doi:10. 3969/j.issn. 1672-4623.2023.01.006

基于轨迹大数据的司机出行知识图谱构建与实现

李　璇[1]，吴　雷[1,2]，姜　梦[1]，潘　晓[1*]
（1. 石家庄铁道大学，河北 石家庄 050043；2. 燕山大学 信息科学与工程学院，河北 秦皇岛 066004）

摘　要：蕴含静态路网特性和动态交通特性的轨迹大数据具有多源异构的特点，给交通知识分析和挖掘带来了困难。近年兴起的知识图谱可对多源异构数据进行有效融合、对齐和加工。基于轨迹大数据，采用自底向上的方法在结构化数据中构建司机出行知识图谱，还原轨迹数据的时空关系和语义关联。具体来说，通过RDF Mapping进行知识抽取，将结构化数据映射为RDF数据；再利用Neo4j存储对静态路网拓扑进行知识补齐和融合。司机出行知识图谱的构建为交通出行知识查询和推理奠定了良好基础；同时，从结构化数据中自底而上构建知识图谱的方法可被扩展应用于其他领域的结构化数据，对推广知识图谱应用起到推进作用。

关键词：知识图谱；轨迹大数据；图数据库；RDB2RDF
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2023）01-0032-04

Construction and Implementation of Driver Travel Knowledge Graph Based on Trajectory Big Data

LI Xuan[1], WU Lei[1,2], JIANG Meng[1], PAN Xiao[1]
(1. Shijiazhuang Tiedao University, Shijiazhuang 050043, China;
2. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

Abstract: The trajectory big data containing static road network characteristics and dynamic traffic characteristics has the characteristics of multi-source heterogeneous, which brings some difficulties to traffic knowledge analysis and mining. In recent years, knowledge graph is an effective means to fuse, align and process multi-source heterogeneous data. Based on the trajectory big data, we realized the construction of bottom-up driver travel knowledge graph from the structured data, and restored the spatio-temporal relationship and semantic association in the trajectory data. Specifically, we used RDF Mapping to extract knowledge and map structured data to RDF data. Then, we stored the extracted knowledge in Neo4j, complemented the new knowledge, and integrated it into the static road network topology. The effective construction of driver travel knowledge graph has laid a good foundation for traffic travel knowledge query and reasoning. At the same time, the explored method of constructing the knowledge graph from the structured data can be extended to the structured data in other fields, which promotes the application of the knowledge graph.

Key words: knowledge graph, trajectory big data, graph database, RDB2RDF

随着移动定位设备、大数据技术的不断发展，人类活动每时每刻都会产生大量带有位置坐标和时间标签的轨迹大数据。轨迹大数据中蕴含着丰富的路网特性与交通特性，来源广泛，数据格式多样。轨迹大数据中的时空维度动态关联特性被离散化表示，语义特征被弱化，使得从轨迹大数据中挖掘有价值的信息变得困难[1-3]。基于异构图结构的知识图谱为海量、异构、动态的数据表达、组织、管理提供了一种有效的方式，更接近于人类的认知思维[3]。目前，知识图谱的构建方式主要包括自底向上和自顶向下两种[4]，两种方式均没有标准且统一的构建流程，其中自顶向下的方式多用于借助百科类网站、专家知识等数据源，从高质量数据中提取本体和模式信息，再加入到知识库中，采用志愿者众包编辑数据页面或咨询专家的方式来构建，费时费力费财[5]。另外，数字化时代中每个企业或组织在其业务系统中都积累了大量数据，原有系统已不能满足人们迅速、准确、智能获取信息的需求；更多的用户希望数据得到更好的组织，进而为数据拥有者提供快速准确的知识抽取、加工，帮助智能决策。领域知识图谱可采用自底向上的方式构建，从

