Computer Science and Application 计算机科学与应用, 2026, 16(3), 572-582
Published Online March 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.163085

Hans汉斯

基于LDA主题模型的数据素养评价指标体系研究

郑　浩*，邓海生

西京学院计算机学院，陕西　西安

收稿日期：2026年1月27日；录用日期：2026年2月26日；发布日期：2026年3月5日

摘　要

在信息化快速发展的背景下，大学生数据素养已成为衡量其数字社会适应能力的重要指标。为精准评估其水平，本研究基于中国知网(CNKI)近十年1291篇相关文献，运用潜在狄利克雷分配(LDA)模型进行分析，揭示了数据分析、数据安全、数据素养教育等研究热点，并构建涵盖5个一级指标和23个二级指标的大学生数据素养测评体系。该体系为数据素养教育的优化与课程改革提供了科学依据，推动高质量教学发展，并为后续研究提供了新的视角与方向。

关键词

信息化时代，大学生数据素养，评价指标体系，LDA主题模型，教育技术

A Study on the Data Literacy Evaluation Index System Based on the LDA Topic Model

Hao Zheng*, Haisheng Deng

College of Computer Science, Xijing University, Xi'an Shaanxi

Received: January 27, 2026; accepted: February 26, 2026; published: March 5, 2026

Abstract

Against the backdrop of the rapid development of informatization, college students' data literacy has become a crucial indicator for measuring their adaptability to the digital society. To accurately evaluate their data literacy level, this study conducted an analysis based on 1291 relevant literatures published in the past decade retrieved from China National Knowledge Infrastructure (CNKI),

*通讯作者。

文章引用: 郑浩, 邓海生. 基于 LDA 主题模型的数据素养评价指标体系研究[J]. 计算机科学与应用, 2026, 16(3): 572-582. DOI: 10.12677/csa.2026.163085

using the Latent Dirichlet Allocation (LDA) model. It revealed major research hotspots including data analysis, data security, and data literacy education, and constructed a college students' data literacy evaluation index system covering 5 first-level indicators and 23 second-level indicators. This system provides a scientific basis for the optimization of data literacy education and curriculum reform, facilitates the development of high-quality teaching, and offers new perspectives and directions for subsequent research.

Keywords

Information Age, College Students' Data Literacy, Evaluation Index System, LDA Topic Model, Educational Technology

1. 前言

在信息化与数字化转型背景下，数据已成为关键生产要素与学习、科研和治理的重要资源，信息环境的复杂性进一步凸显数据素养的战略价值。国务院相继出台《促进大数据发展行动纲要》《科学数据管理办法》，并在《“十四五”规划和2035年远景目标纲要》中明确数字化转型方向，为数据素养培育提供了政策指引。作为信息素养的延伸，数据素养体现为个体对数据的获取、评估、处理与运用能力，涵盖采集、管理、分析、共享与伦理等维度。鉴于大学生受专业与背景差异影响，数据素养水平存在不均衡现象，本文拟构建契合当代大学生特征的数据素养评估体系，以科学刻画其现状，并为高校课程建设与素质教育提供理论与实践参考。

2. 大学生数据素养测评体系构建

在构建大学生数据素养测评体系过程中，本文采用LDA (Latent Dirichlet Allocation)主题模型对大量相关文献进行挖掘，以提炼关键主题与核心要素。作为无监督的主题概率模型，LDA能从非结构化文本中识别潜在语义结构，已广泛应用于文本分析。下文将阐释 LDA 的基本原理及其在本研究中的具体应用。

2.1. LDA 主题模型

主题概率模型无需预设文本结构或依赖语法规则，适用于大规模稀疏文本，在主题发现与提炼方面具有优势。相关方法包括LDA、STM、PLSI和Unigram等。其中，Blei等人提出的LDA以稳定性与有效性著称，应用最为广泛。本文采用LDA对文献数据进行主题识别与分析，以揭示大学生数据素养研究的主要内容与结构，其基本原理如图1所示。

