出台了一系列政策文件，为 AI 在图书馆领域的应用提供了强有力的政策支撑和方向指引：

《新一代人工智能发展规划》明确提出要“发展智能图书馆，构建智慧化知识服务体系”。《“十四五”公共文化服务体系建设规划》中强调要“推动公共文化机构数字化、网络化、智能化发展”。《关于推进实施国家文化数字化战略的意见》中则提出要“建设国家文化大数据体系，促进文化资源数字化、文化生产智能化、文化传播网络化”，在国家政策的积极引导下，各地图书馆纷纷探索 AI 应用场景，高校图书馆也成为其中重要的一份子[1]。

随着高校图书馆信息化程度的不断提高，图书馆积累了海量的读者行为数据。然而，如何从这些数据中挖掘有价值的信息，并将其应用于图书馆服务优化，一直是图书馆管理人员面临的挑战。传统的解决方案依赖于专业的数据分析人员，这不仅成本高昂，而且效率低下。如何让非技术背景的图书馆工作人员能够方便快捷地利用这些数据，提升服务质量，成为智慧图书馆建设中的一个关键问题[2]。

过去，图书馆工作人员在进行数据分析时，需要打开数据库软件，手动编写复杂的 SQL 查询语句，或者依赖数据分析人员的协助。而对于学校管理人员，往往只能通过科研部门或相关部门定期发布的数据报告，来了解和掌握既定的一些数据，一旦需要查询的信息不

项目 / 基金信息：本文系本文系中国传媒大学校级科研业务费课题《智能图书馆：基于 AI 的数据分析与服务优化（项目编号：CUC24WH08）》阶段性研究成果之一（University-level Research Project ‘Intelligent Library: AI-Based Data Analysis and Service Optimization’ (Project Number: CUC24WH08).）。

作者简介：王璇，中国传媒大学图书馆技术部 工程师。

在数据报告之中，仍然需要求助专门的数据分析人员，这种方式效率低下，且难以应对快速变化的需求。

本文采用的自然语言转 SQL 方法（即 Text-to-SQL），旨在借助 AI 技术，让图书馆工作人员能够使用日常语言进行数据分析，极大地降低了技术门槛，提高了数据分析的效率和实时性，从而进一步提升图书馆服务的智能化水平。

一、研究背景和研究意义

（一）研究背景：人工智能推动技术成熟

Text-to-SQL 的概念最早可以追溯到 1970 年代，当时的研究人员开始探索如何让计算机理解自然语言并将其转换为结构化查询语言。然而，由于当时的技术限制，Text-to-SQL 的研究进展缓慢，近年来随着人工智能大爆发，才取得长足进步。它的具体发展历程如下 [3]：

早期探索（1970s-1990s）：这一阶段的研究主要集中在基于规则和语法的转换方法。由于自然语言的复杂性和多样性，这些方法难以处理复杂的查询和表达。

机器学习方法兴起（2000s-2010s）：随着机器学习技术的发展，研究人员开始尝试使用统计机器学习模型进行 Text-to-SQL 转换。这些方法在一定程度上提高了转换的准确性，但仍然存在一些问题，如需要大量的标注数据、泛化能力弱等。

深度学习爆发（2010s 至今）：深度学习技术的兴起为 Text-to-SQL 带来了革命性的突破。基于深度学习的模型，如 Seq2Seq、Transformer 等，在 Text-to-SQL 任务上取得了显著的进展。特别是近年来，大型语言模型（LLM）的出现，如 GPT 系列、Deepseek 等，更是将 Text-to-SQL 的性能推向了新的高度。

Deepseek 等大型语言模型的出现，让人工智能具有了强大的语义理解和生成能力，这也让 Text-to-SQL 这项技术发展到了可应用的阶段，利用这些通用大模型，就能够更好地理解自然语言的意图，并生成更准确的 SQL 查询语句；同时越来越多的 Text-to-SQL 数据集被构建出来，为模型的训练提供了充足的语料。这些数据集涵盖了各种领域和场景，使得模型能够更好地适应不同的数据和查询需求；最后，深度学习技术日新月异的发展，为 Text-to-SQL 提供了更有效的模型和算法。这些模型能够更好地捕捉自然语言和 SQL 之间的复杂关系，从而提高转换的准确性。[4]

