基于LDA主题模型与复杂网络的土石坝病险特征诊断模型研究

李陈瑶 [1,2] 王芳 [1,2] 张铸 [1,2]

1. 南京水利科学研究院，中国 ·江苏 南京 210029

2. 水利部东北寒区长距离有压供水工程野外科学观测研究站，中国 ·江苏 南京 210029

摘要：水库大坝病险的合理诊断是有效开展除险加固工作的重要基础。针对当前土石坝除险加固工作中存在的诊断效率低、历史知识难以系统沉淀与复用等问题，本文提出了一种融合 LDA 主题模型与复杂网络分析的土石坝病险特征智能诊断模型。该模型通过对历史病险文本资料的挖掘分析，实现了对土石坝病险知识的系统提取和可视化表达。研究以 20 座土石坝工程的安全鉴定资料为样本，运用 LDA 主题建模方法成功识别出防洪安全、渗流安全、结构安全、抗震安全及金属结构安全五类主题及其特征词。在此基础上，构建了主题词复杂网络，利用中心性指标深入解析了病险特征与除险措施之间的关联机制，形成了可解释的决策规则。研究表明，本文构建的模型能够有效实现土石坝病险特诊的系统挖掘，为工程实践提供数据驱动的决策支持。该研究方法为水利工程安全管理从经验驱动向“经验－数据”双驱动转型提供了新的技术途径，具有重要的理论价值和实践意义。

关键词：土石坝；文本挖掘；病险诊断；LDA 主题模型；复杂网络分析

Research on an Intelligent Diagnosis and Decision Support Model for Earth–Rock Dam Defect Characteristics Based on LDA Topic Modeling and Complex Networks

Li Chenyao[1,2], Wang Fang[1,2], Zhang Zhu[1,2]

1. Nanjing Hydraulic Research Institute, China Jiangsu Nanjing 210029

2. Northeastern Cold Region Long-Distance Water-Supply Project Observation and Research Station, Ministry of Water Resources, China Jiangsu Nanjing 210029

Abstract: Rational diagnosis of dam defects is essential for effective rehabilitation. To address the limitations of low efficiency and poor knowledge reuse in the current expert experience-based approach for earth-rock dam rehabilitation, this study developed an intelligent diagnostic model that integrates Latent Dirichlet Allocation (LDA) topic modeling and complex network analysis. This model enables systematic extraction and visualization of defect-related knowledge from historical textual records. Using safety assessment reports from 20 earth-rock dam projects as a sample, the LDA topic modeling successfully identified five safety themes—flood control, seepage, structural, seismic, and metal structure safety—along with their characteristic keywords. A complex network of thematic keywords was constructed, and centrality metrics were employed to analyze the association mechanisms between defect characteristics and reinforcement measures, forming interpretable decision rules. The results demonstrate that the proposed model effectively facilitates systematic mining of earth-rock dam defect characteristics and provides data-driven decision support for engineering practice. This research offers a new technical pathway for transitioning hydraulic engineering safety management from an experience-driven paradigm to an integrated "experience-data" dual-driven approach, possessing significant theoretical and practical value.

Keywords: Earth-rock dams; Text mining; Defect diagnosis; LDA topic model; Complex network analysis

0 引言

我国水库大坝安全是国家水安全的重要保障。截至 2024 年底，全国已建水库近十万座，其中土石坝占比超过 90%，且普遍存在老化、病险问题[1]。尽管自 1998 年起国家系统性推进除险加固，成效显著，但近年仍出现了加固后或加固过程中溃坝的案例（如 2020 年内蒙古永安、新发水库）[2-3]，表明除险加固工作具有长期性和复杂性。根据《全国病险水库除险加固实施方案（2025—2027 年）》，未来将对约 5000 座病险水库进行整治，土石坝因其数量众多成为重点对象。在极端气候频发的背景下，发展高效、精准的智能诊断方法以提升决策效率与科学性，已成为行业迫切需求。

当前，土石坝病险诊断主要依赖“一案一策”的专家经验模式[4]。该模式虽行之有效，但在面对海量工程时，存在效率低下、历史知识难以系统沉淀与复用等瓶颈。过去数十年积累的安全鉴定与除险加固文本资料，构成了蕴含丰富经验的“知识富矿”，然而其非结构化形态使得传统方法难以进行深度挖掘。文本挖掘技术，如 LDA 主题模型，能有效识别文本中的潜在主题结构[5]，但传统模型在分析短文本时存在语义关联挖掘不足的局限。而复杂网络分析则擅长揭示实体间的深层关联。现有研究尚缺乏将二者有效融合，以系统挖掘土石坝病险特征与除险措施间关联规则的方法。