LDA模型基于三层贝叶斯生成框架，假设语料库由多篇文档组成，每篇文档由若干潜在主题按权重混合生成，而每个主题又由一组紧密关联的特征词构成。通过分析词语共现关系，模型可揭示文档的潜在主题结构。其计算公式如(2-1)：

$$p(w|\alpha,\beta)=\int p(\theta|\alpha)\left\{\prod_{n=1}^{N}\sum_{z_n}p[z_n|\theta]p(w_n|z_n,\beta)\right\}\mathrm{d}\theta \tag{2-1}$$

其中，α 和 β 分别是主题分布和特征词分布的先验分布超参数。当参数大于1时，分布更平滑；小于1

时则更稀疏。模型通过调整参数并结合上下文信息，可自动识别文档潜在主题结构。文档的主题分布如公式(2-2)：

$$p(D|\alpha,\beta)=\int p(\theta|\alpha)\left\{\prod_{n=1}^{N_d} p[z_{dn}|\theta]\,p(w_{dn}|z_{dn},\beta)\right\}\mathrm{d}\theta \tag{2-2}$$

其中，N_d 表示文档 d 中的词汇数量，z_{dn} 是文档 d 中第 n 个词的主题分配，w_{dn} 是该词的具体内容。模型据此对语料库文档进行主题推断，从而识别各文档的核心主题。最后，LDA 模型的词汇分布可如公式(2-3)：

$$p(w|\beta)=\sum_{z} p(z|\beta)p(\omega|z,\beta) \tag{2-3}$$

通过 LDA 主题模型分析，可识别语料库中所有文档的主题结构，提取各文档的主要主题及其核心特征词。

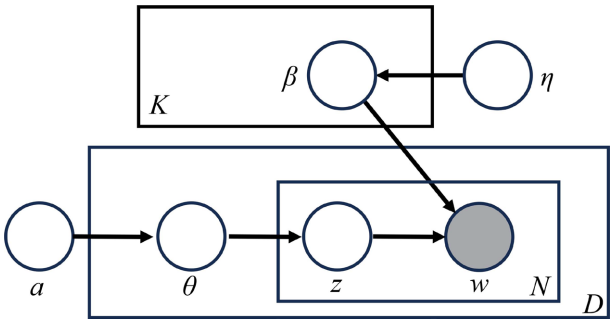

Figure 1. Underlying mechanism of LDA topic modeling
图 1. LDA 主题分析内在机理

2.2. 数据采集

本研究以中国知网(CNKI)为主要数据来源，以“数据素养”为检索主题，检索时间范围为 2010 年 4 月至 2025 年 12 月，共获得 3135 篇文献。经模糊筛选与人工复核，剔除与研究无关及非学术性文献后，最终保留 1291 篇有效样本，其中包括学术期刊 1073 篇、学位论文 154 篇、学术会议论文 29 篇、报纸文章 10 篇及特色期刊论文 25 篇。

2.3. 主题模型分析

LDA 主题模型分析包括三个主要环节：文本预处理以确保数据质量，模型构建以提取潜在主题结构，及基于输出结果的主题特征抽取与评价指标构建。以下将依次介绍具体实施流程。

1. 数据预处理

数据预处理是 LDA 主题分析的关键环节，用于降低噪声并提升主题提炼质量。本文选取论文题目与摘要构建语料库，借助 Python (NLTK、jieba 等)完成清洗与分词向量化处理：统一大小写、去除特殊符号与无意义字符，并结合哈工大与 Gensim 停用词表剔除无实义高频词、保留核心术语。

2. LDA 主题建模

数据预处理完成后，将结果通过 NLTK 导入 Gensim 库以进行 LDA 主题分析。模型需设定三个超参数：α、β 和 K。这里，α 是反映主题流畅度的参数，β 是反映特征词流畅度的参数，其计算公式如(2-4)：

$$l(\alpha,\beta)=\sum_{d=1}^{M}\log p(w_d|\alpha,\beta) \tag{2-4}$$

其中，α 值越小，文档越倾向于集中于单一主题；而 β 值越小，特征词越可能归属于单一主题。Blei 等研究表明，当 α 和 β 取 0.1 时，模型可获得更清晰且语义一致的主题结构，因此本文将二者均设定为 0.1。

超参数 K 表示语料库中的主题数量，其确定通常基于困惑度与一致性指标。困惑度用于衡量模型对语料的拟合效果，值越小说明模型泛化能力越强；一致性则通过计算特征词共现频率来评估主题语义关联性，值越高说明主题区分度与语义一致性更好。