收稿日期：2022-07-15。
项目来源：国家自然科学基金资助项目（61472340、61303017）；河北省自然科学基金资助项目（F2021210005）；河北省省级科技计划资助创新能力提升计划资助项目（21550803D）；河北省教育厅青年拔尖资助项目（BJ2021085）；河北省科技厅大学、中学在校生科技创新能力培育专项资助项目（DXS202106、22E50118D）；国家级大学生创新创业训练计划资助项目（202110107015）。
（*为通信作者）

大量的业务数据中提取置信度高的资源模式，方便查询、发现和挖掘知识[6-7]。因此，本文基于已有的网约车轨迹大数据，采用知识图谱表达交通实体和关系的多样性。基于存储在MySQL中的网约车轨迹和订单信息等业务数据，自底向上构建司机出行交通知识图谱，提取司机出行特点与规律，发现和挖掘隐藏知识，为交通预测和决策提供更多的依据。

综上所述，本文贡献在于，基于新的知识需求探索出了从关系型数据库到知识图谱自动转换与构建的路径；在真实数据集上构建了司机出行知识图谱，可通过Neo4j的Cypher语言进行交通知识查询，在目标领域的分析上具有实用价值。

1　数据集简介

本文采用滴滴出行盖亚数据开放计划提供的2020年8月深圳市网约车的真实业务订单数据，具体包括路网拓扑数据和订单行程信息，路网拓扑数据存储在CSV文件中，结构为<link id,relation,link id>。经统计，一个月的订单数量为8 562 059条，涵盖了80 399位网约车司机。将数据源中的订单行程数据存储在MySQL中，结构如表1~3所示。

表1　订单信息表

列名	数据类型	说明
order id	string	订单编号
ata	float	订单真实行程总时间
distance	float	订单路线距离
simple eta	float	预估通行时间
driver id	int	司机编号
slice id	int	出发时刻时间片
date	string	日期

表2　订单下轨迹的道路小段信息表

列名	数据类型	说明
link id	int	道路小段编号
link time	float	出发时刻道路小段平均通行时间
link ratio	float	道路小段通行比例
link current status	int	出发时刻道路小段路况状态
link arrival status	int	到达时刻道路小段路况状态
order id	string	订单编号

表3　订单下轨迹的路口信息表

列名	数据类型	说明
cross id	int	路口编号
cross time	float	路口通行时间
order id	string	订单编号
rlink	int	路口上游道路小段编号
clink	int	路口下游道路小段编号

1.1　司机活跃度

对该月内每位司机的接单数量进行统计分析发现，该月只接单一次的司机人数最多，活跃度最低；活跃度最高的司机在该月接单603次，如图1所示，司机接单数量呈长尾右偏分布。

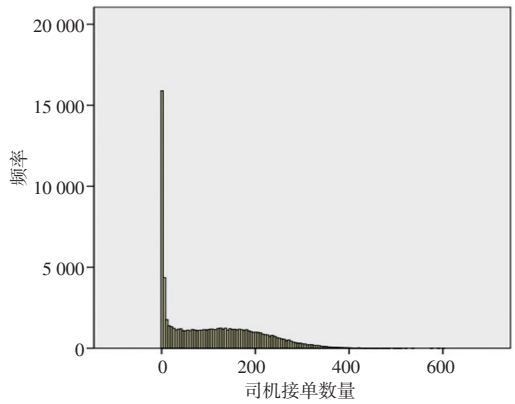


图1　司机活跃度分布直方图

1.2　订单流量分布

网约车司机的订单数量一定程度上反映了司机在某交通区域内的通行频率。本文选取活跃度位于Top5%的司机，从每天上午8:00开始，以每5 min为一个时间戳、24 h为一循环，时间戳的范围为0~288，查看所选司机在不同时间段的订单流量分布情况，如图2所示，可以看出，司机接单呈现出早高峰和晚高峰的特点，大量订单产生在7:00—23:00时间段内，符合通勤规律。