为此，本研究构建了一种融合 LDA 主题模型与复杂网络分析的智能诊断模型。通过对 20 座土石坝安全鉴定文本进行挖掘分析，生成可解释的“病险 – 措施”决策规则，为快速筛查与精准除险提供数据驱动支持，推动水利工程安全管理向“经验 – 数据”双驱动模式转型。

1 研究方法

1.1 土石坝病险文本的结构化预处理

本研究以 20 份内蒙古、宁夏地区的土石坝安全鉴定报告为样本，其安全评价范畴遵循《水库大坝安全评价导则》（SL 258-2017），涵盖防洪、渗流、结构、抗震及金属结构五大安全主题。为从非结构化文本中提取有效信息，研究进行了以下关键预处理步骤：

（1）构建专用词表：包括停用词表（过滤“工程”“属于”等无意义高频词）、自定义词典（基于《水利水电工程技术术语》（SL 26-2012）确保专业术语准确切分）及替换词表（统一同义词表述），以提升分词准确性。

（2）分割语料集：将每份报告中的“病险诊断”与“除险措施”内容分离，建立两个独立语料集，避免主题建模时的语义干扰。

（3）分词与清洗：运用 Python 的 Jieba 库加载上述词表进行分词，得到纯净的词语序列。

通过上述流程，原始文本被转化为可供 LDA 模型分析的结构化数据。

1.2 基于 LDA 模型的文本特征挖掘

LDA 主题模型是一种无监督的生成式概率模型，其核心思想是将文档视为一系列潜在主题的概率混合，而每个主题又是词语的概率分布（图 1）。该模型能有效实现文本的语义降维与特征提取，将高维的文档 – 词语矩阵分解为低维的文档 – 主题矩阵和主题 – 词语矩阵，从而挖掘出文本集合中隐含的语义结构。

在建模过程中，主题数量的确定至关重要。为寻求语义解释力与模型复杂度之间的最佳平衡，本研究采用困惑度（Perplexity）作为评估指标。通过绘制不同 K 值下的困惑度曲线（图 2），并选取其拐点处的数值作为最优主题数，以避免主题语义混杂或过拟合，LDA 主题模型生成过程见图 2 所示。

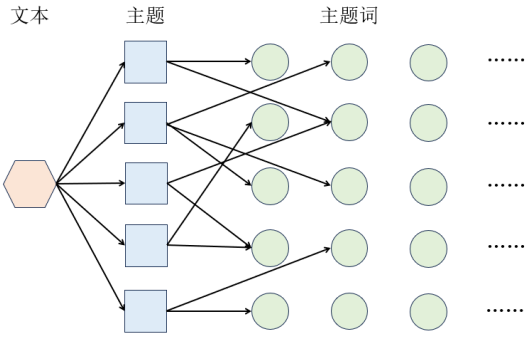

图1 三层贝叶斯概率结构框架

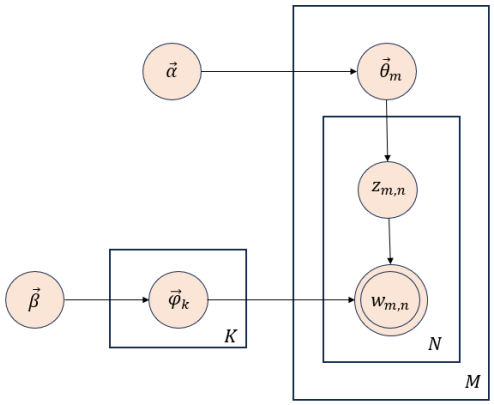

图2 LDA主题模型生成图

1.3 基于复杂网络的特征关联机制分析

为克服传统 LDA 模型在短文本分析中语义关联挖掘不足的局限，本研究进一步引入复杂网络分析方法。该方法以 LDA 提取的主题词为节点，以其在文本中的共现关系为边，构建无向加权网络，从而揭示词语间的深层语义关联。

在网络中，节点大小表征词频高低，边权重对应共现强度。为识别网络中的关键节点，本研究选取度中心性、中介中心性、接近中心性和特征向量中心性四项指标进行量化分析。这些指标分别从局部连接性、桥梁作用、中心位置及邻居影响力等维度评估节点重要性，共同揭示土石坝病险特征与除险措施之间的核心关联路径。