困惑度的计算公式如(2-5)：

$$\text{perplexity}=\exp\left(-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right) \tag{2-5}$$

其中，w_d 是文档 d 中的词汇，N_d 是文档 d 中的词汇数量。

一致性计算公式如(2-6)：

$$\text{Coherence}(v)=\sum_{(v_i,v_j)\in v}\text{Score}(v_i,v_j,\epsilon) \tag{2-6}$$

其中，v 是主题中的特征词集合，(v_i,v_j) 是其中任意两个特征词对，其评分函数定义如(2-7)：

$$\text{score}(v_iv_j,\epsilon)=\log\left[p(v_iv_j)+\frac{\epsilon}{p(v_i)p(v_j)}\right] \tag{2-7}$$

通过结合 Scikit-Learn 和 LDAvis 工具对不同 K 值(1~30)进行分析，如图 2 可以看到，当主题数为 23 时，困惑度较低且一致性较高，模型性能最优。

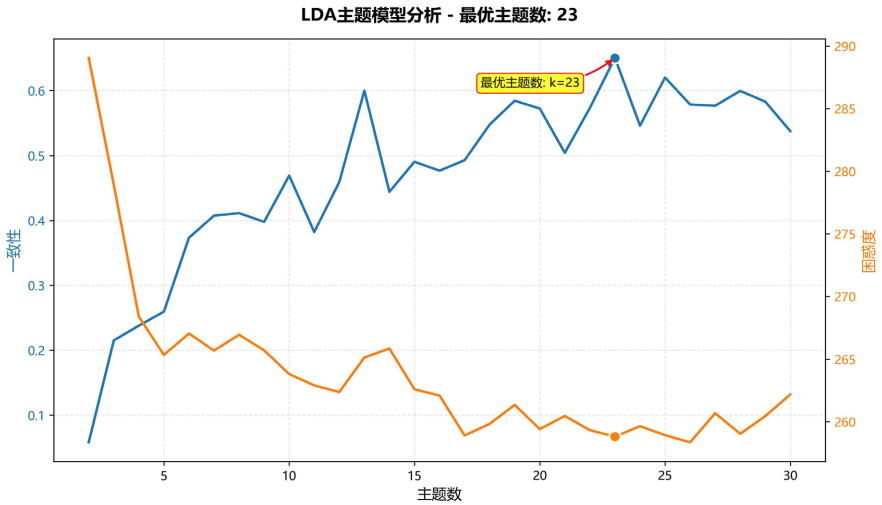

Figure 2. Perplexity and coherence analysis
图 2. 困惑度与一致性分析

3. 主题特征抽取

主题模型的输出本质为“主题－词项”概率分布，既包含代表性强但语义泛化的高频词，也包含区

分度高但可能低频的特征词。若仅按概率排序，易出现代表性有余而区分性不足；若仅按区分度排序，则可能引入稀疏且解释性不稳的词。为兼顾两者，本文在主题词提取中同时考虑词项在主题下的条件概率及其相对全局语料分布的提升程度，采用“相关性”综合排序进行加权计算，以增强主题解释的科学性与稳定性。其计算公式如(2-8)：

$$\text{relevance}(w,k)=\lambda\cdot\log p(w\mid k)+(1-\lambda)\cdot\log\text{lift}(w,k) \tag{2-8}$$

其中，λ 用于平衡高概率词与高区分度词的权重。本文取 $\lambda=0.6$，以兼顾主题词的代表性与区分性，避免仅由通用高频词主导而削弱解释力。最终，从每个主题中选取相关性得分最高的前 10 个词作为该主题的特征词集合。

