文章编号：0253-374X(2025)04-0611-08

DOI：10.11908/j.issn.0253-374x.23307

交通事故致因知识图谱构建及风险因素挖掘

王占中，张书源，杨　萌，兰若冰，吴智豪

（吉林大学 交通学院，长春 130025）

摘要：利用交通事故调查报告中的数据，构建交通事故致因知识图谱并分析风险因素。首先，基于微调通用信息抽取统一框架预训练模型，构建适用于低数据量的交通事故致因命名实体识别模型，并生成实体集；其次，通过结构化处理和本体构建，利用图数据库Neo4j存储交通事故致因知识图谱，实现可视化；再次，基于专家经验和预训练语言文本分类模型，对交通事故致因实体进行标准化；最后，构建基于交通事故致因图谱的风险因素分析方法，通过分析标准化实体的类型分布和度分布，挖掘各因素对事故的触发特征与贡献，并进行关联规则挖掘。这些方法和分析结果提供了对历史事故风险因素的深入理解与探索。

关键词：交通运输；知识图谱；致因分析；数据挖掘；命名实体识别

中图分类号：U491.31　　**文献标志码**：A

Traffic Accident Causation Knowledge Graph Construction and Risk Factor Mining

WANG Zhanzhong, *ZHANG Shuyuan*, *YANG Meng*, *LAN Ruobing*, *WU Zhihao*

(Transportation College, Jilin University, Changchun 130025, China)

Abstract: In this paper, we use data from traffic accident investigation reports to construct a traffic accident causation knowledge graph and analyze risk factors. Firstly, we construct the recognition model of named entities of traffic accident causation applicable to low data volume based on the fine-tuned UIE pre-training model for the generation of the entity set. Secondly, through the structured processing and ontology construction, the graph database Neo4j is used to store the traffic accident causation knowledge graph for visualization. Thirdly, based on the expert experience and pre-trained language text classification model, the traffic accident causation entities are standardized. Finally, a risk factor analysis method based on the traffic accident causation graph is constructed to mine triggering characteristics and contributions of each factor by analyzing the type distribution and degree distribution of standardized entities, and to perform the association rule mining. The results of these methods and analyses provide an in-depth understanding and exploration of historical accident risk factors.

Keywords: transportation; knowledge graph; causal analysis; data mining; named entity recognition

交通事故在造成巨大经济损失的同时，已经严重威胁到了人们的生命安全。由此，《“十四五”全国道路安全规划》提出了将大数据、人工智能、5G等新技术充分应用于道路交通安全管理工作的现代化治理规划目标，针对交通的大数据智能技术成为当下解决问题的关键[1]。

知识图谱作为一种结构化、图形化的知识表达方式，已在医疗、金融、新闻等众多领域得到深入研究。Ernst等[2]构建了健康领域知识图谱“KnowLife”。Rotmensch等[3]基于电子病历构建了疾病与症状相关的临床决策知识图谱。Mohamed等[4]通过图嵌入技术为生物应用知识图谱赋予预测和分析能力。杨晓梅等[5]提出了基于遥感大数据的地学知识图谱计算框架。然而，在交通领域，知识图谱的应用仍处于起步阶段。传统交通事故风险分析方法依赖于系统性事件分析，存在专家数据获取困难、难以提取事故链条中隐藏因素等问题，且利用文本数据和知识图谱结构进行交通事故因素分析的研究较少。

以文本数据为源数据的事故图谱构建和风险关系抽取仍处于研究发展的初期。Ali等[6]提出了一种

收稿日期：2023-09-06
基金项目：吉林省自然科学基金面上项目(20230101112JC)
第一作者：王占中，教授，博士生导师，工学博士，主要研究方向为物流资源优化技术。E-mail：wangzz@jlu.edu.cn

论文拓展介绍

基于社交网络的实时监测框架,利用本体和潜在迪利克雷分配(OLDA)与双向长短期记忆网络(Bi-LSTM)进行交通事故检测和状态分析。国内在此研究领域发展相对较快。Liu等[7]通过文本挖掘与文本增强技术构建了铁路事故知识图谱,并对图谱进行了危险因素分析。贾熹滨等[8]基于新浪交通事故数据源提取风险因素,提出改进的多值属性先验(MA-Apriori)算法,挖掘事故发生的多种因素组合,并建立了贝叶斯网络交通事故风险预测模型。韩天园等[9]基于事故成因的社会网络中心分析,构建了重大交通事故成因网络模型。程宇航等[10]使用字典模式将文本数据分词,结合Word2vec和sigmoid激活函数构建交通事故词向量模型,获得事故特征与致因属性的关键词,并进行可视化分析。樊海玮等[11]利用双向编码器表示模型(BERT)对文本字符进行动态向量映射,并利用双向门控循环单元(BiGRU)提取向量化后的特征,显著提高了对交通事故文本数据提取的准确率。

