DOI:10.13203/j.whugis20200201 文章编号:1671-8860(2020)06-0816-10

COVID-19病例活动知识图谱构建——以郑州市为例

陈晓慧[1] 刘俊楠[1] 徐 立[2] 李 佳[1] 张 伟[1] 刘海砚[1]

1 信息工程大学数据与目标工程学院,河南 郑州,450001
2 信息工程大学地理空间信息学院,河南 郑州,450001

摘 要:目前,随着全球新型冠状病毒肺炎(coronavirus disease 2019,COVID-19)病例数量不断增加,疫情时空传播过程变得越来越复杂。传统的传播过程研究主要是在宏观上研究传染病的整体传播规律或趋势,不能在个体层面分析具体病例之间的传播关系,无法精准定位疫情传播路径,很难支持传染病的精准防控,亟需兼顾时空和语义特征研究传染病传播过程。首先在解析COVID-19病例数据基础上,利用知识图谱技术提出了构建适应多样化描述方式的COVID-19病例活动知识图谱;然后从传播事件角度设计了COVID-19病例活动知识图谱本体规则,完成了模式层的构建;并以流行病调查数据为基础,对病例数据进行解析、事件实体识别和数据存储,完成了数据层的构建;最后,通过图数据库和B/S端构建原型系统进行实验验证。结果表明,通过COVID-19病例活动知识图谱对传播过程推理、关键节点分析和活动轨迹回溯等层面进行验证,方法较为有效,且具有一定可行性。

关键词:新型冠状病毒肺炎;传染病调查;活动知识图谱;知识图谱可视化

中图分类号:P208 **文献标志码**:A

尽管经济的发展和医疗水平不断提高,传染病对人类的威胁并没有降低。近年来,埃博拉病毒、登革热、中东呼吸综合征等恶性传染病频繁暴发,对人类健康、经济发展和社会稳定构成严重威胁。2019年12月底和2020年初暴发的新型冠状病毒肺炎(coronavirus disease 2019,COVID-19)传染性强,影响范围广泛,引起了全世界的关注。截至2020-05-28,在中国境内的COVID-19疫情基本得到控制的情况下,其他国家的疫情呈快速大规模扩散态势。经验表明,通过研究传染病传播规律和路径能够有效控制传染病,因此,以多源病例活动数据为基础,通过病例的时空和语义特征研究流行病传播规律,已成为政府和学术界关注的焦点,同时也是当今世界迫切需要解决的重大问题。

随着COVID-19病例数量不断增加,疫情时空传播路径变得越来越复杂。目前已有的传染病传播规律研究主要包括利用数学传播模型和仿真传播模型方面的研究。其中,数学传播模型的经典模型是一种常微分方程组[1]以及在此基础上增加的随机模型[2],很多学者在此基础上针对特定传染病的传播规律进行了研究[3-4]。由于数学传播模型是一种理想化的模型,与实际存在一定差距,为克服数学传播模型的不足,有学者通过建立仿真传播模型对传染病传播规律进行了研究[5-7]。然而,数学传播模型和仿真模型基于均质空间假设,主要研究传染病例者数量随时间变化的关系,侧重数理统计分析,忽略了时空特征对传染病传播的影响;随后,很多学者从时空演变的角度对传染病的传播趋势展开研究[8-9],在时空维度上很大地改进了传播模型。然而基于时空信息的传播模型研究主要考虑人口的空间分布特征,而忽略了病例之间社会关系、语义关系和时序关系的表达。以上3种模型研究主要是宏观上研究传染病的整体传播规律或趋势,而不能在个体层面表达具体病例之间的传播关系,无法精准定位疫情传播路径,很难支持传染病的精准防疫。

收稿日期:2020-04-29

项目资助:国家自然科学基金(41801313,41901397)。

第一作者:陈晓慧,博士,副教授,主要从事时空数据可视化与知识图谱相关研究。cxh_vrlab@163.com

通讯作者:刘海砚,博士,教授。liu2000@vip.sina.com

此外,传染病医学领域引入知识图谱相关技术描述生物医学和临床医学领域相关概念及概念之间的关系,对传染病学知识的理解和共享提供了概念模型支持[10];但是从医学角度所构建的传染病本体规则主要应用于疾病治疗的知识建模,难以在时空维度展示传染病的传播防控过程,且无法展示病例的人群活动趋势和传播路径;因此,需要结合人群活动事件的分析方法,从事件角度对传染病传播进行知识建模,更符合传染病传播分析的需求。目前,以传染病为概念本体构建的医疗领域知识图谱被广泛应用于COVID-19防治的研究,例如OpenKG(http://openkg.cn/dataset/covid-19-concept)从网络文本中采集了与COVID-19疾病相关的实体和关系,进一步融合了百度百科、维基百科等知识库,构建了COVID-19概念知识图谱(http://openkg.cn/dataset/covid-19-epidemiology),重点刻画传染病学的基本概念,为后续的研究提供了重要的理论基础。但该研究仅侧重于COVID-19流行病医学角度的信息,缺少病例的时序和空间关系概念体系,无法适应大数据时代病例信息的多样化描述方式,也难以表达病例实体的活动关系。

现有传染病传播规律的研究各具特点,已在不同领域发挥了重要作用。但传染病传播过程与人群活动事件密切相关,即人群活动直接影响COVID-19的时空扩散途径;针对COVID-19传播时空规律分析研究主要关注统计层面分析,对于传播事件语义分析较少,需将二者结合进一步探索传播动态演变路径。因此,亟需以传染病病例为中心,结合知识图谱前沿技术,兼顾时空和语义特征的数据组织形式对病例数据进行建模。