复杂网络分析超越简单的词频统计，通过挖掘“病险特征—除险措施”之间的强关联规则，形成可解释的决策知识，为土石坝病险的快速诊断与方案制定提供数据驱动的辅助支持。

2 LDA 主题模型的构建与结果分析

2.1 模型构建与参数确定

在完成文本预处理后，本研究将处理后的“病险特征”与“除险措施”语料分别导入 LDA 主题模型进行训练。为确定最优主题数量，通过计算困惑度进行评估。如图 3 所示，当主题数 K=5 时，模型困惑度降至最低（84.2），表明此条件下模型具有最佳的语义区分度与拟合效果。据此，设定主题数为 5，并配置超参数 α=0.01、β=0.0001，以平衡主题分布的稀疏性与主题内词语的集中程度。模型经过 1000 次迭代后趋于收敛，为后续分析提供了稳定基础。

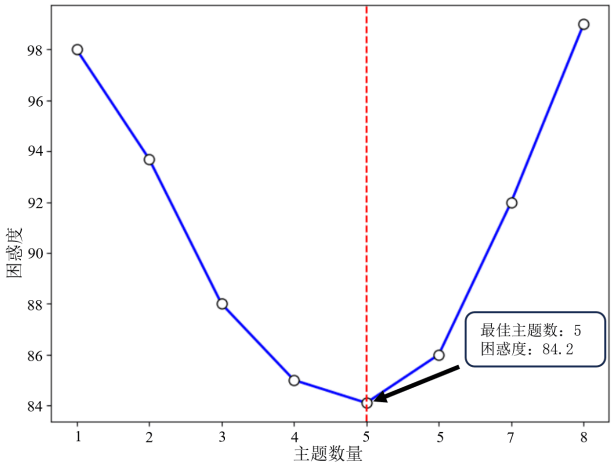

图3 困惑度曲线图

2.2 主题识别与语义解析

基于主题词分布特征并参照《水库大坝安全评价导则》（SL 258-2017），本研究定义了防洪安全、渗流安全、结构安全、抗震安全及金属结构安全五类主题，其对应的关键词分别揭示了病险特征与除险措施的核心内容。

病险特征主题词提取结果。分析表明：防洪安全主题的核心病险集中于泄洪能力不足、防洪标准不足及高程不足，主要体现于溢洪道、泄洪洞等建筑物的冲蚀；渗流安全以渗漏为核心，主要表现为坝基渗漏、绕坝渗漏及接触渗漏；结构安全主要体现为裂缝、不均匀沉降、冻融破坏等混凝土结构损伤；抗震安全重点关注渗漏通道及穿坝建筑物等薄弱环节；金属结构安全则主要表现为闸门启闭设备的锈蚀、老化及漏水。

除险措施主题词提取结果。相应分析表明：防洪安全措施以重建、培厚、补强泄洪设施为主，并注重应急预案编制；渗流安全遵循“上截下排”原则，以帷幕灌浆、反滤排水及防渗墙为核心措施；结构安全重点采用加固、修复及高压喷射灌浆等技术；抗震安全以加固、防渗墙及强夯等地基处理为主；金属结构安全则侧重于金属结构的更换、除锈、防腐及加装监测系统。

3 土石坝病险特征的复杂网络分析

3.1 数据准备与网络构建

在完成 LDA 主题识别后，依据所划分的防洪安全、渗流安全、结构安全、抗震安全及金属结构安全五类主题，将文本数据重新整合为对应语料集，每个语料集包含相应主题下的病险特征与除险措施文本。在此基础上，为揭示主题词间的深层语义关联，进一步构建主题词共现网络。具体流程为：首先统计词对在上下文中的共现频次，选取频次分布前 10% 的阈值以筛选强语义关联词对；继而构建共现矩阵，矩阵值反映词对关联强度，累计值则体现节点的整体影响力。该矩阵为后续网络可视化与中心性分析提供了数据基础。

本研究构建了无向加权网络，节点表示主题词，边表示其共现关系，权重对应共现频次。采用无向网络旨在识别病险特征与除险措施之间的整体关联模式，而非判定因果方向，从而有效揭示其间稳定的共生关系与关键节点。

3.2 网络中心性分析

为综合评估节点影响力，对中心性指标进行归一化处理并计算综合评分，识别出各主题下排名前十的关键节点，结果见表 1（以防洪安全主题为例）。

（1）防洪安全：度中心性分析表明，防洪标准不足、泄洪能力不足及坝顶高程不足是主要病险。中介中心性与接近中心性较高的节点（如泄洪能力不足、渗漏）需优先处理，而特征向量中心性突出的节点（如防渗处理、加高培厚）代表关键除险措施。