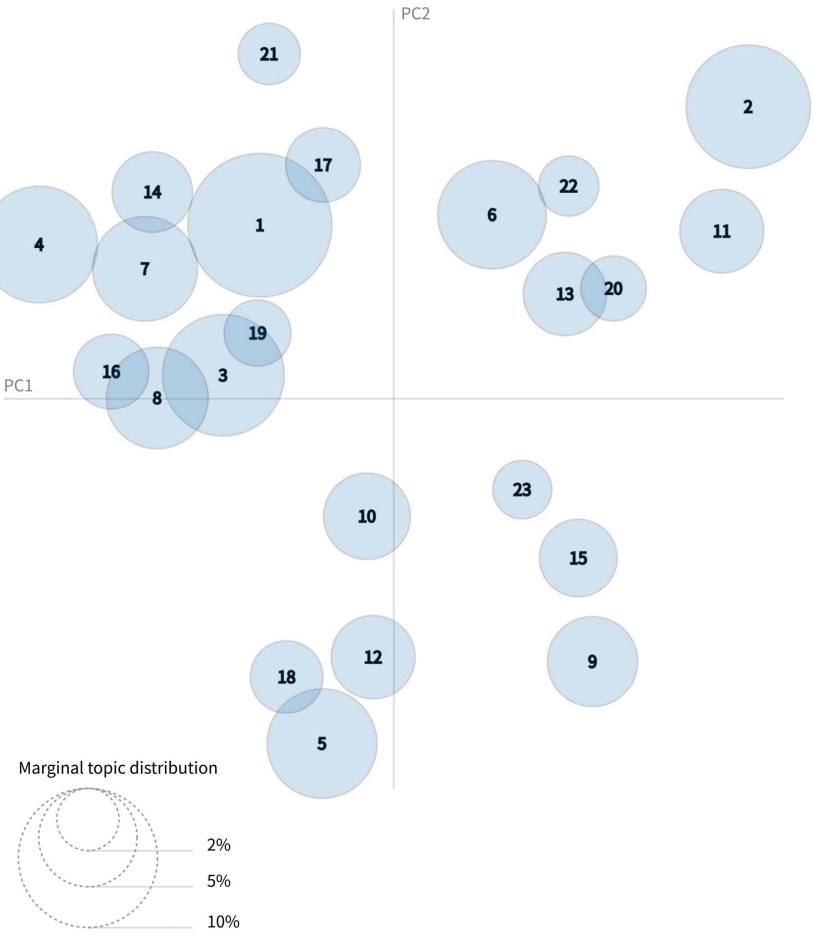

Figure 3. Intertopic distance map
图 3. 主题间距离图

如图 3 和表 1 所示，各主题的前 10 个特征词能够清晰地呈现主题的内涵与边界。例如，V1 的特征词集中于“评价、指标体系、德尔菲法”等，体现了“数据评价”方法论导向；V5 围绕“数据安全、风险、个人信息”等，强调了“数据保护”在隐私与法律层面的重要性；V16 则以“数据挖掘、关联规则、聚类”等为核心，反映了“数据挖掘技能”的技术取向。这种结合代表性与区分性的特征抽取方法，不仅能够“明确揭示各主题的核心内容”，还能够“有效区分它们与其他主题的差异”，为后续研究提供了更加精准的参考依据。

Table 1. LDA-derived topics and representative keywords
表 1. 基于 LDA 生成的主题及代表性关键词

主题编号	前 10 个特征词	主题名称
V1	评价、评价指标体系、指标体系、评价指标、指标、权重、问卷、调查、专家、德尔菲法	数据评价
V2	项目、教学模式、设计、教学实践、活动、探究、实验、行动研究、情境、技能	数据项目实践
V3	数据处理、分析、影响因素、因素、预测、关联规则、检验、结构方程模型、spss、特征	数据分析
V4	数据治理、科学数据管理、数据质量、一致性、模式、数据服务、服务、合作、融合、融合创新	数据管理
V5	数据安全、风险、安全、监管、法律、责任、个人信息、原则、数据安全治理、数据泄露	数据保护
V6	理念、数据思维、体系、课程体系、专业、培育、融合、路径、教学改革、导向	数据素养培养
V7	检索、知识图谱、知识库、关系、分类、语义、算法、结构化、文档、文章	数据检索
V8	数据可视化、可视化、图表、python、citespace、数据分析、大数据分析、语言、编程语言、信息	数据可视化
V9	教师数据素养、教师、数据知识、数据驱动教学、数据驱动决策、教学、教育信息化、教学决策、教育数据、精准教学	教育数据决策
V10	用户隐私、个人信息、策略、行为数据、信息、设计、方法、加密、问题、用户数据	数据隐私
V11	教师数字素养、数字素养、教师、培训、技能、数字技术、数字化教学、提升策略、提升路径、智慧教育	教师数字素养
V12	数据伦理、公平性、政策、社会、大数据环境、学生、科学素养、教学方法、教学效果、情境	数据伦理
V13	企业、核心、决策、领导力、数字化转型、转型、升级、体系、产业结构、商业	业务推动
V14	数据治理、建设、体系、全过程、对策、战略、层面、数智、智慧校园、教学资源	数据治理
V15	大数据、大数据时代、素养、智慧城市、人工智能技术、数据应用、模态、大数据技术、云计算、大数据思维	大数据应用意识
V16	数据挖掘、数据挖掘技术、关联规则、聚类、预测、预测模型、模型构建、规律、分析、sql	数据挖掘技能
V17	存储、数据存储、数据管理、架构、方案、分布式、云计算、节点、区块链、安全	数据存储

