

面向事故报告与指标项关联挖掘的根原因分析^①



陈鹏运¹, 房敏营¹, 侯孝波², 杜军威¹

¹(青岛科技大学 数据科学学院, 青岛 266061)

²(中石化安全工程研究院有限公司, 青岛 266104)

通信作者: 杜军威, E-mail: dujunwei@qust.edu.cn

摘 要: 为充分挖掘专家历史事故根原因分析经验, 提升事故根原因分析的准确性与全面性, 减少化工安全事故发生, 本文提出一种基于历史事故报告与根原因指标体系关联挖掘的分析方法. 通过构建事故报告与指标体系的关联矩阵, 采用预训练模型对事故文本和指标项文本进行表征, 基于注意力机制融合二级和三级指标项的信息, 最后运用图卷积神经网络进行事故根原因分析. 通过在 1351 份样本上进行验证, 实验结果表明: 该方法显著提高了根原因预测准确性, 能够有效利用专家历史分析经验辅助当前事故分析, 并揭示历史事故分析不完整的问题. 同时, 该方法也能通过不完整事故描述准确挖掘事故根原因, 该方法的应用将提升事故预防和安全生产风险管理的水平.

关键词: 化工安全; 根原因分析; 图卷积神经网络; 潜在关联关系

引用格式: 陈鹏运, 房敏营, 侯孝波, 杜军威. 面向事故报告与指标项关联挖掘的根原因分析. 计算机系统应用, 2025, 34(2): 272-280. <http://www.c-s-a.org.cn/1003-3254/9746.html>

Root Cause Analysis Based on Association Mining Between Accident Report and Indicator

CHEN Peng-Yun¹, FANG Min-Ying¹, HOU Xiao-Bo², DU Jun-Wei¹

¹(School of Data Science, Qingdao University of Science & Technology, Qingdao 266061, China)

²(SINOPEC Safety Engineering Institute Co. Ltd., Qingdao 266104, China)

Abstract: This study proposes an analysis method based on association mining between historical accident reports and a root cause index system to fully leverage experts' experience in root cause analysis of past accidents and enhance the accuracy and comprehensiveness of such analysis, thereby reducing chemical safety incidents. By constructing an association matrix between accident reports and the index system, this method utilizes a pre-trained model to represent accident and index texts. It integrates secondary and tertiary index information based on an attention mechanism and finally employs a graph convolutional neural network for root cause analysis. Validation on a dataset of 1351 samples demonstrates that this method significantly improves the accuracy of root cause prediction, effectively utilizing expert analysis of historical accidents to analyze current accidents and uncover the limitations in previous accident analysis. Additionally, this method accurately identifies the root causes of accidents even with incomplete incident descriptions. The application of this method will enhance accident prevention and risk management in occupational safety.

Key words: chemical safety; root cause analysis (RCA); graph convolutional neural network (GCNN); latent related relationship

由于化工生产过程多涉及大量易燃、易爆、有毒的危险化学品, 任何过程上的“小事件”都可能引发重

特大灾害事故. 深入分析这些事故的背后原因, 发现多是由于责任人员安全意识淡薄、企业安全管理制度不

① 基金项目: 山东省自然科学基金(ZR2021MF092); 山东省重点研发计划(2018GGX101052)

收稿时间: 2024-06-22; 修改时间: 2024-07-18; 采用时间: 2024-08-01; csa 在线出版时间: 2024-12-16

CNKI 网络首发时间: 2024-12-17

完善等管理原因导致的^[1]。博德事故连锁理论认为人的不安全行为或物的不安全状态是引发事故的直接原因,间接原因是个人因素及与工作有关的因素,根源原因是管理失误^[2]。根源原因是指导致事故发生的最深层次的原因,通常与管理、操作和维护等系统性问题有关,通过根原因分析能够挖掘事故背后的企业管理缺陷。