（2）渗流安全：渗漏在度中心性中居核心地位，主要发生在坝基，并常由裂缝引发。中介中心性显示止水失效与排水淤堵会加剧渗漏，特征向量中心性较高的节点（如帷幕灌浆、反滤排水）既为关键措施，也反映特定渗漏形态。

（3）结构安全：裂缝是主要病险，多见于溢洪道、护坡等部位。中介中心性指出冻融破坏与不均匀沉降是其主要成因，特征向量中心性则凸显灌浆、补强等加固措施的重要性。

（4）抗震安全：渗漏为抗震安全的核心问题，易诱发结构裂缝与接触带软化。中介中心性表明穿坝建筑物（如输水涵管）是薄弱环节，特征向量中心性验证了结构加固与渗漏路径管控（如防渗墙、回填）为主要措施。

（5）金属结构安全：闸门与启闭设备的老化、锈蚀是

表1 防洪安全主题网络中心性分析

节点	度中心性	中介中心性	接近中心性	特征向量中心性
渗漏	20	95.050	0.744	1.000
溢洪道	16	55.687	0.617	0.861
防洪标准不足	13	89.257	0.580	0.644
泄洪能力不足	13	100.000	0.580	0.628
高程不足	12	27.917	0.460	0.417
泄洪洞	11	71.870	0.558	0.419
防渗处理	10	0	0.537	0.607
岸坡	10	0	0.537	0.607
加高培厚	10	0	0.537	0.607
编制调度规程	10	0	0.537	0.607
…	…	…	…	…

主要病险。中介中心性显示其功能性退化与止水破损易导致渗漏，综合中心性分析表明设备更换、防腐及智能监测系统是保障长期稳定的关键措施。

4 结语

（1）本研究构建了融合 LDA 主题模型与复杂网络分析的土石坝病险智能诊断方法，实现了从非结构化文本中自动识别五类安全主题及其特征词。通过中心性分析发现“渗漏”是串联多重病险的关键节点，揭示了病险特征与除险措施间的内在关联机制。该方法突破了传统经验模式的局限，为水利工程安全管理提供了可解释的数据驱动决策支持。

（2）相较于已有研究主要采用单一文本挖掘技术的做法，本文创新性地将 LDA 主题模型与复杂网络分析相结合，不仅识别了主题特征，更深入挖掘了病险与措施间的复杂关联。这一方法拓展了文本挖掘在水利工程领域的应用深度，为行业知识沉淀和复用提供了新途径。

（3）本研究目前仍存在对低频病险识别能力有限、样本依赖性较强等不足。未来将通过扩充样本库、引入迁移学习技术，并结合监测数据等多源信息，进一步提升模型的泛化能力和实用价值，推动水利工程病险诊断向智能化方向发展。

参考文献：

[1] 盛金保，李宏恩，盛韬桢. 我国水库溃坝及其生命损失统计分析[J]. 水利水运工程学报，2023(01): 1-15.

[2] 盛金宝，刘嘉炘，张士辰等. 病险水库除险加固项目溃坝机理调查分析[J]. 岩土工程学报，2008(11): 1620-1625.

[3] 张士辰，李宏恩. 近期我国土石坝溃决或出险事故及其启示[J]. 水利水运工程学报，2023(01): 27-33.

[4] 胡学同，周杏鹏，李雷等. 基于案例推理的土石坝病险智能诊断系统[J]. 计算机工程，2003(07): 163-165.

[5] 袁军鹏，朱东华，李毅等. 文本挖掘技术研究进展[J]. 计算机应用研究，2006(02): 1-4.

基金项目：国家重点研发计划项目（2024YFC3210604）；国家自然科学基金项目（U2443231）；中央级公益性科研院所基本科研业务费专项资金项目 (Y724003，Y724008，Y725006，Y725008，Y725009)；引绰济辽工程科研项目（YC-KYXM-11-2024）；南京水利科学研究院研究生学位论文基金（Yy725012）；世界一流的三峡集团水库水电站大坝安全管理能力建设咨询项目（Hj724122）。

作者简介：李陈瑶（2001-），女，汉族，湖南省株洲市，硕士研究生，研究方向：主要从事水库大坝安全评估与病险诊断方面研究。