目前,基于知识图谱的文本数据交通事故分析方法多集中于新闻数据,且研究主要关注事故因素及组合的提取。然而,基于文本数据的交通事故风险分析研究仍显不足,利用知识图谱进行事故因素分析的研究尚未深入展开。本文基于小批量交通事故调查报告数据,提出一种基于预训练交通事故命名实体识别和事故文本标准化方法的交通事故致因知识图谱构建及风险分析模型,能够充分利用历史专家经验和事故结构,为交通事故分析提供针对性方法。

1 交通事故命名实体识别模型

1.1 基于网络爬虫技术的交通事故调查报告文本数据获取

数据来源于安全管理网,通过车辆伤害事故栏目获取交通事故调查报告文本数据。网站提供文本格式和PDF格式的报告。利用Python爬虫爬取文本格式的报告,下载并以文本文档形式存储,共获取1 478篇小规模交通事故调查报告文本数据。

1.2 交通事故文本数据集标注

采用“BIOES”五元序列标注方法对事故文本数据进行人工标注,以满足监督式学习算法对大量标注数据的需求。BIOES标注使用B(开始)、I(内部)、O(非实体)、E(尾部)和S(单个实体)5种标签对命名实体序列进行标注。通过标注分析发现,交通事故实体具有以下特点:实体序列较长、形容词较少、泛化性较高。标注后的数据如图1所示,标注效果如图2所示。经过分割获得交通事故调查报告实体句共3 543条。

2	·	谭	路	金	驾	驶	的	车	辆	在	禁	停	路	段	违	法	停	车	。
O	O	O	O	O	O	O	O	O	O	B-ddes	I-ddes	I-ddes	I-ddes	I-ddes	I-ddes	I-ddes	I-ddes	E-ddes	O

图1 事故数据标注结构

Fig.1 Accident data annotation structure

2021年9月14日0时52分许(•time),惠安县涂寨镇惠崇线高铁桥下T型交叉路口(•loc)(S312线18KM+200M路段)发生一起重型半挂牵引挂车与普通二轮摩托车(•desc)相碰撞,造成一人死亡(•cons)的道路交通事故。依据《生产安全事故报告和调查》

图2 实体标注实例

Fig.2 Example of entity labeling

1.3 基于BiLSTM-CRF的交通事故实体识别模型

非预训练模型BiLSTM-CRF(bidirectional long short-term memory-conditional random field)由双向长短期记忆网络(BiLSTM)和条件随机场(CRF)组成。BiLSTM通过正向和逆向捕捉序列的上下文信息,而CRF用于标签之间依赖关系的处理,确保输出标签序列的全局最优性。

在BiLSTM输出层添加CRF层以维护交通实体标签转移矩阵,能够为模型增加先验知识与规则,提高对标签预测的准确率。通过输入训练集的标签序列CRF层可维护标签转移概率,转移概率通过维特比解码算法进行计算。

CRF层与BiLSTM输出层组合后的交通实体识别模型如图3所示。其中,w0表示句子中第一个词元,输出结果中B-ddes一行代表该词元应标记为B-ddes标签的概率,以此类推。

1.4 基于ERNIE-UIE的交通事故实体识别模型

Lu等[12]提出了一个基于文本到结构映射的通用信息抽取UIE模型,该模型基于ERNIE(enhanced representation through knowledge integration)架构。UIE模型通过构建text-to-structure映射能力、结构能力和语言能力进行预训练,能够满足多种下游任务,包括实体抽取、关系抽取、事件抽取、评论观点抽取、情感分类以及跨任务抽取等。

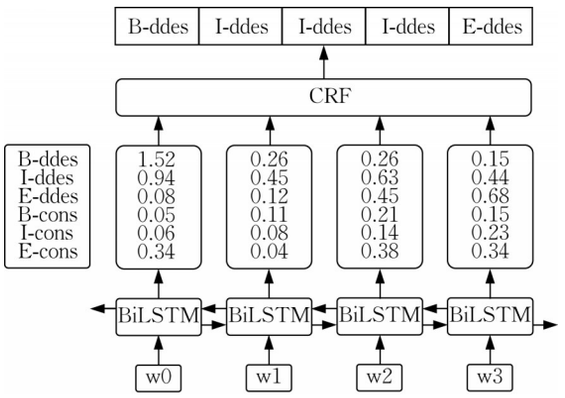

图3　交通事故实体BiLSTM-CRF模型

Fig.3　BiLSTM-CRF model of traffic accident entities

针对文本数据量小导致监督模型训练效果不佳的问题，提出一种基于预训练自然语言处理模型的交通事故相关实体提取方法。设计了UIE模型的微调方案及结构，构建了基于微调UIE模型的交通事故命名实体识别模型，并与BiLSTM-CRF模型进行对比，通过指标确定最优交通事故命名实体识别模型。