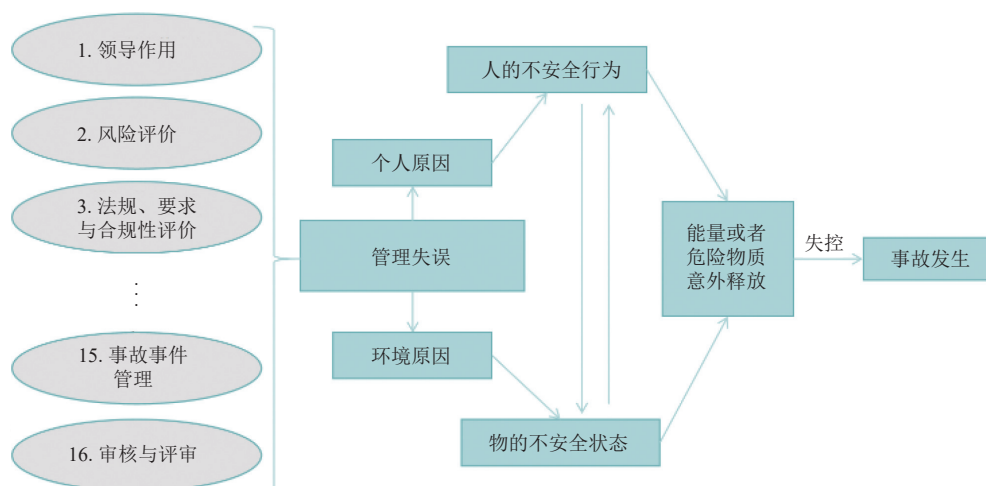


图1 事故根原因分析图

目前用于化工事故根原因分析的传统方法有多种^[3],例如逻辑树法^[4]、事故树分析法^[5]、故障树分析法^[6-9]、安全检查表法^[10]和因果图法^[11]等。虽然上述方法为化工安全提供了一些有效信息和经验,但仍然存在种种问题^[12-15]。例如,安全检查表法^[10]和因果图法^[11]依赖于有限的事案例和专家知识,缺乏足够的先验知识导致其无法挖掘事故与指标项之间的关联关系,并难以涵盖所有的事案例情况。逻辑树法^[4]、事故树分析法^[5]和故障树分析法^[6-9]没有充分利用大规模数据和现代分析技术^[3,16],降低了数据处理效率和效果,如何解决这些问题成为事故案例智能化分析的瓶颈。同时,上述方法仅适用于对单一事故进行分析,并依赖专家经验来推断事故成因演化机理。由于事故成因的复杂性和案例间的差异性,这种依赖于单一事故情景和专家经验的方式使得大量已分析的历史案例难以直接借鉴^[17]。此外,由于根原因一般是通过事故发生后的定性分析得出,缺少量化指标,导致其难以用于企业安全生管理。因此,需要设计一种特定的指标项来代表根原因分析中的各个方面。具体来说,每个指标项对应一个可能导致事故发生的具体问题或风险因素。

随着技术和理论不断发展,图卷积神经网络展现出了强大的推理能力,因此可以使用人工智能方法

如图1所示,展示了事故分析的过程,强调了对不安全行为、状态及管理漏洞的综合评估,涵盖了安全管理、设备状况、操作程序和员工培训等多维度指标。因此,进行事故根原因分析 (root cause analysis, RCA) 是企业识别生产过程中潜在安全隐患、找出管理漏洞,从而提升安全管理水平、预防和减少事故的重要方法。

从大量历史数据中提取规律,揭示历史事故与指标项之间的潜在关联关系^[18,19]。基于上述内容阐述,本文做出了以下几点贡献。

(1) 本文提出一个多级指标项体系,以定量化研究事故根原因,并采用表示学习技术,通过融合多阶信息来深度挖掘指标项的三阶树形结构信息,提升了对根原因体系之间结构信息的表达能力。

(2) 提出了一种基于图卷积神经网络的事事故根原因分析方法模型,通过分析历史事故报告案例与事故指标项之间的图关系,来深度挖掘事故报告和指标项潜在关联关系。

1 相关工作

1.1 化工事故根原因分析国内外研究现状

在化工安全领域,根原因分析是一种有效识别并解决问题根源的方法。过去的化工事故根原因分析主要依赖于人工分析和统计方法,缺乏系统性和实时性。2016年,邹吉然等人^[7]研究了故障树分析法 (fault tree analysis, FTA) 在化工安全评价中的应用,阐述了该方法如何帮助提高化工行业的安全性,为后续的系统可靠性分析和故障诊断提供了有效的理论工具和实践方法。事故树分析法 (event tree analysis, ETA)^[5]主要用于

评估一个初始事件如何在一系列后续的条件和选择中进展,从一个已知的初始事件开始,然后探索可能的后续事件链和不同的结果.与事故树分析法不同,故障树分析法从顶部的不良事件开始,逐层探究可能的原因.2007年,卜全民等人^[5]探讨了事故树分析法的应用研究,提出了该方法在事故预防和诊断中的价值,但上述方法没有充分利用大规模数据和现代分析技术^[3,16],降低了数据处理效率和效果.2005年,王婕等人^[11]探讨了因果图分析法在液氯泄漏事故分析中的应用,揭示了该方法在事故原因分析中的有效性,因果图法通过绘制“鱼骨”结构图,列举分析问题的所有潜在原因,以找出其根本原因.2020年,张静涛等人^[10]研究了安全

检查表法在危险化学品罐区安全评价中的应用,强调了该方法如何助力于化学品存储区的安全管理,安全检查表法是根据事故细节,基于相关法规和技术标准,针对危险化学品罐区的潜在风险进行评估的工具,旨在检查设备、设施的安全性及日常管理有效性,并提供改进建议,但以上两种方法过度依赖于有限事故案例和专家知识,导致无法挖掘事故与指标项之间的潜在关联关系,难以涵盖所有的事故情况.事故分析是专家依据标准化的评估模板对事故进行系统性评估,全面分析事故的影响和未来风险,并将评价指标作为评分依据,运用评分表对事故进行定量化分析,形成一个事故-指标项评分表,如图2所示.

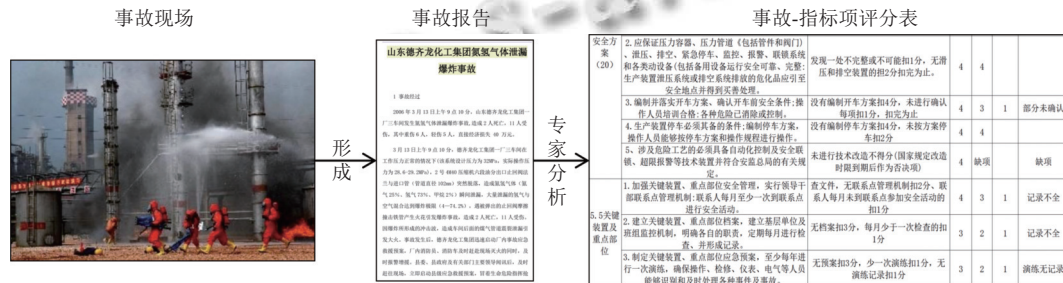


图2 事故分析流程图

1.2 量化评估指标

本实验数据来源于中石化安全工程研究院,数据内容包括安全量化评估指标体系的详细信息以及其他相关数据.安全量化评估指标体系由一级、二级、三级这3大类指标项构成,分别包含16个、92个、534个.该体系是参考多个国外评估机构的指标体系,并结合我国化工安全领域实际情况自行修改完成的,具体关系如图3所示.