（二）研究意义

国内外的研究者一直在积极探索如何利用 AI 技术提升图书馆的服务质量和效率。例如，有研究利用自然语言处理（NLP）技术优化图书馆咨询服务 [5]，通过机器学习算法提升用户查询响应的速度和准确性。也有研究关注用户行为数据的分析，运用机器学习模型进行个性化的图书推荐服务。美国一些公共图书馆已经整合语音识别和聊天机器人技术，为用户提供 24/7 的咨询服务。还有一些研究者则聚焦于数据驱动的服务优化，研究如何通过分析用户数据来改进图书馆的空间规划和资源配置 [6]。

然而，目前国内图书馆在利用 AI 进行数据分析方面的实践还相对有限，特别是在使用 Text-to-SQL 技术处理日常数据查询和分析任务方面鲜有研究。传统的数据分析方法往往需要工作人员掌握复杂的 SQL 语法和数据库知识，这不仅增加了培训成本，也容易出现查询错误。引入 Text-to-SQL 技术能够让图书馆员通过自然语言描述直接获取所需数据，既可以提高工作效率，又能确保数据分析的准确性，最后还能拓展业务边界。

本研究不仅在技术上具有创新性，而且在应用上具有重要的实践意义。通过降低数据分析的技术门槛，本研究能够帮助图书馆工作人员更好地利用数据，从而优化服务、提升用户满意度，为智慧图书馆建设提供有力支撑。同时，该研究也将为其他公共服务机构运用 AI 技术进行数据分析提供有益的参考和借鉴。

二、系统设计目标与实现

本研究旨在设计一个基于大型语言模型（LLM）的实时查询系统，通过 Text-to-SQL 功能实现对图书馆数据的自然语言实时查询，并优化图书馆数据报告的生成流程 [7]。当前，图书馆每月需要人工对借阅数据、空间利用数据、网络资源利用数据及出入馆数据进行大规模分析与整理，进而生成格式化的月度和年度报告。然而，这一过程面临诸多挑战，包括数据庞杂、校对繁琐、效率低下以及输出形式有限等问题。为解决上述问题，本系统设计了一个文本对话窗口，通过自然语言处理技术，实现用户以自然语言形式实时查询馆内各项数据资料，从而提升数据检索与分析效率，简化报告生成流程。

（一）系统架构

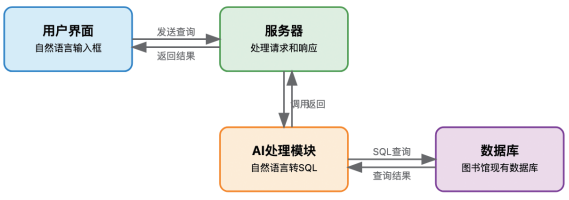
本系统采用常见的 Web 应用架构（B/S 架构），用户可以通过浏览器访问系统，进行数据查询。系统的核心组成部分包括（图 1）：

1. 用户界面：提供自然语言输入框，用户可以在这里输入查询需求。

2. AI 处理模块：利用开源 Deepseek 的 API 负责将用户的自然语言转化为 SQL 查询语句。

3. 数据库：存储图书馆的各种数据，如学生信息、借阅记录、入馆记录等。 需要说明的是，本系统是在现有图书馆数据库的基础上进行开发的，没有对数据库的结构进行修改。

4. 服务器：负责处理用户请求，调用 AI 模块，查询数据库，并将结果返回给用户。



> 图 1 系统架构

> Fig 1 System Architecture

（二）项目逻辑结构

系统的代码按照功能模块进行组织，如图2所示：

```
CopyInsert
wxproject/
├── app.py                  # 主应用程序
├── templates/              # 前端模板
│   └── index.html          # 主页面
├── check_db.py             # 数据库检查工具
├── test_query.py           # SQL查询测试工具
├── test_api.py             # API测试工具
├── requirements.txt        # 依赖配置
└── .env                    # 环境变量配置
```

> 图2 逻辑结构

> Fig 2 Logical Structure

(1)`app.py`：主程序，负责启动服务器，处理用户请求。

(2)`templates/`：存放网页模板文件，例如用户界面的HTML代码。

(3)`check_db.py`：数据库检查工具，用于确保数据库连接正常。

(4)`test_query.py`：SQL查询测试工具，用于验证SQL语句的正确性。

(5)`test_api.py`：AI接口测试工具，用于测试AI模块的可用性。

（三）关键技术实现：

1. 提示工程优化

为了提高AI模块的准确性，本项目中，作者针对图书馆数据的特点，对AI模块进行了“提示工程”优化。简单来说，就是告诉AI模块数据库的结构和数据的含义。

在提示词模块里可以告诉AI：`studentInfo`表包含学生的信息，例如学生姓名、学生ID、性别、学院等。`indoorInfo`表包含学生的入馆记录，例如入馆时间、学生ID等。在提示词中也需要一些模糊匹配概念，比如当输入‘计算机学院’时，自动匹配为数据库中的‘计算机与网络安全学院’，以达到更为精确的查询结果。通过提前载入提示词的这种方式，可以有效地提高AI模块对用户查询意图的理解能力，从而生成更准确的SQL查询语句。