利用UIE-base基础模型进行未微调的交通事故实体抽取。确定实体提取schema，选定UIE任务流类别，并将数据输入预训练模型，输出基于预定义schema的文本片段及概率。实体提取流程如图4所示。

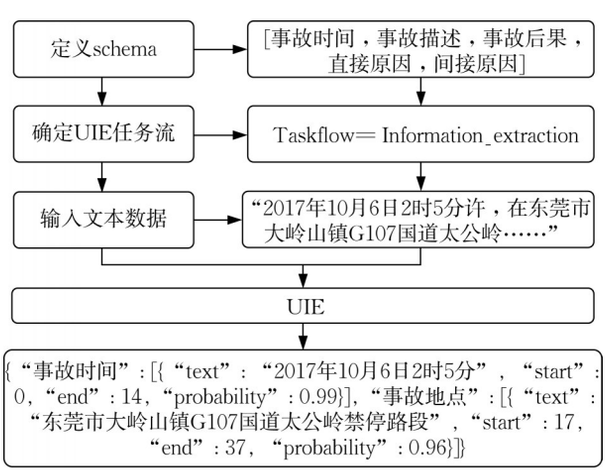

图4　基于UIE模型的交通事故实体提取流程

Fig.4　Traffic accident entity extraction process based on the UIE model

由于未微调的UIE模型可能未对交通事故领域进行训练，且其schema不符合该领域任务理解，直接使用会导致准确度较低。UIE模型支持小样本微调训练，可提升对自定义类别的识别效果。本文采用Doccano标注的425篇交通事故文本数据，按0.8∶0.2比例划分为训练集与测试集，通过调用UIE的微调接口进行训练，以获得更精确的交通事故实体识别模型。微调流程如图5所示。

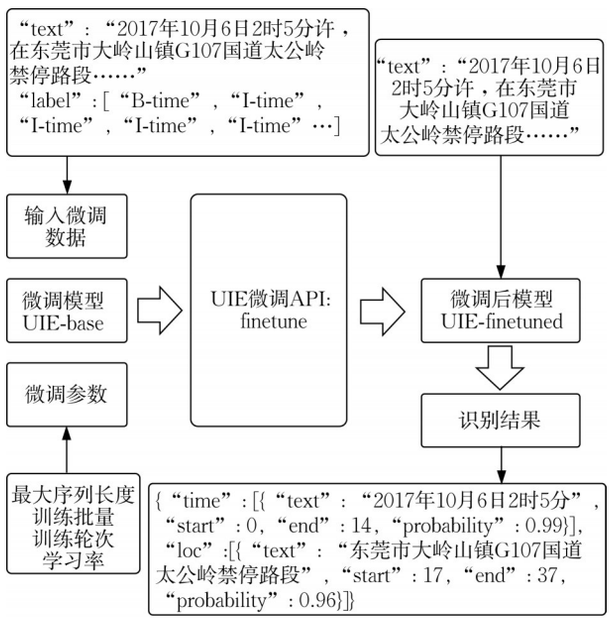

图5　UIE模型微调流程

Fig.5　Fine-tuning process of the UIE model

使用模型指标精确率、召回率、F1分数(F_1)，计算式分别为：

$$P=\frac{T_P}{T_P+F_P} \tag{1}$$

$$R=\frac{T_P}{T_P+F_N} \tag{2}$$

$$F_1=2\times\frac{PR}{P+R} \tag{3}$$

式中：T_P代表预测为正、实际也为正的样本数；F_P代表预测为负、实际为正的样本数；F_N代表预测为负、实际也为负的样本数。

分别对基于BiLSTM-CRF和基于微调的UIE预训练模型进行实验验证，2种命名实体识别模型各运行20个轮次(epoch)，最后得到各模型的指标如表1所示。

表1　模型结果

Tab.1　Results of models

模型	P	R	F_1
BiLSTM-CRF	0.245	0.213	0.225
小样本微调UIE	**0.718**	**0.693**	**0.705**
未微调UIE	0.353	0.301	0.324