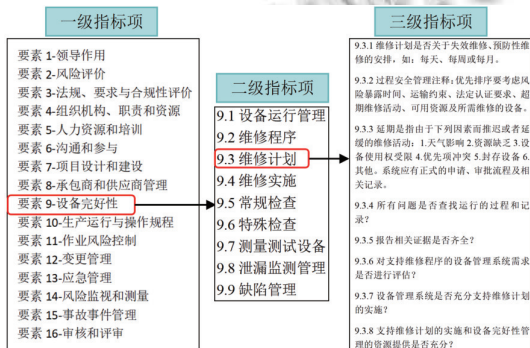


图3 安全量化评估指标项展示

该机构除提供上述评估指标体系之外,还提供了时间跨度为1975–2019年的1351个事故案例样本,该

事故样本中包含文档编号、事故标题、事故内容以及事故报道网址,事故内容包含事故发生的时间、事故涉及的企业类型以及事故类型等详细信息.

2 方法设计

如图4所示,本文模型主要分为以下几个部分.

(1) 文本嵌入表示模块,利用BERT模型对事故文本与指标项文本分别进行语义嵌入表示,将二级指标项与三级指标项的嵌入进行权重融合,如图4(a)所示.

(2) 结构关系捕获模块,将上述得到的事故与指标项语义向量输入到图卷积神经网络模块,以得到最终的融合结构关系的事故与指标项嵌入,如图4(b)和图4(c)所示.

(3) 得分预测模块,使用非线性激活函数和全连接网络获取事故的根原因得分,如图4(d)所示.

2.1 评分关系矩阵

由数据集可以得到一个事故与指标项的评分关系矩阵 $R \in \mathbb{R}^{N \times M}$,该矩阵中 N 表示事故报告文本数量, M 表示二级指标项数量,共有1351条事故案例数据和92条二级指标项数据,如图5所示.

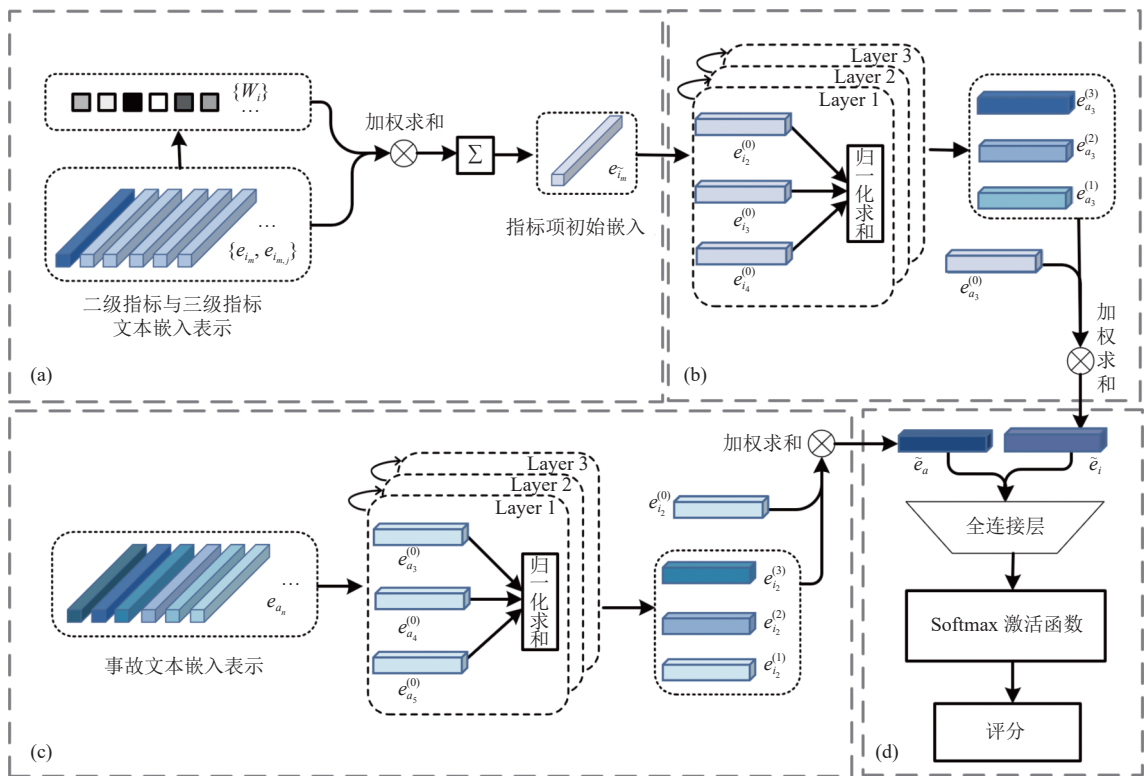


图4 事故-指标项评分关系矩阵图

指示项 事故报告	I_1	I_2	...	I_m	...	I_M
A_1	?	5	...	3	...	?
A_2	1	?	...	?	...	1
...	...	...	...	...	...	...
A_n	?	?	...	4	...	?
...	...	...	...	...	...	...
A_N	2	?	...	?	...	3

图5 融合事故报告及多级指标项的事故评分补全推荐模型

图5中数字表示专家对于指标项 I 与事故 A 发生影响程度的评分等级. 由于该数据中仍存在一些空值以及部分异常数据, 需要对这一部分数据进行清理. 评分等级限制在1-5之间, 1级最低, 5级最高, 同时等级越高则代表该指标项对事故发生的影响越大. 通过这些评分数据, 我们可以构建事故-指标项评分关系矩阵用于接下来的操作.