续表

V18	研究热点、热点、趋势、态势、脉络、关键词、文献、研究现状、视角、研究领域	数据解读
V19	数据处理、数据采集、编码、方法、模块、分析、数据分析、数据技术、信息技术、云计算	数据处理方法
V20	数据表述、仪表板、沟通协作、数据分享、信息共享、数据服务、平台、互联网、融合、培养机制	数据沟通
V21	监测、传感器、数字孪生、影像、空间、区域、特征、要素、模态、分析工具	数据监测
V22	信息化、战略、顶层、体系、信息化建设、数字中国、基础设施、国家、互联网、信息技术	信息化战略意识
V23	法规、政策、机制、利益、对策、逻辑、案例分析、方式、层级、特征	数据法规

本文以国内外关于数据素养内涵与结构维度的研究成果为理论锚点，在系统梳理既有模型的基础上，对一级指标进行整合与重构，如表 2。从国际视角看，联合国教科文组织统计研究所(UNESCO Institute for Statisticsm, UIS)发布的 A Global Framework of Reference on Digital Literacy Skills for Indicator 4.4.2 提出，数字素养是个体在数字环境中有效获取、理解、评价、创造和负责任使用信息与数据的综合能力体系，其中“信息与数据素养”被置于核心位置，并与技术操作、问题解决及伦理责任等维度形成有机整体[1]。总体而言，UNESCO 框架强调能力的通用性与跨情境适用性，适用于宏观监测与国际比较，但其能力划分相对整合，对高等教育情境中数据素养的具体结构刻画仍有细化空间。基于此，本文在遵循其核心理念的基础上，对数据素养结构进行情境化拓展。

在“数据基础能力”维度上，UNESCO 框架将数据获取、组织、管理与质量意识整体纳入“信息与数据素养”。本文结合 Pinto 等人的研究，将其进一步细化为覆盖数据全生命周期的基础支撑能力，突出数据评价、检索、管理、治理、存储与监测等要素[2]，以更好反映高等教育与组织层面数据治理的实践需求。

在“数据工具应用”维度上，UNESCO 框架强调数字工具使用能力，但未区分工具操作与分析建模层次。本文依据徐慧关于高校数据素养培养应以工具利用与分析技能为核心的观点[3]，将数据分析、可视化、挖掘与处理方法整合为独立维度，增强指标体系的可测量性与教学可操作性。

在“数据思维能力”维度上，本文与 UNESCO 框架高度契合。UNESCO 强调对信息与数据的批判性理解与判断，本文结合 Sun 的研究，大学生数据素养需强化“批判性数据思维”和价值判断取向，突出对数据意义的理解、评价与决策应用[4]。进一步将其具体化为数据解读、数据价值理解与基于数据的决策意识，突出数据素养由“技术使用”向“意义理解与决策支持”的认知跃迁。

在“数据伦理规范”维度上，UNESCO 框架将伦理、安全与责任视为数字素养的重要组成部分。本文在此基础上结合 Sun 的研究，数据隐私泄露与算法偏见等风险使得“数据伦理判断”成为不可或缺的评价维度[4]。引入法规意识与制度合规视角，将数据保护、隐私、伦理与法规整合为独立维度，使伦理要求从价值倡议延伸至规范约束，更契合当前数据治理法治化的发展趋势。