微调后的UIE模型在各项评价指标上均优于其余2种模型，其性能足以支持交通事故实体提取任

务。然而,微调后的UIE模型对于sdes和ddes实体的预测效果较差,这可能是由于直接原因与间接原因的文字结构不固定、实体被分隔,或缺乏直接叙述性文字,导致识别结果与实际情况存在差异。尽管如此,微调后的UIE模型结果仍可作为交通事故致因知识图谱的准确数据源。

2 交通事故致因知识图谱可视化

2.1 交通事故本体构建

交通事故实体是交通事故报告中的结构化要素,而构建交通事故致因知识图谱还需要一个固定的交通事故实体关系结构来规范化交通事故事件结构。基于交通事故的事件结构与提取到的实体特征构建了交通事故本体模型,如图6所示。

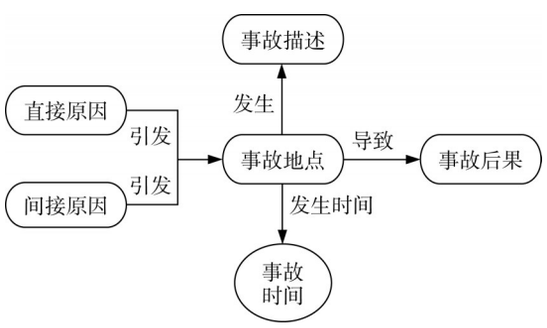

图6 交通事故本体模型

Fig.6 Traffic accident ontology model

事故地点在知识图谱中具有事故时间属性,是每个事故案例的唯一标识。直接原因、间接原因和事故描述被设为独立实体,在知识图谱中各自唯一存在,作为统一元素连接不同事故案例,以标示事故的同质因素。

2.2 交通事故致因知识图谱可视化

Neo4j是构建知识图谱常用的图数据库之一,本文采用Neo4j进行数据存储。设计了具有独立实体的交通事故实体可视化处理流程,创建了因素独立实体与案例实体,最终构建完成Neo4j知识图谱数据库。该数据库包含2 199个节点和2 821条关系,可视化效果如图7所示。

3 交通事故风险因素挖掘

3.1 交通事故实体标准化

通过对历史交通事故事件结构的分析,将交通事故描述、直接原因和间接原因实体划分为如表2~4所示的类别。同一类别的实体具有相似的表现形式,在处理和解决相关问题时也展现出相似性。

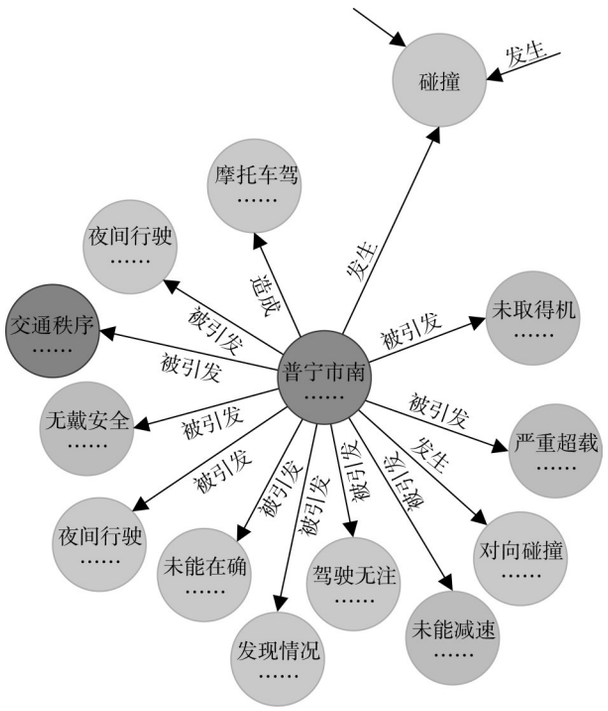

图7 交通事故致因知识图谱可视化

Fig.7 Visualization of traffic accident knowledge graph

表2 事故描述实体类别

Tab.2 Incident description entity category

事故描述类别	类别概述
车辆与非机动车事故	事故是机动车与非机动车之间发生的
多车辆事故	事故是机动车之间发生的
单车辆事故	事故是机动车与环境之间发生的

表3 事故直接原因实体类别

Tab.3 Incident direct cause entity category

事故直接原因类别	类别概述
超员超载	车辆违规超员或拉货超载
非法改装加挂	对机动车做了违法改装或非法牵引其他装置
无证或不符	无证驾驶或准驾车型不符
驾驶员操作不当	驾驶员操作失误或紧张或注意力不集中引发事故
未做安全措施	未系好安全带或带好头盔
超速	车辆超过规定速度行驶
疲劳驾驶	驾驶员疲劳驾驶
未遵守交通信号	闯红灯、跨禁线等无视交通信号标识
车辆隐患	车辆部件不符合安全行车需求
酒驾醉驾	驾驶员饮酒后驾驶机动车