2.2 多级指标项融合嵌入

针对化工事故报告这类上下文强关联性文本, 我们使用利用BERT模型进行处理, 以能够有效地捕捉上下文关系, 同时能对每个词的含义和语境进行更全

面的理解. $A=\{a_1, a_2, a_3, \cdots, a_n\}$ 表示所有事故案例集合, $I=\{i_1, i_2, i_3, \cdots, i_m\}$ 表示所有的二级指标项集合. 对于每一个事故案例 $a_n \in A$ 和指标项 $i_m \in I$, 其嵌入表示被定义为 $e_a \in R^d$ 和 $e_i \in R^d$ 其中 d 表示嵌入空间的维度. 通过嵌入函数 $BERT(\cdot)$ 将 a_n 和 i_m 映射到 R^d , 嵌入表示如下:

$$e_a = BERT(a), \quad e_i = BERT(i) \tag{1}$$

通过嵌入得到二级指标项文本 $E_{I_m} = \{e_{i_1}, \cdots, e_{i_m}\}$ 及三级指标项文本 $E_{I_{m,j}} = \{e_{i_{1,1}}, \cdots, e_{i_{1,j}}, \cdots, e_{i_{m,1}}, \cdots, e_{i_{m,j}}\}$ 的输入序列, 同时, 将事故报告文本也使用嵌入模型进行处理, 得到相应的文本嵌入表示 $E_A = \{e_{a_1}, \cdots, e_{a_n}\}$. 将二、三级指标项嵌入通过自注意力机制进行融合, 得到指标项的初始嵌入表示, 如图4(a)所示.

通过注意力机制得到二级指标项 I_m 与其对应的三级指标项 $I_{m,j}$ 的权重, 以反映每个三级指标项对其父级指标项的重要程度, 然后将添加权重后的嵌入进行拼接, 得到融合多阶指标项信息的指标项嵌入 $e_{i_m}^{(0)}$, 公式如下所示:

$$e_{i_m}^{(0)} = W_0 I_m + \sum_{i,j=1}^{i,j} W_i I_{m,j} \tag{2}$$

其中, W_0 和 W_i 分别表示二、三级指标项通过注意力

机制训练得到的权重, 初始嵌入向量嵌入如下:

$$E_I = [e_{i_1}, \dots, e_{i_m}] \quad (3)$$

我们由此得到了两部分的初始嵌入表示: 一个是指标项的嵌入表示 $e_{i_m}^{(0)}$, 另一个是事故报告的嵌入表示 $e_{a_n}^{(0)}$. 这些初始的嵌入反映了每个指标项和事故报告的语义信息. 在一个包含 N 个事故文本和 M 个融合指标项的文本数据中, 完整的嵌入矩阵被定义为:

$$E = [E_I | E_A] \quad (4)$$

2.3 结构捕获学习模块

事故的发生可能有更复杂的因素相互作用, 过去的方法通常使用描述性特征 (如 ID 和属性) 构建嵌入函数, 但它们无法充分挖掘事故-指标交互中潜在关联关系. 因而, 本文提出通过利用事故和指标项图结构中的高阶连通性来解决上述问题. 如图 6 所示, a 表示事故文本节点, i 表示指标项文本, 两者之间的交互构成了路径连接包含着丰富的语义信息, 这种高阶信息可由某一节点展开的树状结构表示.

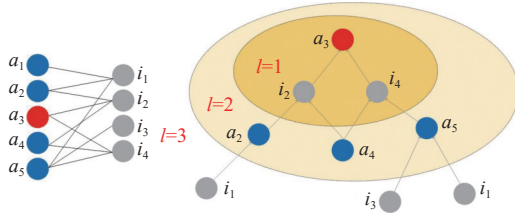


图6 高阶连通表示

如路径 $a_3 \leftarrow i_2 \leftarrow a_2$ 表示 a_3 和 a_2 之间的事事故相似性, 因为两个事故都与 i_2 指标项存在交互; 路径 $a_3 \leftarrow i_2 \leftarrow a_2 \leftarrow i_1$ 表示事故 a_3 很有可能与指标项 i_1 存在关系; 从 $l=3$ 的整体层次来看, 与指标项 i_3 相比, 指标项 i_1 可能与事故 a_3 的关系更大, 因为有两条路径连接 $\langle a_3, i_1 \rangle$, 只有一条路径连接 $\langle a_3, i_3 \rangle$.

基于上述研究, 本文以图卷积神经网络^[20]作为连通性捕捉模块, 如图 4(b) 和图 4(c) 所示. 首先将第 2.2 节中得到的指标项嵌入 $e_{i_m}^{(0)}$ 和事故报告文本的嵌入 $e_{a_n}^{(0)}$ 作为初始嵌入. 之后, 将其输入到图网络特征提取模块中进行处理, 通过叠加多个嵌入传播层, 来增强事故和指标项的嵌入表示, 从而有效地捕获图形中的高阶连通信息.