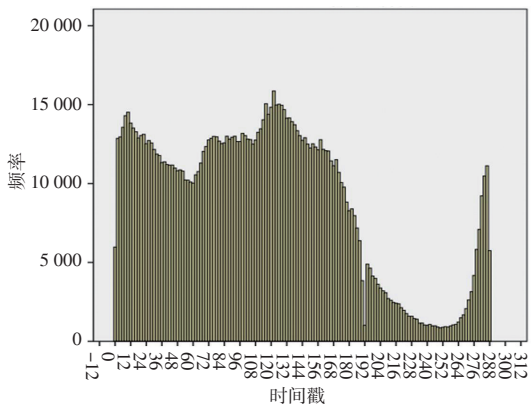


图2　订单流量分布直方图

2　司机出行知识图谱构建

2.1　总体设计思路

自底向上构建知识图谱，即从数据层出发首先归纳实体，再抽取实体、属性和实体间的关系，然后归纳知识并对知识进行存储，最终实现知识图谱的构建。其中，知识归纳的核心是知识抽取、知识融合、知识补全。存储在MySQL数据库中的出租车订单行程

数据包含大量的交通实体、属性和实体间的关系，信息散落在不同的数据表中被割裂，对于数据查询不利，较难从中获取交通知识。本文采用五步法，通过知识抽取、知识存储、知识引入、知识融合和知识补全，实现了从MySQL存储的网约车业务数据到司机出行知识图谱的构建。总体构建流程如图3所示。

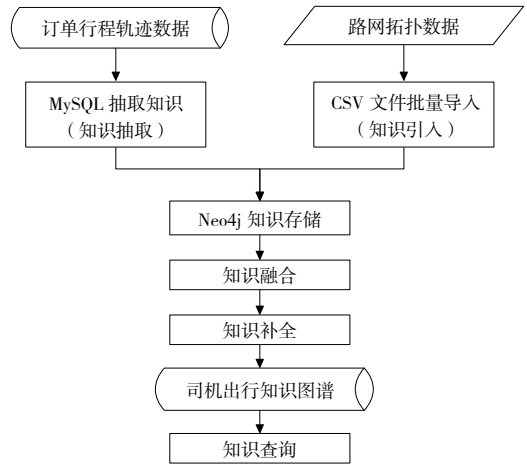


图3 总体构建流程图

2.2 具体构建流程

2.2.1 知识抽取

首先，梳理存储在MySQL订单行程数据中数据的含义和关系，明确交通事件的实体、实体属性以及实体间的关系，根据数据间的关联和特征，初步分析交通轨迹的简要特征；然后，根据所描述对象的不同，分为司机、订单、路段和路口4大类。由表1~3可知，司机与订单具有一对多的关系，订单与轨迹具有一对一和一对多的关系，如视订单为一个实体，订单编号、订单行程总时间、订单路线距离是属性。

本文采用D2R技术从网约车业务数据中抽取知识。D2R技术，即Database to RDF，是一种能从关系型数据库中抽取知识的技术，主要包括D2R Server、D2RQ Engine和D2RQ Mapping语言。D2RQ根据可定制的D2RQ Mapping文件将关系数据库中的数据转化为RDF格式。RDF，即资源描述框架，是一种用于描述事物的方法，其本质是一种数据模型。RDF形式上表示为SPO三元组，即<主体，谓词，宾语>，既可称为一条语句，也可表示知识图谱中的一条知识[8]。在Mapping文件中定义了整个映射的主要模式：数据库中的表映射为一个类，表中的每一行是一个资源或实例，每一列是资源的属性，表与表的外键则映射为关系[9]。本文基于D2RQ工具对MySQL中的订单行程数据进行知识抽取，以一条订单信息为例，说明从二维表数据到RDF数据的转化过程。

MySQL中在2020年8月1日记录了一条订单编号为1612040的订单。通过D2RQ工具得到的RDF数据为：

<file:///D2RQ安装目录/traffic.nt#tbl_order/1612040>
<file:///D2RQ安装目录/vocab/tbl_order_date>
<file:///D2RQ安装目录/traffic.nt#date/20200801>.