对所有实体数据进行分类,将分类结果导入图数据库中,作为直接原因、间接原因、事故描述的属性以及与交通事故地点(交通事故案例唯一标识)的关系属性,完成对交通事故实体的标准化。标准化

表 4　事故间接原因实体类别

Tab.4　Incident indirect cause entity category

事故间接原因类别	类别概述
执法不严	管理行政主体对违规生产活动执法不严
安全培训不足	责任主体的义务安全培训不足
责任制度不健全	责任主体的安全管理制度存在漏洞
安全意识淡薄	责任主体人员安全意识淡薄
主体责任未落实	责任主体未落实安全生产主体责任
生产组织不到位	责任主体的生产规划与安全配置不足
违反有关规章制度	责任主体违反相关监察及安全生产规定

流程如图 8 所示。在实体分类结果中,所有的事故因素实体文本被替换为如表 2~4 所示的标准化实体类别。

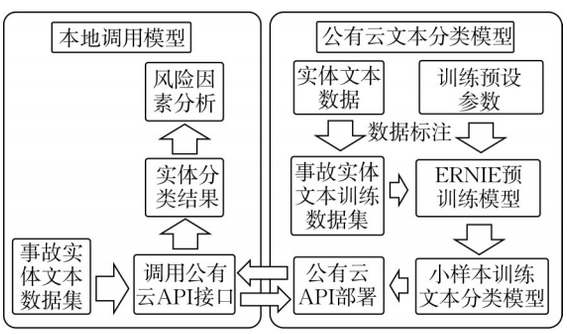

图 8　基于云端与本地端的交通事故实体标准化

Fig.8　Cloud-based and local traffic accident entity standardization

3.2　交通事故风险因素分析

对交通事故类型与直接原因的综合分析可揭示各直接风险因素对不同事故类型的影响。图 9 中的事故类型与直接原因度分布热力图显示,“驾驶员操作不当”是导致多车辆事故和单车辆事故的主要因素,而“超速”和“未遵守交通信号”则是多车辆事故的高发因素。

同样地,可以对间接原因与事故类型进行分析。如图 10 所示,“主体责任未落实”是引发事故的关键间接原因。“违反有关规章制度”和“生产组织不到位”是多车辆事故的次级诱因;“执法不严”对单车辆事故和多车辆事故的影响相当,“安全意识淡薄”则更易导致单车辆事故。

如图 11 所示,在所有搜集的案例中,“主体责任未落实”是最主要的间接原因。与该间接原因相关的直接原因包括“超员超载”“驾驶员操作不当”“超速”“未遵守交通信号”,在事故关联中占据较高比重。

3.3　交通事故关联规则挖掘

通过从交通事故致因知识图谱中提取的事故数据及相关因素构建交通事故因素记录集,并使用

图 9　事故类型与直接原因度分布热力图

Fig.9　Heat map of accident type and direct cause degree distribution

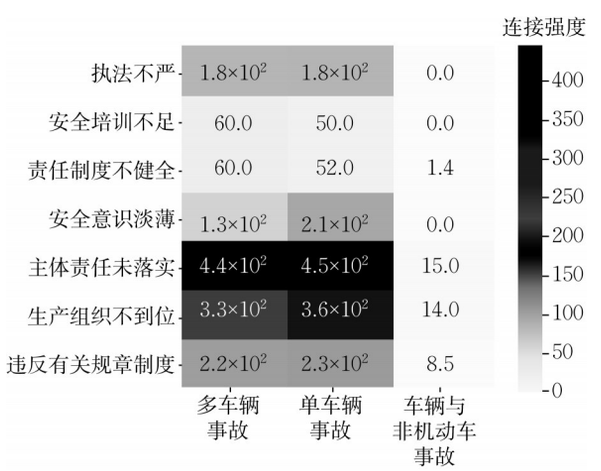

图 10　事故类型与间接原因度分布热力图

Fig.10　Heat map of accident type and indirect cause degree distribution

Apriori 和频繁模式树(FP-Growth)算法进行关联规则挖掘,可以快速有效地分析事故因素之间的关联规则。关联规则挖掘流程如图 12 所示。

在对事故记录进行挖掘分析时,采用 Apriori、FP-Growth 及暴力遍历算法,并设置支持度比例阈值为 0.1、0.2、0.3、0.4。结果表明,3 种算法在挖掘频繁项集的数量上完全一致。进一步分析时间与空间成本,运行 5 次后取平均值,暴力遍历算法的存储空间成本最高,平均空间占用为 34.64 MB;Apriori 算法优化了存储成本,平均空间占用为 29.58 MB,但时间成本有所增加,平均时间为 0.067 2 s;FP-Growth 算法在时间和空间上均表现更优,平均时间为 0.056 0 s,平均空间占用为 29.24 MB。FP-Growth 算法的独特数据结构使其时间复杂度较低,