图网络特征提取模块传播规则定义如下:

$$e_a^{(k+1)} = \sum_{i \in N_a} \frac{1}{\sqrt{|N_a|} \sqrt{|N_i|}} e_i^{(k)} \quad (5)$$

其中, $e_i^{(k)}$ 表示邻居指标项节点 i 在第 k 层的嵌入向量, N_a 、 N_i 分别表示目标事故报告与指标项节点 a 、 i 的邻居节点数量. $e_a^{(k+1)}$ 表示事故报告节点 a 在第 $k+1$ 层的嵌入向量. 对于每个节点, 它的新嵌入是基于其邻居节点的嵌入的加权平均, 其中权重是由邻居节点的数量决定的, 这样可以保证信息在图中均匀传播, 避免了邻居数量多的节点对信息传播的影响过大. 那么将节点在每一层的嵌入向量进行聚合, 就可以得到一个综合了多层信息的全局嵌入, 以此作为事故案例及指标项的嵌入表示, 公式如下:

$$e_a = \sum_{k=0}^K \alpha_k e_a^{(k)}; \quad e_i = \sum_{k=0}^K \alpha_k e_i^{(k)} \quad (7)$$

其中, α_k 表示第 k 层卷积得到的嵌入表示在最终嵌入表示中所占的权重. 在本实验中我们设置 $\alpha_k = 1/K+1$, 其中 $K=3$. 之后, 将得到的节点嵌入表示输入到全连接层中, 并由激活函数将模型的输出结果映射为 1-5 分, 得分越高, 标记为事故根原因的概率就越大.

2.4 损失函数

本文采用常见的贝叶斯个性化损失 (Bayesian personalized ranking, BPR) 来优化事故与根原因之间的对应关系, 公式如下所示:

$$\mathcal{L}_{\text{BPR}} = - \sum_{(a \in A; i, h \in I)} \sigma(\hat{y}_{ai} - \hat{y}_{ah}) + \lambda \|E^{(0)}\|^2 \quad (8)$$

其中, a 为事故文本, i 为正样本指标项, h 为负样本指标项, $\hat{y}_{ai}$ 为正样本指标项与事故的预测得分, $\hat{y}_{ah}$ 为负样本指标项与事故的预测得分, 函数 λ 控制 L_2 正规化强度, $E^{(0)}$ 表示第 0 层的初始嵌入. BPR 损失函数通过 σ 函数转换得分差异, $(\hat{y}_{ai} - \hat{y}_{ah})$ 为一个概率, 这个概率表示模型预测事故文本 a 更倾向于正样本指标项 i 而非负样本指标项 h 的程度.

3 实验基础与结果

3.1 评价指标

在本实验中, 通过实验模型预测得到的是事故文本以及指标项之间的推荐评分情况. 为了模型预测的准确性, 我们使用了 $F1$ 、 $NDCG$ 、 mAP 这 3 个指标来评估模型性能.

(1) $F1$: 综合考虑准确率和召回率, 为两者的调和

平均值。1 表示最好的性能, 0 表示最差的性能, 定义为:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

(2) *NDCG*: 评估排序质量的指标。由于不同的事故根原因可能具有不同的重要性或严重性, 它考虑了每个原因的相关性和位置, 确保更重要的原因得到更高的权重。 rel_p 是排名第 p 的指标项的相关度, 而 rel_p^* 是理想情况下排名第 p 的指标项的相关度, 定义为:

$$DCG@k = \sum_{p=1}^k \frac{rel_p}{\log_2(p+1)} \quad (10)$$

$$IDCG = \sum_{p=1}^k \frac{rel_p^*}{\log_2(p+1)^*} \quad (11)$$

$$NDCG = \frac{DCG}{IDCG} \quad (12)$$

(3) *mAP*: 评估根原因推荐准确率的指标, 当存在多个可能的事故根原因时, 该指标帮助确定哪些原因更有可能导致事故发生, 确保关键原因能够优先得到考虑, 定义为:

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP(q) \quad (13)$$

其中, Q 是事故文本的总数, $AP(q)$ 为第 q 个查询的平均精度。

3.2 基准模型

在本节中, 为了充分地体现本文所提出的方法模型的优势以及更客观地对方法模型进行性能评估, 我们选取了若干传统模型和基于神经网络的推荐模型作为基线方法, 统一使用提供的数据集进行对比测试实验, 将实验结果与我们的模型进行对比。下面将对每个模型进行相应的介绍。

(1) 支持向量机 (support vector machine, SVM)^[21]: 在分类任务中以稳健性和效率著称的成熟机器学习算法, 为模型分类能力提供一个基准。

(2) 决策树 (decision tree, DT)^[22]: 用于分类和回归任务, 具有较好的简洁性和可解释性。

(3) LDA 主题模型^[23]: 该模型是文本建模和自然语言处理中流行的方法。

(4) 矩阵分解 (matrix factorization, MF)^[24]: 使用线性代数技术分解用户-项目评分矩阵, 广泛用于推荐系统以处理数据稀疏性。

(5) 图卷积神经网络 (graph convolutional neural network, GCNN)^[25]: 一种利用图结构的神经网络模型, 主要应用于节点分类和图分类任务中。

(6) 神经图协同过滤 (neural graph collaborative filtering, NGCF)^[26]: 融合了神经网络和图理论的推荐系统模型, 通过显式地建模用户和项目之间的交互图结构, 并提取这种结构的深层特征, 以此来提高推荐的准确性和可解释性。

(7) 解纠缠图协同过滤 (disentangled graph collaborative filtering, DGCF)^[27]: 通过分离用户和物品的多种潜在因素, 提高推荐系统的准确性和可解释性。

为了准确评估所提出方法的性能并与其他基准模型进行全面对比, 我们采取了一系列细致的数据准备、模型训练和预测评估步骤。首先确保训练数据集中包含了完整的事故报告及其对应的根原因标签, 从而允许所有参比模型能够有效学习到事故文本与根原因之间的潜在联系。针对测试数据集, 我们采取了一种模拟实际场景中信息不完整情况的方法, 即故意从事故报告中移除部分根原因。这一步骤旨在测试各模型在面对部分信息缺失的情境下, 对根原因的预测能力。在模型训练阶段, 所有模型均在完整的训练集上进行训练, 以确保它们能够在相同的学习条件下掌握事故报告与根原因间的关系。在预测与评估阶段, 我们利用已修改的测试集对各模型进行评估, 通过 *F1*、*NDCG* 和 *mAP* 等综合性性能指标来衡量它们在缺失根原因信息时的预测准确性。