这条RDF数据表示为订单1612040和日期20200801之间的关系是tbl_order_date，即在2020年8月1日产生了一条订单编号为1612040的订单。

2.2.2 知识存储

通过上述操作，将存储在MySQL中的订单行程数据转化为大量的RDF数据。知识图谱主要包括RDF和图数据库两种存储方式，目前普遍认为，RDF的重点在于数据发布和资源共享，图数据库的重点在于对图数据的高效查询和搜索[10]。因此，本文选择主流图数据库Neo4j来存储知识图谱。

利用Neo4j的Neosemantics插件将生成的大量RDF数据导入到Neo4j中进行存储。在Neo4j中，知识便以图的形式组织在一起，进而形成了基于属性图的知识图谱。利用Neo4j的Cypher语言进行交通实体、关系和相关实体属性的查询，获得与司机出行相关的知识，如某时段路网交通状态、司机历史轨迹和各时段网约车流量等。此外，利用Neo4j的可视化交互前端可以图的形式展示交通实体、关系和属性以及规模。

由MySQL中的订单行程数据构建的图谱如图4所示，以司机节点为核心，关联到与该司机相关的订单节点，再由订单节点关联到该订单行程轨迹，具体包括路段和路口。然而，这些路段和路口在图谱中并没有以轨迹的形式展现，没有还原出原有订单行程的轨迹信息，轨迹序列仍是离散的形式。对于不同的业务订单，其订单行程轨迹可能存在路段交集。一方面，图4中的图谱不能直接显示存在路段交集的轨迹；另一方面，这些路段交集具有相同的link id和拓扑关系，在图谱中会出现多次。

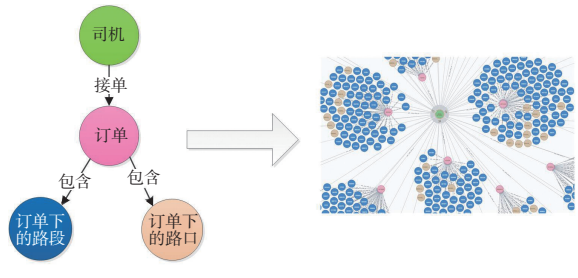


图4 订单行程数据形成的图谱

2.2.3 知识引入

网约车轨迹记录了司机出行过程中车辆移动过程和车辆状态变化，包含网约车司机出行的知识。为了

连续表达订单行程中重要的轨迹信息，完整还原司机出行的全过程，本文将动态的订单业务数据与静态的交通路网相结合，使订单行程轨迹关联到静态路网上。本文利用Neo4j的load CSV功能和Cypher语言批量建立路段节点与路段间的拓扑关系，再将静态的路网知识引入，最终在Neo4j中建立路网拓扑图。具体构建流程如图5所示。