此外，UNESCO 框架虽强调情境化应用与问题解决，但未将实践能力单独上升为结构性维度。本文结合朱芳慧关于数据素养评价应突出真实情境中实践能力的观点[5]，提出“数据践行能力”维度，重点关注数据项目实践、数据沟通与组织应用，强化数据素养从“能力具备”向“能力落地”的转化。

综上，本文构建的“数据基础能力、数据工具应用、数据思维能力、数据伦理规范与数据践行能力”五个一级维度，在价值理念与能力取向上与 UNESCO 数字素养框架保持一致，确保了国际可比性；同时通过结构细化与情境拓展，增强了数据素养评价在高等教育领域的解释力与应用价值。

Table 2. Primary and secondary indicators for assessing college students' data literacy
表 2. 大学生数据素养测评的一级指标和二级指标

一级指标	二级指标	主要内容
T1 数据基础能力	V1 数据评价	能构建数据评价指标体系，合理设置指标与权重，结合问卷调查与专家咨询开展评价并解释结果。
	V4 数据管理	能开展数据质量控制与一致性维护，形成可复用的数据管理流程，支持共享与融合应用。
	V7 数据检索	能高效检索与组织文献及数据资源，理解知识库、知识图谱与语义检索的基本思路。
	V14 数据治理	具备全生命周期数据治理意识，能够设计系统化治理框架，支撑组织数字化建设。
	V17 数据存储	能选择合适的数据存储架构，理解系统设计与安全要求，保障数据可靠与可用。
	V21 数据监测	能理解多源监测数据特征，在数字孪生等场景中支持要素识别与多模态分析。
T2 数据工具应用	V3 数据分析	能进行数据处理与分析，开展预测或检验，使用常用统计工具完成实证分析。
	V8 数据可视化	能运用可视化工具呈现分析结果，并借助相关软件展示信息结构与知识关系。
	V16 数据挖掘技能	能运用数据挖掘方法进行模型构建与规律发现，具备基础数据查询与处理能力。
	V19 数据处理方法	掌握数据采集与处理的基本流程，能根据任务需求选择合适的处理策略。
T3 数据思维能力	V9 教育数据决策	理解教育数据价值，具备数据驱动教学与决策意识，支持教学改进。
	V15 大数据应用意识	具备“大数据 ＋AI”应用视角，理解典型应用场景，形成面向问题的数据思维。
	V18 数据解读	能识别研究热点与发展趋势，对数据与文献进行结构化解读。
T4 数据伦理规范	V5 数据保护	具备数据安全意识，能识别风险并采取相应的防护与治理措施。
	V10 数据隐私	能识别隐私边界，理解数据使用风险，掌握基本隐私保护方法。
	V12 数据伦理	理解数据公平与伦理问题，避免不当数据使用带来的偏差与风险。
	V23 数据法规	了解数据相关法律法规，能够在实践中落实合规要求。

续表

T5 数据践行能力	V2 数据项目实践	能在真实或模拟情境中开展数据项目实践，形成可应用的成果。
	V6 数据素养培养	能从课程与体系视角提出数据素养培养路径与改进方案。
	V11 教师数字素养	关注教师数字素养提升，理解培训与数字化教学应用路径。
	V13 业务推动	理解数据在组织决策与转型中的作用，具备数据支撑业务发展的意识。
	V20 数据沟通	能清晰表达数据结果，支持协作共享与平台化数据服务。
	V22 信息化战略意识	具备信息化战略视角，能将数据工作融入组织发展与制度建设。

