

基于道路交通事故领域的知识图谱构建及可视化

张丽岩¹ 王 燕¹ 马 健¹ 谢雨晴¹ 周欢生²

ZHANG Liyan WANG Yan MA Jian XIE Yuqing ZHOU Huansheng

摘要

近年来,知识图谱技术进展迅速,领域知识图谱价值得到认可,各领域专家就本领域知识图谱的构建展开了广泛研究,并成功将其应用于知识检索、知识问答、知识推荐等服务中。以道路交通事故数据为主要数据源,经过概念模型设计、知识抽取等步骤提取出相关信息,与原结构化数据融合,借助 Neo4j 图数据库,构建融合道路交通事故数据和道路法规的道路交通事故领域的知识图谱,实现相关数据的存储和可视化。

关键词

知识图谱; 交通事故; 知识抽取; Neo4j; 可视化

doi: 10.3969/j.issn.1672-9528.2024.01.008

0 引言

随着我国经济社会的快速发展,道路交通设施建设步伐明显加快、机动车保有量持续快速增长,这给道路交通秩序与安全带来重大挑战,严重威胁人们的生命安全。根据《中国统计年鉴—2021》,2020 年我国发生 244 674 起道路交通事故,交通事故致死率高达 19.75%,这说明我国道路交通安全形势仍然十分严峻^[1]。

通过构建道路交通事故领域知识图谱,让道路交通事故领域内结构化、半结构化和非结构化等分散冗余的数据处理成统一的结构化数据。与传统的数据分析方法相比,图结构更能增强庞杂数据之间的关联性,同时通过图数据库查询语句对道路交通事故知识进行可视化查询和分析。

1 知识图谱架构

知识图谱在架构上可以分为逻辑架构和技术架构。

1.1 逻辑架构

知识图谱在逻辑上主要划分为数据层和模式层^[2]。模式层在数据层之上,包括概念模式和关系模式的设计,是知识图谱构建的核心,是对现实事物的概念以及关系进行抽象化映射。数据层存储真实的数据,是由一系列真实的道路交通

事故领域信息构成的,每个事实最终以三元组的方式进行表示,并借助 Neo4j 图数据库实现相关数据的存储。

1.2 技术架构

从技术层来看,知识图谱涉及到的相关技术包括实体抽取、关系抽取、实体消歧、信息存储、知识查询、智能问答等,技术框架如图 1。

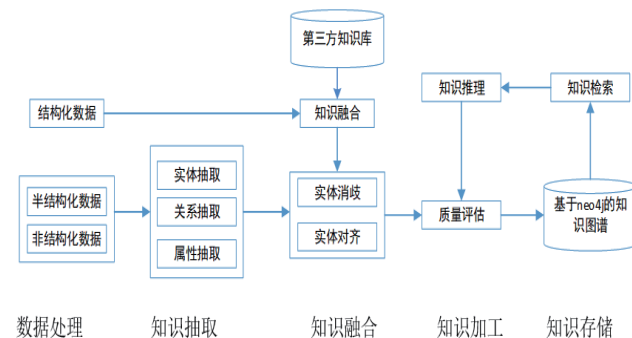


图 1 知识图谱技术框架

2 道路交通事故领域知识图谱构建

2.1 数据来源及构建思路

本文以某市 2021 道路交通事故数据为研究对象,该数据是由相关部门提供的实例数据,数据的准确性及真实性较高,价值密度较大。共有 6176 条交通事故数据,其中有 3583 条选取了由表格数据记录的数据,另外选取了 2593 条经典道路交通事故的交通事故责任认定书。通过对交通事故的数据进行整理,从中归纳出 9 种实体、8 种关系。

为进一步拓展图谱的适用性,将 ZG122 公众号发布的基于 GA/T 标准的交通违法行为代码,包括违法行为的法律规范、处罚依据、处罚种类和法定幅度等作为数据源,构建融

1. 苏州科技大学土木工程学院 江苏苏州 215011

2. 江苏江南商贸集团有限责任公司 江苏苏州 215500

[基金项目] 江苏省建设系统项目(2020ZD14、2018ZD258) 江苏省自然科学基金项目(BK20151201、BK20160357); 江苏省高校哲学·社会科学项目(2018SJA1348); 江苏省研究生科研与实践创新计划项目(SJCX20_1117); 江苏高校哲学社会科学研究一般项目(2023SJYB1420)