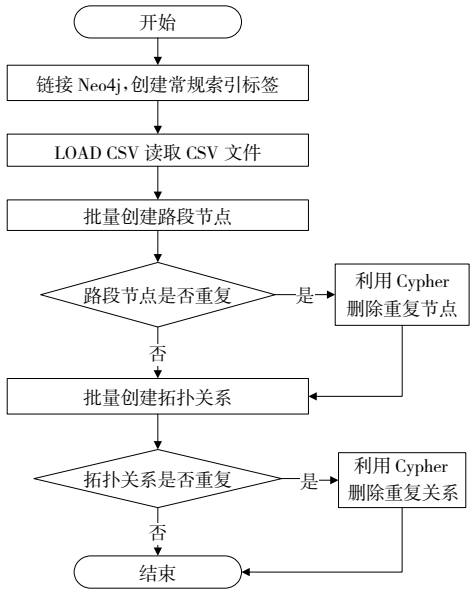


图5　路网图谱构建流程图

2.2.4　知识融合

如图4所示，仅由订单轨迹形成的图谱不足以表示司机从接单、行驶再到目的地的全过程，且难以体现轨迹特点。因此，本文将订单行程数据形成的图谱与路网拓扑形成的图谱进行知识融合，使订单行程轨迹关联到静态路网上。本文将具有相同link id的订单行程轨迹路段与路网中的路段建立关系，表达为路网中的路段在不同订单轨迹下的状态。状态信息包括在此订单下路段的通行时间、通行比例和路况状态。经过上述处理后，离散的订单行程轨迹通过路网拓扑被连续表示；同时，静态的路网拓扑具有在不同订单下的路段状态信息。订单下的路段与静态路网的连接关系如图6所示。

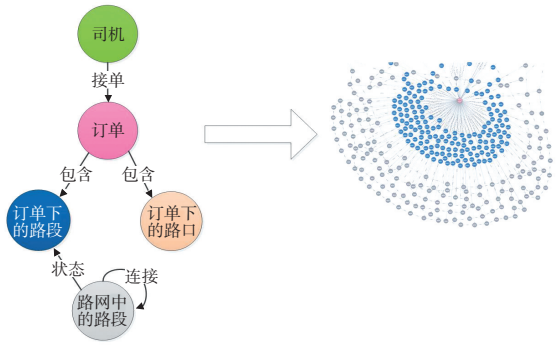


图6　知识融合后的图谱

2.2.5　知识补全

仅交通路网不足以体现轨迹中蕴含的交通知识，还需将路口等信息补充到交通路网中。订单行程下的路口节点包含其所属订单编号、路口编号、路口通行时间和路口上下游路段的link id。基于路网拓扑关系，利用Python计算每个link节点的出入信息，综合订单行程下产生的路口信息，再通过Cypher语言将路口知识补充到知识图谱中。

2.2.6　交通知识查询

在构建的司机出行知识图谱上，可利用Neo4j的Cypher语言进行交通知识查询，可查询司机历史轨迹并通过Neo4j可视化界面对网约车轨迹进行展示。对于某订单的轨迹，可进一步查询轨迹中各路段在司机行驶时的状态，显示相关状态信息，如轨迹中各路段的通行时间、通行比例、路况和路口通行时间等。

3　结　语

本文构建的司机出行知识图谱将路网结构与司机出行知识相结合，支持交通出行知识的查询。本文采用自底向上的方式从结构化数据中构建司机出行知识图谱，还原了轨迹大数据中的时空关系和语义关联，探索出了从关系型数据库到知识图谱自动转换与构建的路径，可被扩展应用于其他领域的结构化数据，对推广知识图谱应用起到了推进作用。

未来考虑从司机的视角出发，结合构建的司机出行知识图谱，还原司机从接单到行驶、再到达目的地的全过程，从中挖掘出有关于司机的交通知识。在司机接单推荐过程中，提供交通知识进行决策；可查询目标起终点历史轨迹，利用知识图谱的异构图结构并结合图算法进行轨迹推荐和时间预估等交通出行预测，使出行更加智慧、更加个性化。

参考文献

[1]　牟乃夏，张恒才，陈洁，等. 轨迹数据挖掘城市应用研究综述[J]. 地球信息科学学报，2015，17(10)：1 136-1 142

[2]　李仕峰. 多源异构数据时空融合关键技术研究与应用[J]. 地理空间信息，2021，19(10)：19-21

[3]　毛建，王晓霞，徐志远，等. 基于浮动车数据的交通出行特征库构建研究[J]. 地理空间信息，2020，18(5)：46-48

[4]　李涛，王次臣，李华康. 知识图谱的发展与构建[J]. 南京理工大学学报，2017，41(1)：22-34

[5]　李涓子，侯磊. 知识图谱研究综述[J]. 山西大学学报(自然科学版)，2017，40(3)：454-459

（下转第130页）