4. 大学生数据素养研究热点与趋势

基于 LDA 主题模型与年度活跃度分析，本文计算了各年份主题的平均出现概率，以揭示大学生数据素养研究的演进趋势。

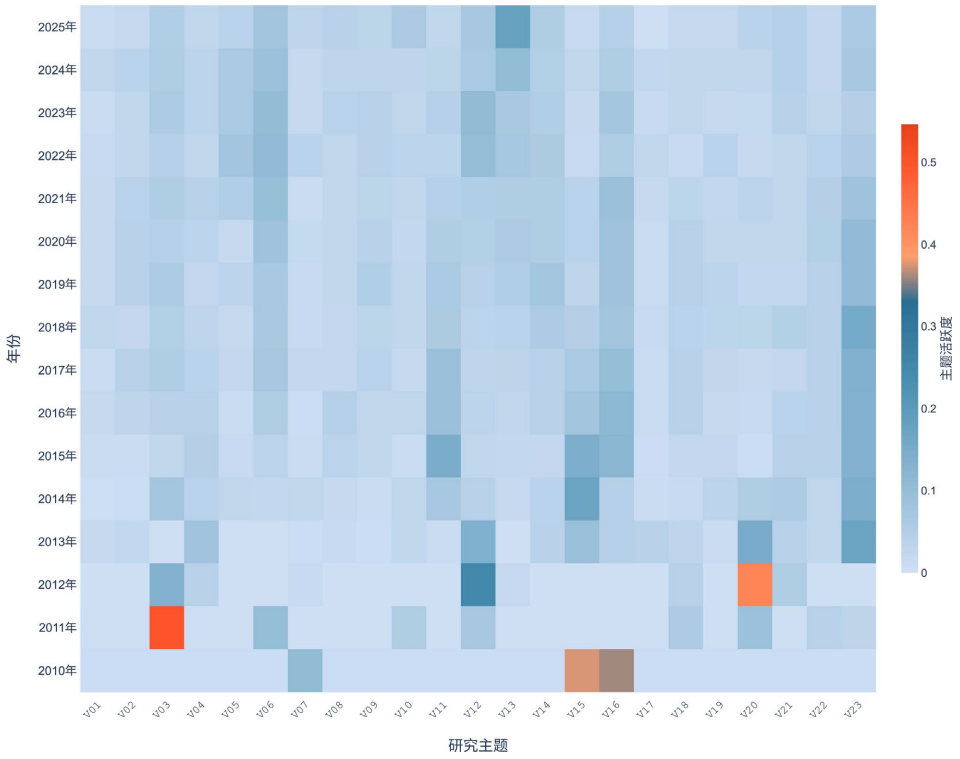

Figure 4. Heatmap analysis of topic evolution over time
图 4. 主题时间演进的热力图分析

由图 4 和图 5 可见，近年来，随着数字化转型的加速，数据分析、数据安全与数据治理等技术主题迅速升温，成为大学生数据素养研究的核心内容，反映出社会对数据能力与安全意识的高度重视。与此同时，教育领域中关于教师数据素养与数据驱动教学的研究也显著增长，凸显数据在提升教学质量与决策中的关键作用。自大数据与人工智能广泛应用以来，数据驱动决策和智能分析逐渐成为学术与产业研究的热点。总体来看，大学生数据素养的研究呈现出技术能力、数据伦理与教育实践相结合的多维趋势，体现出在数字化时代全面提升数据素养的迫切需求。