3.3 结果分析

实验结果如表 1 所示, 我们的模型在各指标项方面均优于传统的模型, *F1* 的得分比传统统计分析提升约 10%, 相比于支持向量机和决策树模型也有约 5% 的提升。此外, 支持向量机和决策树模型相比于传统模型性能有所提升, 推测是因为两者都是监督学习模型, 它们能够从数据中学习复杂的模式和关系, 这使得它们在处理复杂的、非线性的问题时, 比基于规则或简单相关性的传统统计方法有更强的性能。

进一步分析发现, LDA 主题模型通过挖掘文本中的隐含主题来增强对事故根原因的理解, 提升模型的性能; 而图卷积神经网络等模型则利用其强大的图结构数据处理能力来捕捉事故数据中的复杂关系, 使得它在推荐事故根原因时相比其他模型表现更好。我们的模型不仅使用了图卷积神经网络, 同时还增加了多

阶指标项嵌入, 节点嵌入表示更加丰富, 使得模型性能进一步提升。

表1 与基准方法性能对比

方法	<i>F1</i>	<i>NDCG</i>	<i>mAP</i>
传统统计分析	0.6887	0.5633	0.8018
支持向量机	0.7115	0.5967	0.8326
决策树	0.7492	0.6223	0.8546
LDA主题模型	0.7522	0.6266	0.8658
矩阵分解	0.7509	0.6248	0.8673
图卷积神经网络	0.7545	0.6274	0.8739
神经图协同过滤	0.7661	0.6925	0.8812
解纠缠图协同过滤	0.7817	0.7364	0.8867
本文模型	0.7931	0.8892	0.9013

3.4 消融性分析

为了理解每个组件对推荐系统性能的贡献, 我们从4个方面进行了消融实验, 结果如表2所示。

表2 消融实验结果对比

消融部分	<i>NDCG</i>	<i>F1</i>	<i>mAP</i>	<i>p</i>
指标项三阶融合	0.6274	0.7545	0.8739	0.021
图卷积神经网络	0.6266	0.7522	0.8739	0.018
事故原因文本	0.5952	0.7193	0.8169	0.023
事故经过文本	0.5523	0.6886	0.7781	0.014
本文模型	0.8892	0.7931	0.9013	0.011

(1) 去除融合二阶和三阶指标项的文本嵌入部分, 仅使用二阶指标项的嵌入进行实验分析。在相同的数据集上分别训练并测试评估和比较两个模型的性能。

(2) 去除图卷积神经网络模块, 直接利用初始嵌入表示进行预测分析, 并与我们的模型进行对比, 以此评估图卷积神经网络在事故根原因识别中的作用和重要性。

(3) 去除事故报告中的“原因”部分文本描述, 以探究原因部分文本信息对模型预测准确度的影响。

(4) 去除事故报告的“经过”部分文本描述, 目的是评估事故经过部分文本信息对模型准确性的影响。

通过上述实验, 我们可以看出, 三阶指标项的融合、图卷积神经网络以及事故报告的“原因”和“经过”部分都是模型中至关重要的组成部分。每个组件都以其独特的方式优化了模型的性能, 提高了预测的准确性和全面性。

分析实验结果, 当从模型中去除对二阶和三阶指标项融合的嵌入表示后, 性能有所下降, 但仍然保持在一个相对较高的水平。这表明三阶指标项融合更多层级的指标项, 提供了额外的信息, 有助于捕获事故根原因的更复杂的层次结构; 当去除图卷积神经网络模块后, 性能有一定程度的下降, 可能是因为该模块能够利

用事故报告和指标项之间的图结构关系, 捕获更高阶的关联性, 从而提升模型性能; 当去除事故报告中“原因”部分后, 性能明显下降, 可能因为事故报告的“原因”部分直接关联到了事故的根本原因, 提供了关于事故发生原因的直接线索, 去除它会直接影响模型的预测能力; 去除“经过”部分后, 性能下降最为显著, 准确率、召回率和 *F1* 以及 *mAP* 都有大幅度的下降。推测原因是事故的经过描述包含了关键的上下文信息, 这些信息对于理解事故的全貌和确定事故根原因非常重要。例如, 事故的经过可能包括了导致事故的具体操作、环境条件或者序列事件, 这些细节对于准确预测根原因至关重要。

为了评估性能差异是否具有统计学意义, 我们也进行了显著性检验。实验结果显示, 考察的组件在显著性检验中均显示出显著性差异, 因为它们的 *p* 值都小于 0.05, 这表明去除这些组件都会导致模型性能的显著下降。事故报告的“经过”部分的去除导致了最显著的性能下降, 暗示这部分信息对于模型的预测准确性至关重要。相比之下, 二阶和三阶指标项的融合对模型性能的影响虽然也是统计显著的, 但相较于“经过”文本描述的去除, 其影响稍微温和。