2.上下文管理

为了提高转换准确性，系统实现了智能的上下文管理，主要包括：保持数据库结构信息，在多次对话中，系统会一直保留数据库的结构信息，避免重复传递，减少计算量。记录字段映射关系，系统会记录自然语言中的关键词与数据库字段的映射关系，例如将“学生姓名”映射到`studentInfo.studentname`字段，提高后续查询的准确性。维护查询历史记录，系统会维护用户的查询历史记录，以便用户可以参考之前的查询结果，进行更复杂的分析。

三、实践成果和效果分析

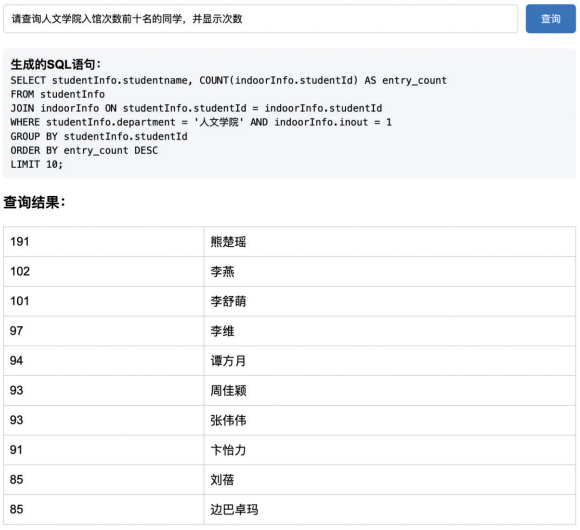
（一）实践成果

笔者基于HTML和JavaScript实现了Web界面的查询系统，同时为了验证系统的有效性，本系统使用中国传媒大学图书馆某一个月份入馆的实际数据，进行混淆后进行系统测试。人工智能模块接入Deepseek的API进行Text-to-SQL的技术处理。最终实际呈现效果如图3所示：

> 图3 查询界面

> Fig 3 Query Interface

系统页面简洁易懂，并给出示例查询，用户可以根据示例的提示进行查询，也可以根据实际需求进行测试查询，例如，在查询框输入“请查询人文学院入馆次数前十名的同学，并显示次数”，点击查询后，得到的内容有两个部分：生成的SQL语句和查询结果。



> 图4 效果展示

> Fig 3 Performance Demonstration

（二）实际效果评估

在系统搭建完成后，测试阶段本系统主要关注2个指标：

1.转换准确率：AI模块成功将自然语言转换为正确的SQL语句的比例。

2.查询执行时间：系统完成一次查询所需要的时间。

实验结果表明，系统在各种查询场景下都取得了较高的转换准确率，平均准确率达到95%以上。而系统的平均查询执行时间小于2秒，可以满足图书馆的日常数据分析需求。

（三）效果对比分析

在传统模式下，图书馆的数据分析工作高度依赖数据分析人员。以中国传媒大学图书馆为例，每月初需抽调两个部门共三名工作人员，耗时数天完成上月数据报告的撰写，包括数据查询、整理、分析和校对等环节。这种模式不仅耗费大量人力与时间，

且效率较低。相比之下，本研究设计的系统通过与数据工作相结合，能够使任何图书馆工作人员在短时间内独立完成相同任务，显著提升了工作效率，如图5所示：

比较项目	传统模式	智能系统
人力需求	需要3名工作人员	任一工作人员即可
时间消耗	数天	一天内
数据范围	固定格式报告	支持多维度分析
查询灵活性	仅能查看预设数据	支持自定义查询
使用门槛	需要专业人员操作	自然语言交互

> 图5 效果对比

> Fig 5 Performance Comparison

此外，传统数据报告受限于固定格式，仅能提供有限的数据模块和结果，例如关于入馆次数的数据，你只能看到“全校入馆次数前二十”，且后续细分仅涵盖本科生、硕士生和博士生，数据分析的深度和广度均受到较大限制。而本系统基于自然语言处理与 Text-to-SQL技术，能够支持更为复杂和多样化的查询需求。系统可执行交叉查询及细化的个性化查询，比如计算机学院的领导想查看本学院任何的数据，均可以在系统中输入例如：“计算机学院中，研究生入馆次数最多的人是谁？入馆多少次？”，“计算机学院中，在图书馆借书最多的同学是谁？”。这种能力不仅扩展了数据分析的范围，还显著提升了数据报告的精细化程度与实用性，为图书馆管理与决策提供了更全面的数据支持。