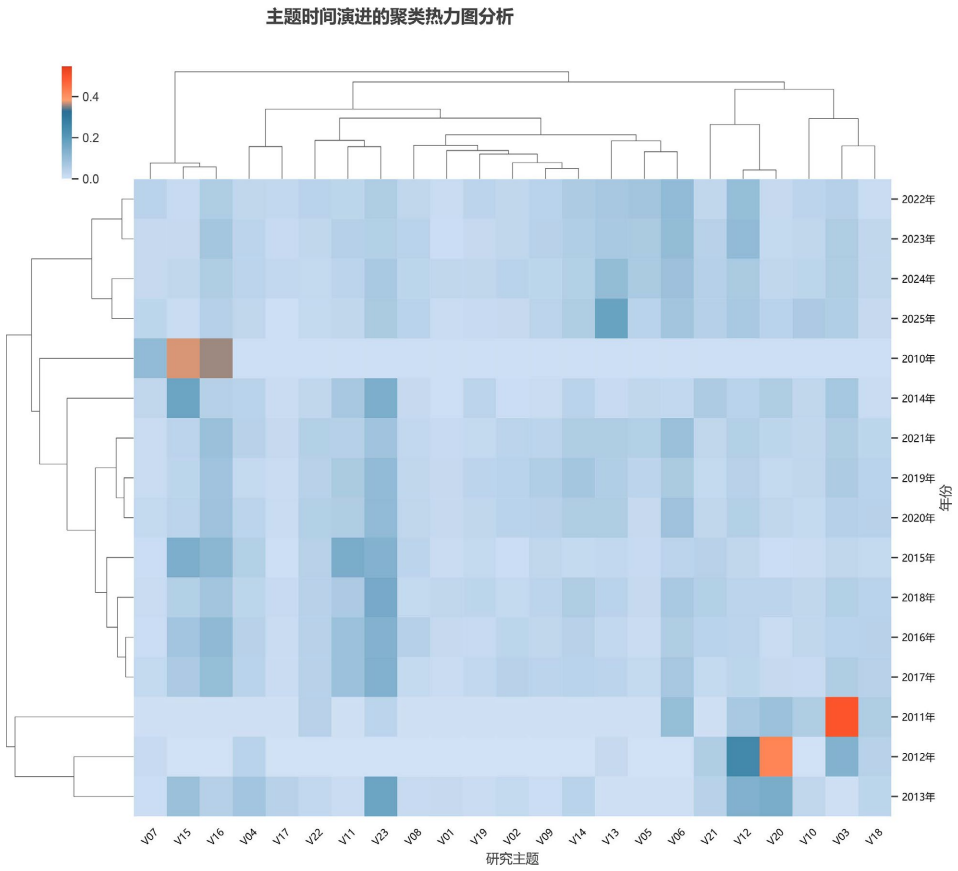

Figure 5. Hierarchically clustered heatmap of topic evolution over time
图 5. 主题时间演进的聚类热力图分析

5. 总结与展望

本研究立足信息化时代数据素养的战略价值，针对大学生数据素养差异显著及既有测评体系适配性不足的问题，运用 LDA 主题模型对近十年相关文献进行系统分析，构建了包含数据基础能力、工具应用、思维能力、伦理规范与践行能力五个维度、23 项指标的大学生数据素养评价体系。该体系以数据驱动的量化结果为依据，融合国内外理论框架，覆盖数据全生命周期核心能力，为高校数据素养测评与课程建设提供了参考。

未来可在此基础上进一步拓展指标内涵，将生成式 AI、数字孪生等新兴技能纳入评估；开展面向不同学科与年级群体探索差异化培养路径；开发配套教学与测评工具，推动数据素养教育与专业课程深度融合。同时，加强国际比较与本土化适配，借鉴先进经验优化教育模式，推动数字时代复合型人才培养。

伦理规范

本人郑重申明，所开展的“基于 LDA 主题模型的数据素养评价指标体系研究”严格遵循学术伦理规范与相关法律法规，秉持诚信严谨的科研态度。研究原始数据真实可追溯，隐私信息得到严格保护；规范引用与署名，无任何学术不端行为；已充分披露潜在利益冲突，切实维护学术诚信。

基金项目

本文系 2023 年度教育部人文社会科学研究规划基金项目“大学生数据素养评价体系理论及实证研

究”(23YJAZH022)；中国民办教育协会 2025 年度规划课题“智慧软件在大学生人工智能素养培养中的应用研究”(CANQN250577)；西京学院 2025 年度教育教学改革研究项目“人工智能素养培养的‘积木式’教学工具开发与创新教学模式研究”(JGYB2527)的研究成果。

参考文献

[1] Law, N., Woo, D., de la Torre, J. and Wong, G. (2018) A Global Framework of Reference on Digital Literacy Skills for Indicator 4.4.2. UNESCO Institute for Statistics.

[2] Pinto, M. and Segura, A. (2025) Toward a Conceptual Framework on Mobile Information Literacy in Higher Education. *The Journal of Academic Librarianship*, **51**, Article 103051. https://doi.org/10.1016/j.acalib.2025.103051

[3] Sun, C. (2024) Investigation on the Data Literacy of College Students and the Promotion Strategies in the Era of Big Data in China. *Exploration of Educational Management*, **2**, 82-87.

[4] 徐慧. 大数据时代大学生的数据素养教育[J]. 新闻战线, 2017(4): 126-127.

[5] 朱芳慧. 基于高校师范生数据素养培养的表现性评价体系优化设计[J]. 高教研究与实践, 2023, 40(5): 90-98.