式中： A 为网络层输出的转移分数矩阵；元素 A_{ij} 代表标签从 i 转移为标签 j 的转移分数； P 为得分矩阵；元素 P_{i,y_i} 代表文本中第 i 个字符的 y_i 标签的得分； n 为文本字符数。CRF 模型计算标签序列的概率分布具体实现为：

$$P(Y | X) = \frac{\exp(\text{score}(x, y))}{\sum_{y'} \exp(\text{score}(x, y'))}$$

(2)

2.4 实验与结果分析

实验过程中将 BiLSTM 模型、BiLSTM-CRF 模型、BERT-BiLSTM-CRF 模型进行了对比，为检验模型优劣，本文使用精确率 P 、召回率 R 和 F_1 值进行评估。其中 TP 表示真实值与预测值均为真， TN 表示真实值与预测值均为假， FP 表示真实值为假、预测值为真， FN 表示真实值为真、预测值为假^[7]。 F_1 值具有平均精确率和召回率的得分优势，故命名实体识别任务常用上述评价指标。计算公式为：

$$P = \frac{TP}{TP + FP}$$

(3)

$$R = \frac{TP}{TP + FN}$$

(4)

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

(5)

通过对比表 2 中可以看出使用组合模型的效果优于使用单一模型的效果，模型的 P 值、 R 值、 F_1 值，都随着模型的改变而逐渐升高。其中 BERT-BiLSTM-CRF 模型的三个评价指标都是所有模型中值最高的，分别达到 95.13 %、95.74 %、95.43 %，所以本文选取该模型做道路交通事故领域的命名实体识别。

表 2 三种不同模型评价指标

模型	准确率 P	召回率 R	F_1 值
BiLSTM	0.939 5	0.946 9	0.940 6
BiLSTM-CRF	0.946 9	0.944 9	0.945 9
BERT-BiLSTM-CRF	0.951 3	0.957 4	0.954 3

2.5 数据细分与扩充

基于上述方法首先抽取到的交通事故的发生地点、时间、交通方式、成因、形态等关键信息，能达到非结构化到结构化数据的预处理效果，然后将其与原有的表格数据融合，将部分数据进一步扩充与细分，细分与扩充后的数据示例如表 3 所示。

表 3 细分与扩充后的数据

事故要素	原提取数据	细分与扩充后新增数据
事故时间	2021 年 1 月 1 日 10 点 23 分	1 月，星期五，10:00—11:00
事故地点	方洲路南施街交叉路口处	四支分叉口
当事人	410526198608082029	31 ~ 40 岁、女

3 基于 Neo4j 道路交通事故领域的知识图谱存储及可视化

Neo4j 是一个可以在网络上而非表中存储结构化数据的高性能 No SQL 图数据库^[8]。其依赖自身特性及数据格式，具有伸缩性和灵活性的特点，设计道路交通事故领域数据库时能随时对节点和边进行更新、查询、增删等操作，不会影响原数据的正常使用，且图搜索模式能快速检索和可视化反馈所需知识，因此可利用其构造道路交通事故领域知识图谱。

如图 5 所示是该知识图谱的部分示意图。图中同一实体对应的节点是统一的样式，如事故 ID 实体对应的节点颜色为红色，事故位置实体对应的节点颜色为黄色，这两实体对应的关系 Accident_Location 节点颜色为橘色。



图 5 部分道路交通事故领域知识图谱示意图

3.1 事故画像查询

在道路交通事故领域知识图谱中，基于领域知识图谱的交通事故画像，可以描述发生事故时间、事故时间、事故时间、天气、形态、责任判定等，以某一事故画像为例，如图 5 所示。2021 年 1 月 1 日 14 时 20 分，在星万盛街与钟园路交叉路口处发生的一起交通事故中，因当事人陈某转弯时未让直行的车辆先行的行为，承担本次事故的全部责任，当事人乔某无导致该事故发生的过错行为，不承担此次事故的责任。从当事人的属性中，可以看到当事人年龄、性别、身份证等。