3.5 超参数分析

本节分析图卷积神经网络层数对我们模型性能的影响, 我们系统地分析各层数图卷积神经网络对模型性能的影响情况, 综合实验结果见图7。

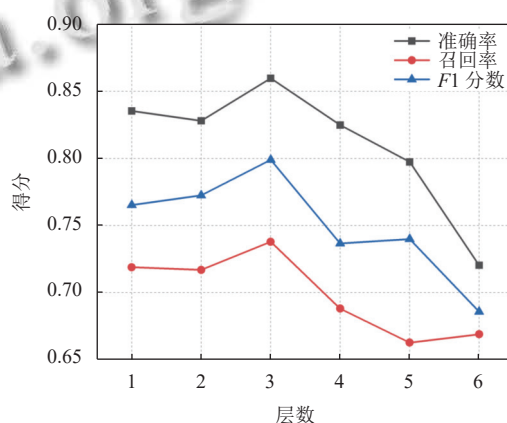


图7 图卷积神经网络层数性能分析

具体来说, 3层图卷积神经网络在捕捉化工事故数据中的复杂关系和细节方面表现出明显的优势, 较1层和2层图卷积神经网络显示了更强的特征提取能力。然而, 当网络层数增加超过3层时, 出现了性能下

降的情况,此现象主要归因于过平滑问题,这是深层图网络常见的挑战.在多层卷积过程中,节点的特征表示开始变得越来越相似,最终导致不同节点间的区分度降低,从而导致过平滑问题,影响了模型的预测效果.此外,化工事故数据的稀疏性也加剧了这一问题.由于多层图卷积过程中节点特征不断聚合其邻居特征,当层次增多时,有效信息的衰减速度会更快,从而加速了过平滑的发生.

3.6 实例分析

为了进一步验证本文提出的模型在实际应用中的有效性和鲁棒性,本节进行化工事故案例分析实验.在此实验中,我们需要首先确保训练集中的所有事故文本和对应的根原因标签是完整无缺的,以保证模型在学习过程中能够接触到全面的数据和信息.测试实例

如图8所示,2019年在内蒙古东兴化工发生的“4·24”爆炸事故,导致4人死亡,3人重伤,33人轻伤,直接经济损失4154万元.针对该事故案例,根据专家深度分析,事故的根原因已经较为准确地确定为6大方面:(1)安全教育培训不足;(2)安全设备安装使用不符合规定;(3)对历史事故教训的忽视;(4)安全风险辨识能力不足;(5)隐患排查治理不彻底;(6)应急处置措施不到位.过去传统方式对该事故的原始报告进行分析后,确定了以下可能的根原因:(1)操作员疏忽;(2)安全设备安装使用不符合规定;(3)对历史事故教训的忽视;(4)安全风险辨识能力不足;(5)气象环境条件的影响.综上,传统分析结果认为由于操作员的疏忽,以及设备不符合规定,加之外部大风环境因素,进而导致的爆炸事故发生.

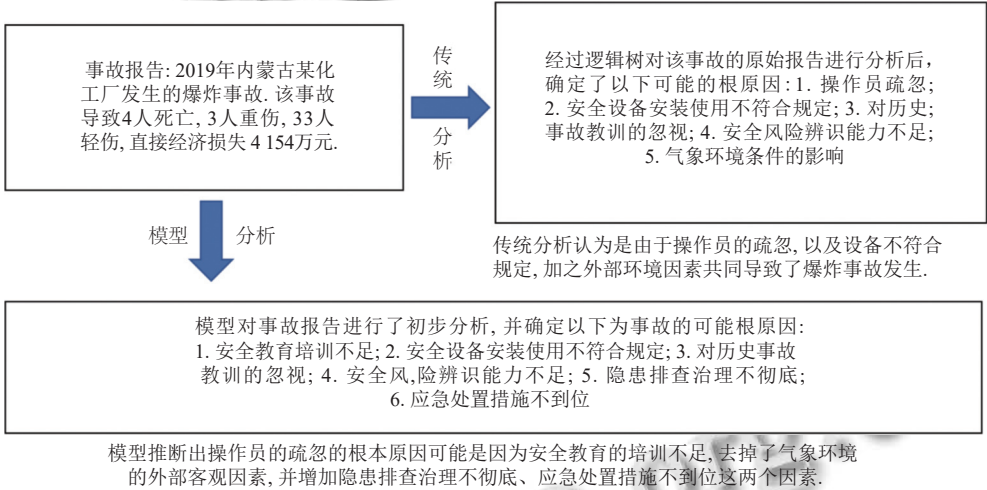


图8 事故案例分析

我们直接将该事故的完整报告输入模型进行分析.模型成功识别出了专家分析所确定的所有根原因,证明了本文方法在捕捉复杂事故背后的深层原因方面的有效性.两种方法相对比,所提模型额外捕捉到了“隐患排查治理不彻底”和“应急处置措施不到位”这两个关键因素,并将“操作员疏忽”这一根因修正为“安全教育培训不足”,并去掉了“气象环境条件的影响”这一客观因素.虽然传统的分析具有丰富的经验和专业知识,但在复杂的事事故境中,我们的模型可以提供一个全面和深入的视角,帮助揭示可能被忽视的事故原因.

为了进一步测试模型的发现和补全能力,我们故意从报告中隐藏了部分已知根原因.模型在这种信息缺失的情况下仍然成功地发现了大部分被隐藏的根原

因,进一步突显了其在处理不完整信息时的鲁棒性和准确性.此外,我们还对网络上收集的其他化工事故案例进行了类似的验证.模型不仅能够准确识别出已明确的根原因,而且还能在事故报告中故意遗漏的根原因测试中,成功地补全或发现这些关键因素.这一结果验证了模型在实际应用中的潜力,能够为化工安全领域提供一种高效、全面的事故根原因分析工具.