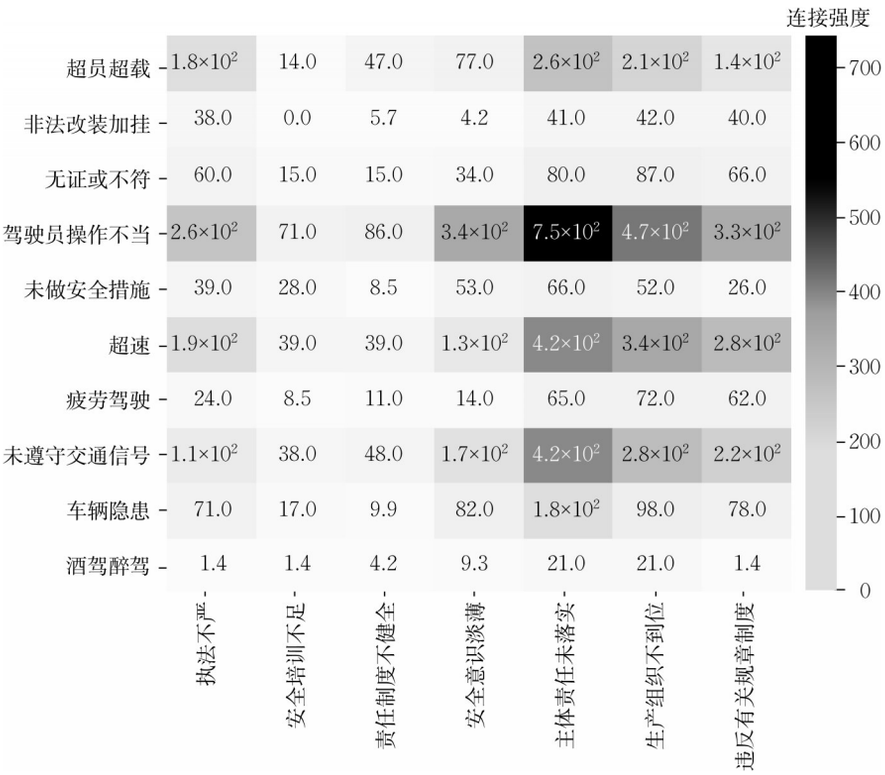

图11 直接原因与间接原因度分布热力图

Fig.11 Heat map of the degree distribution of direct and indirect causes

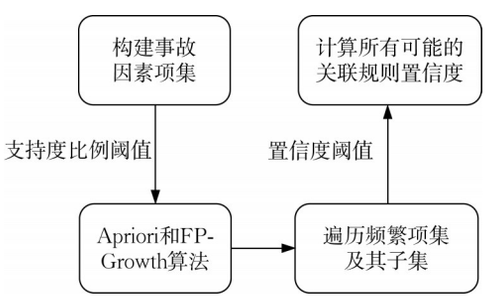

图12 关联规则挖掘流程

Fig.12 Association rule mining process

更适合处理大数据集。相比之下,暴力遍历算法时间复杂度为$O(MN)$,数据量增加时时间成本迅速膨胀,难以应对大数据环境。

支持度比例阈值为0.4时,Apriori与FP-Growth算法挖掘出的频繁项集均为1项集,无法提高支持度,且无法计算关联规则置信度,挖掘结果如表5所示。支持度比例阈值为0.3时挖掘出的频繁项集如表6所示。

表5 支持度比例阈值为0.4时的频繁项集

Tab.5 Frequent term set at a support proportion threshold of 0.4

项集	支持度
单车辆事故	719
超速	580
驾驶员操作不当	861

表6 支持度比例阈值为0.3时的频繁项集

Tab.6 Frequent term set at a support proportion threshold of 0.3

项集	支持度
单车辆事故	719
超速	580
主体责任未落实	503
未遵守交通信号	550
多车辆事故	432
驾驶员操作不当	861
{单车辆事故,驾驶员操作不当}	437