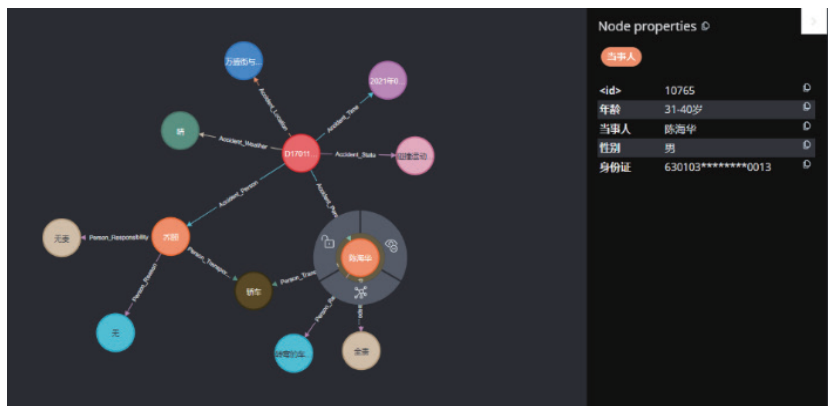


图6 基于交通事故知识图谱的某事故画像

4 结语

通过规范整理道路交通事故数据,分析数据结构化程度和语言特点,采用了基于BERT-BiLSTM-CRF和规则两种方法进行实体抽取、属性抽取,将相关非结构化知识转化为结构化知识,并与原有的结构化数据融合。利用Neo4j构建融合道路交通事故数据和道路法规的知识库,进行交通事故相关数据可视化分析;道路法规相关知识检索,帮助驾驶员提高安全意识和驾驶技能;拓展图谱的适用性,打破数据壁垒,实现数据共享。

3.2 事故统计查询

基于Neo4j的Cypher查询语言,可以根据需要筛选需要的知识数据,从而发现语言学术语的某些发展规律。通过Cypher语言查询,输入Cypher命令: MATCH (n:`事故时间`) RETURN n.时间段 AS 事故发生时间段, COUNT(*) AS 事故起数 ORDER BY 事故起数 DESC, 可以呈现出每个时间段发生的事故起数。MATCH (n:`事故时间`) RETURN n.时间段 AS 事故发生时间段, COUNT(*) AS 事故起数 ORDER BY 事故起数 DESC, 可以呈现出按照星期几分布的事故发生起数。数据以表格的形式显示返回结果,如图7所示。

```
experience$ MATCH (n:`事故时间`) RETURN n.时间段 AS 事故发生时间段, COUNT(*) AS 事故起数 ORDER BY 事故起数 DESC
```

事故时间段	事故起数
"08:00-09:00"	386
"18:00-19:00"	262
"09:00-10:00"	261
"17:00-18:00"	250
"07:00-08:00"	232
"16:00-17:00"	203

(a) 按时间段分布

```
experience$ MATCH (n:`事故时间`) RETURN n.星期 AS 事故发生星期, COUNT(*) AS 事故起数 ORDER BY 事故起数 DESC
```

事故发生星期	事故起数
"星期五"	497
"星期六"	474
"星期三"	469
"星期二"	459
"星期四"	455
"星期日"	450

(b) 按星期分布

图7 事故按时间段分布查询结果

参考文献:

- [1] 田振中,孙振雷.我国道路交通事故死亡人数影响因素及管理对策研究[J].中国人民公安大学学报(自然科学版),2022,28(2):38-44.
- [2] 姚剑,白伟,杨国元,等.铁路客站设备健康管理知识图谱构建与应用研究[J].铁道运输与经济,2023,45(3):94-102.
- [3] 罗丽.地铁运营安全事故知识图谱的构建及分析[D].徐州:中国矿业大学,2023.
- [4] 李金润.面向飞机维修实体关系构建与故障推理算法的研究与应用[D].成都:电子科技大学,2023.
- [5] 唐进君,庾昊南,刘佑,等.基于BERT-Bi-LSTM-CRF模型的自主式交通系统参与主体识别方法[J].交通信息与安全,2022,40(5):80-90.
- [6] 宫法明,董文吉,袁向兵.基于知识图谱的潜油电泵井故障诊断[J].计算机系统应用,2023,32(5):87-96.
- [7] 李韧,李童,杨建喜,等.基于Transformer-BiLSTM-CRF的桥梁检测领域命名实体识别[J].中文信息学报,2021,35(4):83-91.
- [8] 肖飞龙,张爽,胡志凌.基于Neo4j的疾病预防与控制知识图谱研究[J].电子技术与软件工程,2021(22):180-182.

【作者简介】

张丽岩(1978—),女,黑龙江齐齐哈尔人,博士,高级实验师,研究方向:交通大数据、交通规划。

王燕(1998—),女,江苏泰州人,硕士研究生,研究方向:交通大数据、交通规划。

马健(1979—),通信作者,男,江苏扬州人,博士,副教授,研究方向:交通大数据、交通仿真与控制。

谢雨晴(1998—),女,安徽淮南人,硕士研究生,研究方向:交通大数据。

周欢生(1983—),男,江西吉水人,学士,副部长,研究方向:数字政府、交通大数据。

(收稿日期:2023-10-07)