4 结论

(1) 本文提出了一种基于图卷积神经网络的事事故根原因分析方法,能够通过结合树形结构的指标项和二部图结构的数据,有效地挖掘化工事故报告与指标项之间的潜在关联关系.该方法在理解和分析事故根原

因方面相较于传统方法显示出更高的准确性和全面性。

(2) 通过一系列对比实验, 证明了我们提出的基于图卷积神经网络事故根原因分析方法在准确率、召回率和 $F1$ 等关键性能指标上优于传统事故分析方法。实验结果支持了我们所提出的方法在化工安全领域的事故根原因分析上的有效性和优势。

(3) 本文中的实例分析进一步验证了模型的有效性。模型不仅能够识别出已知的根原因, 还能发现在传统分析方法中可能被忽略的潜在原因, 如“安全教育培训不足”等, 这一发现突显了我们的模型在深入挖掘复杂化工事故根原因的优势。

参考文献

- 1 Wang JH, Fu G, Yan MW. Comparative analysis of two catastrophic hazardous chemical accidents in China. *Process Safety Progress*, 2020, 39(1): e12137. [doi: 10.1002/prs.12137]
- 2 王媛媛, 师立晨, 王如君, 等. 化工和煤矿行业安全管理对比分析与对策建议研究. *中国安全生产科学技术*, 2021, 17(S1): 172–176.
- 3 徐国飞. 关于国内化工企业安全评价方法探讨. *当代化工研究*, 2022(1): 10–12. [doi: 10.3969/j.issn.1672-8114.2022.01.004]
- 4 王延平, 卢均臣, 张婷, 等. 化工事故逻辑树法研究与应用. *安全、健康和环境*, 2015, 15(9): 1–4. [doi: 10.3969/j.issn.1672-7932.2015.09.001]
- 5 卜全民, 王涌涛, 汪德耀. 事故树分析法的应用研究. *西南石油大学学报*, 2007, 29(4): 41–144.
- 6 Kabir S. An overview of fault tree analysis and its application in model based dependability analysis. *Expert Systems with Applications*, 2017, 77: 114–135. [doi: 10.1016/j.eswa.2017.01.058]
- 7 邹吉然, 侯蕊, 朱炳旭. 故障树分析法在化工安全评价中的应用. *山东工业技术*, 2016(16): 46.
- 8 American Institute of Chemical Engineers. *Guidelines for Chemical Process Quantitative Risk Analysis*. 2nd ed., New York: American Institute of Chemical Engineers, 2000.
- 9 Ruijters E, Stoelinga M. Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools. *Computer Science Review*, 2015, 15–16: 29–62. [doi: 10.1016/j.cosrev.2015.03.001]
- 10 张静涛. 安全检查表法在危险化学品罐区安全评价中的应用. *中小企业管理与科技*, 2020(8): 162–163.
- 11 王婕, 林柏泉, 茹阿鹏. 因果图分析法应用于液氯泄漏事故分析. *安全与环境工程*, 2005, 12(3): 78–80, 90. [doi: 10.3969/j.issn.1671-1556.2005.03.023]
- 12 李艳昌, 刘哲含, 贾进章. 2006–2015 年我国化工企业生产事故统计分析. *应用化工*, 2017, 46(8): 1620–1623. [doi: 10.3969/j.issn.1671-3206.2017.08.042]
- 13 胡万吉. 2009–2018 年我国化工事故统计与分析. *今日消防*, 2019, 4(2): 3–7.
- 14 万婧, 容振乾, 赵远飞, 等. 2010–2019 年我国化工行业较大以上生产安全事故统计分析及其启示. *工业安全与环保*, 2021, 47(5): 59–63.
- 15 张圣柱, 王旭, 魏利军, 等. 2016–2020 年全国化工和危险化学品事故分析研究. *中国安全生产科学技术*, 2021, 17(10): 119–126.
- 16 马丽. 化工事故技术分析方法. *化工设计通讯*, 2018, 44(1): 235. [doi: 10.3969/j.issn.1003-6490.2018.01.214]
- 17 杨小峰. 化工企业生产事故分析及对策研究. *化工设计通讯*, 2022, 48(12): 160–162. [doi: 10.3969/j.issn.1003-6490.2022.12.054]
- 18 武妍, 王守觉. 一种通过反馈提高神经网络学习性能的新算法. *计算机研究与发展*, 2004, 41(9): 1488–1492.
- 19 陈国华, 安霆, 陈培珠. 危险化学品事故多米诺效应历史数据研究评述. *中国安全生产科学技术*, 2015, 11(4): 64–70. [doi: 10.11731/j.issn.1673-193X.2015.04.010]
- 20 He XN, Deng K, Wang X, et al. LightGCN: Simplifying and powering graph convolution network for recommendation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 2020. 639–648.
- 21 吴红, 王晓明, 王斌, 等. PSO-SVM 模型在化工事故风险预测中的应用研究. *工业安全与环保*, 2022, 48(4): 70–73.
- 22 杨静, 张楠男, 李建, 等. 决策树算法的研究与应用. *计算机技术与发展*, 2010, 20(2): 114–116, 120. [doi: 10.3969/j.issn.1673-629X.2010.02.030]
- 23 杨杰, 牛立杰. 基于 LDA 主题模型的建设安全管理研究. *建设科技*, 2022(24): 59–63.
- 24 Lekshmi Priya T, Sandhya H. Matrix factorization for recommendation system. *Proceedings of the 2019 International Conference on Artificial Intelligence and Data Engineering*. Singapore: Springer, 2019. 267–280.
- 25 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907*, 2016.
- 26 Wang X, He XN, Wang M, et al. Neural graph collaborative filtering. *Proceedings of the 42nd International ACM SIGIR conference on Research and Development in Information Retrieval*. New York: ACM, 2019. 165–174.
- 27 Wang X, Jin HY, Zhang A, et al. Disentangled graph collaborative filtering. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 2020. 1001–1010.

(校对责编: 张重毅)