通过支持度比例阈值计算得到的交通事故因素频繁项集,能够进一步计算交通事故因素之间的关联关系。关联关系置信度计算式为

$$C(X\rightarrow Y)=\frac{\delta(X\cap Y)}{\delta(X)} \tag{4}$$

式中:$\delta(X\cap Y)$代表项目X与Y同时出现的次数;$\delta(X)$代表项目X单独出现的次数。

关联规则可用于因素间触发关系的表达,但需依据交通事故调查报告进行约束。如图13所示,所有触发关系需符合3级触发规则,包括间接原因自触发、间接原因触发直接原因、直接原因自触发、直接原因触发事故类型。基于此规则,对数据库中关

联规则进行筛选并计算置信度。

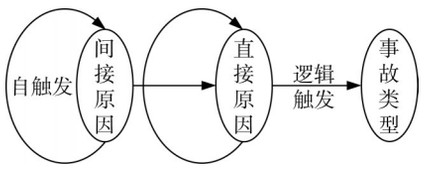

图 13　触发关联规则

Fig.13　Triggering association rules

支持度比例阈值为0.3、置信度阈值为0.5时事故原因关联关系计算结果如表7所示。

表 7　支持度比例阈值0.3和置信度阈值0.5时的关联规则

Tab.7　Association rules under a support proportion threshold of 0.3 and a confidence threshold of 0.5

关联关系	置信度
[驾驶员操作不当]→[单车辆事故]	0.507

上述关联关系表明,单车辆事故和驾驶员操作不当相关事故的发生概率超过30%;在触发规则约束下,驾驶员操作不当导致单车辆事故的概率接近50%。

当支持度比例阈值为0.2时,计算出的频繁项集数量较少;当置信度阈值为0.5时,所有关联规则均满足该条件,降低置信度后不会产生新的关联规则。支持度比例阈值为0.2、置信度阈值为0.2时事故因素关联关系计算结果如表8所示。

表 8　支持度比例阈值0.2和置信度阈值0.2时的关联规则

Tab.8　Association rules under a support proportion threshold of 0.2 and a confidence threshold of 0.2

关联关系	置信度
[主体责任未落实]→[驾驶员操作不当]	0.658
[超速]→[驾驶员操作不当]	0.608
[驾驶员操作不当]→[单车辆事故]	0.507

除去重复关联规则后,在该阈值组下得到的关联规则如下:

[主体责任未落实]→[驾驶员操作不当],表示企业或政府未落实主体责任的间接原因有20.0%的概率触发驾驶员操作不当,且主体责任未落实时,驾驶员操作不当的发生概率为65.8%。[超速]→[驾驶员操作不当]表示超速状态下2种直接原因共同导致交通事故的发生概率为20.0%,且在超速状态下驾驶员操作不当的概率为60.8%。

降低置信度能够挖掘出因素之间更低概率的触发关系,以支持度比例阈值0.1、置信度阈值0.2的阈值组进行触发关联规则挖掘,如表9所示。

表 9　支持度比例阈值0.1和置信度阈值0.2时的关联规则

Tab.9　Association rules under a support proportion threshold of 0.1 and a confidence threshold of 0.2

关联关系	置信度
[安全意识淡薄]→[驾驶员操作不当]	0.771
[车辆隐患]→[驾驶员操作不当]	0.618
[违反有关规章制度]→[驾驶员操作不当]	0.593
[驾驶员操作不当]→[单车辆事故]	0.507
[主体责任未落实,驾驶员操作不当]→[单车辆事故]	0.513
[超速,驾驶员操作不当]→[单车辆事故]	0.470
[未遵守交通信号,驾驶员操作不当]→[单车辆事故]	0.532

在该阈值组下得到的关联规则如下:

[主体责任未落实,驾驶员操作不当]→[单车辆事故]表示当企业或政府未落实主体责任的间接原因与驾驶员操作不当的直接原因同时出现时,触发单车辆事故的概率为51.3%;在所有历史事故中,超过10.0%的事故包含这3种要素。

[超速,驾驶员操作不当]→[单车辆事故]表示在超速状态下驾驶员操作不当同时出现时,发生单车辆事故的概率为47.0%。

通过调整支持度比例阈值,管理者可快速查询历史事故中某因素或因素组合的占比;通过调整置信度阈值,可基于现有因素预测可能触发的下级因素或事故类型,或反推未查明的隐含因素。此类隐含关系难以通过其他分析方法发现,且基于历史关联规则的推断能为未来事故分析提供经验补充,弥补风险因素分析的不足。

4　结 语

针对交通事故报告内容复杂的问题,提出基于命名实体识别的提取模型,通过对比选择最佳模型提取事故报告中的关键信息。基于事故事件结构与实体数据,构建交通事故本体模型,并利用Neo4j图数据库通过实体检索与关联完成交通事故致因知识图谱及可视化。为了解决数据来源差异,提出基于ERNIE2.0的交通事故致因实体标准化模型,实现图谱实体节点的分类与标准化。最后,提出基于图谱节点关系的事故风险因素分析方法,通过统计分析挖掘事故因素间的关联规则,揭示事故与风险因素的规律,辅助风险问题的分析。