四、结论与展望

（一）研究结论和不足

本研究成功地将人工智能应用于图书馆数据分析，实现了自然语言到 SQL的转换，显著提升了数据分析的效率和可用性。该系统降低了数据分析的技术门槛，为图书馆服务优化提供了新的手段，证明了 AI技术在图书馆领域的巨大潜力，为其他图书馆提供了可借鉴的经验。

然而，尽管系统已具备基础功能，其在成熟度与应用层级上仍存在诸多不足，需进一步优化与完善。包括但不限于以下几个方面：

1.数据量级与性能优化

当前系统在测试环境下的表现与实际数据量级存在显著差异。由于图书馆数据规模庞大，系统在处理大规模数据时可能出现性能瓶颈。未来需通过代码优化、算法改进及分布式计算等技术手段，提升系统的处理效率与稳定性。

2.数据安全问题

系统在调用外部 API时可能面临数据泄露或滥用风险。为确保数据安全，未来需加强数据加密、访问控制及隐私保护机制的设计，同时探索本地化部署方案，减少对外部服务的依赖。

3.多源数据统一化查询

图书馆数据库结构复杂，涵盖借阅、空间利用、出入馆记录、网络资源利用等多类数据。如何实现多源数据的统一化查询是当前系统的主要挑战之一。未来需进一步优化 AI提示词工程，提升模型对复杂查询语句的理解与转换能力，同时探索多数据库融合查询的技术路径。

4.用户体验与功能扩展

当前系统的用户界面与交互设计仍有改进空间。未来可结合用户反馈，优化文本对话窗口的设计，提升用户体验。同时，可探索更多功能扩展，如自动化报告生成、数据可视化及智能决策支持等，进一步增强系统的实用价值。

（二）未来展望

综上所述，本研究为图书馆数据管理与分析提供了新的技术路径，拓展了图书馆在人工智能领域应用的场景，未来可预想的场景有如下几个方面：

图书馆的空间利用优化：抓取精细化的空间使用数据，包括空间、灯光、设施使用数据等，基于使用数据的分析来优化空间布局，提高空间利用率；个性化阅读推荐：通过关联借阅记录、学生信息和图书分类数据，分析不同群体的阅读偏好。从而为不同专业的学生提供定制化的推荐书单，识别专业核心阅读和拓展阅读的模式，优化馆藏资源配置；数据驱动的采购决策：基于系统分析结果调整采购策略，将预算向高需求学科倾斜，同时减少低借阅率类别的采购。提升整体馆藏利用率，提高采购预算效益，满足了更多读者的需求[8]。

基于人工智能未来逐步进入这些应用场景，我们将持续的进行技术迭代与实践验证。未来，笔者将继续致力于系统的优化与升级，紧跟国内外大模型的研究步伐，推动其在图书馆及其他领域的广泛应用，为数据驱动的管理与决策提供有力支持。

参考文献

[1] 初景利，任娇菡，王译晗. 从数字图书馆到智慧图书馆［J］. 大学图书馆学报，2022,40(2)：52-58.

[2] 王曦. 人工智能赋能智慧图书馆发展的作用机制［J］. 图书情报知识，2024,41(6)：94-101，165.

[3] 林毅炜. 基于深度学习的中文自然语言生成复杂 SQL语句技术研究［D］. 湖北：华中科技大学，2021.

[4] 赵志超，游进国，何培蕾，等. 数据库中文查询对偶学习式生成 SQL语句研究［J］. 中文信息学报，2023,37(3)：164-172.

[5] 肖舒玥，孙守强，李青青. AI大模型驱动的智慧图书馆服务体系研究［J］. 图书馆理论与实践，2024,(03):54-61.

[6] 汪聪，张春红，高楠，等. 面向人工智能生成内容时代的图书馆服务适应与创新［J］. 情报理论与实践，2024,47(7)：17-24.

[7] 许辉. 基于大数据分析的高校图书馆服务系统设计研究［J］. 信息与电脑（理论版），2023,35(07):158-160.

[8] 储节旺，杜秀秀，李佳轩. 人工智能生成内容对智慧图书馆服务的冲击及应用展望［J］. 情报理论与实践，2023,46(5)：6-13.