本文模型效果有待进一步提升,实体分类与风险分析方法亦需在未来不断完善,但基于交通事故

致因知识的结构化表达与风险因素分析提出了新的技术路径。通过结合事故历史经验与事理逻辑结构,深入挖掘专家知识与潜在风险关系,为理解和预防交通事故提供了有力支持。

作者贡献声明:

王占中:研究思路和方法提出,研究结果分析,结论总结,论文修改。

张书源:构建和完善研究思路和方法,构建模型,分析处理实验结果,撰写论文。

杨 萌:采集事故文本数据集,构建实体标注数据集。

兰若冰:组织相关文献分析,提出研究方向和思路。

吴智豪:分析交通事故致因实验结果,绘制实验流程图表。

参考文献:

[1] 李佳芯. 以人为本 夯实基础 推进"十四五"道路交通管理科技发展:访公安部道路交通安全研究中心主任、中国道路交通安全协会会长王长君研究员[J]. 道路交通管理, 2021(3): 28.
LI Jiaxin. People-oriented, compact foundation, promoting the development of road traffic management science and technology in the 14th Five-Year Plan: interview with Researcher Wang Changjun, Director of the Road Traffic Safety Research Center of the Ministry of Public Security and President of the China Road Traffic Safety Association [J]. Road Traffic Management, 2021(3): 28.

[2] ERNST P, SIU A, WEIKUM G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences [J]. BMC Bioinformatics, 2015, 16(1): 157.

[3] ROTMENSCH M, HALPERN Y, TLIMAT A, *et al*. Learning a health knowledge graph from electronic medical records[J]. Scientific Reports, 2017, 7(1): 5994.

[4] MOHAMED S K, NOUNU A, NOVÁČEK V. Biological applications of knowledge graph embedding models [J]. Briefings in Bioinformatics, 2021, 22(2): 1679.

[5] 杨晓梅, 王志华, 刘岳明,等. 遥感智能信息处理的发展及技术前景[J]. 同济大学学报(自然科学版), 2023, 51(7): 1025.
YANG Xiaomei, WANG Zhihua, LIU Yueming, *et al*. Development and technical prospect of remote sensing intelligent information processing [J]. Journal of Tongji University (Natural Science), 2023, 51(7): 1025.

[6] ALI F, ALI A, IMRAN M, *et al*. Traffic accident detection and condition analysis based on social networking data [J]. Accident Analysis & Prevention, 2021, 151: 105973.

[7] LIU C, YANG S. Using text mining to establish knowledge graph from accident/incident reports in risk assessment [J]. Expert Systems with Applications, 2022, 207: 117991.

[8] 贾熹滨, 叶颖婕, 陈军成. 基于关联规则的交通事故影响因素的挖掘[J]. 计算机科学, 2018, 45(S1): 447.
JIA Xibin, YE Yingjie, CHEN Juncheng. Influence factors mining of traffic accidents based on association rules [J]. Computer Science, 2018, 45(S1): 447.

[9] 韩天园, 田顺, 吕凯光,等. 基于文本挖掘的重特大交通事故成因网络分析[J]. 中国安全科学学报, 2021, 31(9): 150.
HAN Tianyuan, TIAN Shun, LÜ Kaiguang, *et al*. Network analysis on causes for serious traffic accidents based on text mining [J]. China Safety Science Journal, 2021, 31(9): 150.

[10] 程宇航, 张健钦, 李江川,等. 交通行业事故文本数据的可视化挖掘分析方法[J]. 计算机工程与应用, 2021, 57(21): 116.
CHENG Yuhang, ZHANG Jianqin, LI Jiangchuan, *et al*. Visual mining and analysis method of text data in traffic accident [J]. Computer Engineering and Applications, 2021, 57(21): 116.

[11] 樊海玮, 秦佳杰, 孙欢,等. 基于BERT与BiGRU-CRF的交通事故文本信息提取模型[J]. 计算机与现代化, 2022(5): 10.
FAN Haiwei, QIN Jiajie, SUN Huan, *et al*. Traffic accident text information extraction model based on BERT and BiGRU-CRF fusion [J]. Computer and Modernization, 2022(5): 10.

[12] LU Yaojie, LIU Qing, DAI D, *et al*. Unified structure generation for universal information extraction [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational Linguistics, 2022:5755-5772.