1　COVID-19病例活动知识图谱

流行病学调查是一个极其复杂和繁琐的过程,但有助于人们了解疾病传播风险,进而制定有效的防治与控制策略。通过流行病学调查可以明确每一个患者的感染路径,确定可能感染的人群,在后续的疫情防控中发挥巨大作用。针对数万病人进行传染病调查、溯源、密切接触者追踪是一项耗时、耗力的工作,而通过构建病例活动知识图谱可以为医护人员和疾病防控人员提供技术支撑,加速传染病调查研究。

1.1　人群活动要素解析

传染病传播过程与人群活动事件密切相关,即人群活动直接影响着COVID-19的时空扩散范围。活动由一系列事件构成,是什么人(角色)、什么时候(时间)、在哪里(地点)以及干什么(目的)的组合[11]。

Peuquet[12-13]概括活动由空间(Where)、时间(When)和对象(Who)3部分组成。Orellana和Renso[14]进一步将活动描述为活动对象与所在环境之间相互作用的集合,并将活动的组成部分扩充为6要素集合:何事(What)、何人(Who)、时间(When)、地点(Where)、原因(Why)和途径(How)。因此,为准确详实地描述COVID-19"人传人"的传播过程,以活动六要素模型(5W1H)为基础,分析病例活动要素组成,如图1所示。其中,Who是活动参与对象集合,包括活动主动参与对象和被动承受对象,具有显著时空特征,是描述病例活动的必要元素,据此可掌握活动的人-物、人-人和物-物关系;What指病例活动参与对象现在的状态信息,包括疑似、确诊、治愈和死亡;When指病例活动发生的时间相关信息,包含疑似时间、确诊时间、病程时间和出院时间,可展现病例活动的时间维度信息,进而体现传播程度;Where指病例活动发生的地点,包括定性语义描述和定量坐标范围,也可转换为一系列轨迹数据集,据此展现病例活动波及地区和蔓延范围;Why指病例活动发生原因,体现病例患病的因果逻辑关系;How指病例活动的途径,可作为推测活动类型和预测波及范围的基础。

1.2　病例活动知识图谱

事件描述人群活动的基本构成单元。文献[15]于2018年提出事理知识图谱概念,从新闻中提取并抽象泛化事件,并通过知识图谱技术描述事件演化模式,以实体节点表示事件,利用边表示事件间语义关系,其中节点、边相互链接构成有向循环图,表示事件间的顺承、因果关系,进而描述事件的演化模式。但是,事件包含时间、地点等大量信息,现有事理模型无法表示事件的时空关系,难以预测传播事件时空变化过程。传染病传播与人群的时空活动密不可分,故本文在事理知识图谱的基础上,利用人群活动描述刻画病例活动事件,更易满足传染病传播趋势分析的需要。

将活动知识图谱描述为一张图G(图2),其模式层由传染病学模式层G_s、病例活动模式层G_m、数据图G_d及其之间的关系R组成,即:

$$G=<G_s, G_m, G_d, R>$$

其中,病例活动模式层G_m包括事件子图G_{event}及其之间的时序关系R_t,即:

$$G_m=<G_{\text{event}}, R_t>$$

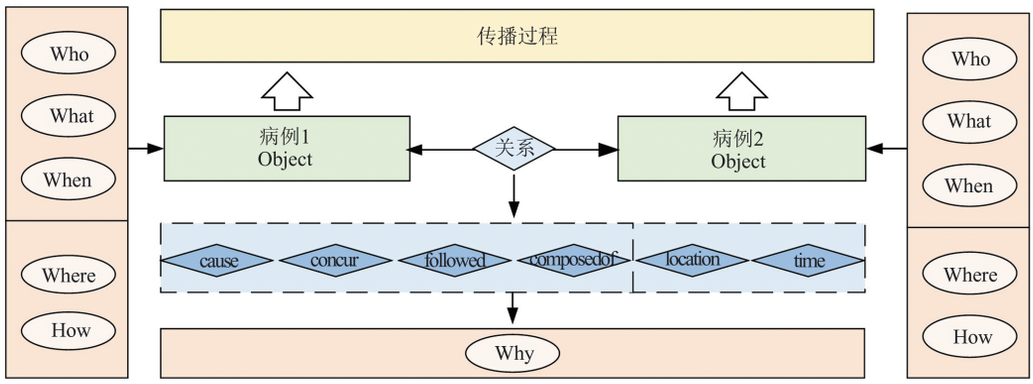


图1 传染病病例构成要素分析

Fig. 1 Analysis of Infectious Disease Cases Elements

事件子图G_{event}由一个五元组$<N_{When}, N_{Where}, N_{Who}, N_{What}, N_{How}>$组成，$N_{When}$、$N_{Where}$、$N_{Who}$、$N_{What}$、$N_{How}$分别对应活动要素的何时、何地、何人、何事和途经，即：

$$G_{event} = <N_{When}, N_{Where}, N_{Who}, N_{What}, N_{How}>$$

将传染病学模式层G_s引入现有传染病本体，表示病例的基本情况(年龄、性别等)，此外引入时间本体表示事件子图之间的时序关系R_t，进而展现“疑似→确诊→治愈”等活动记录的时序关系。

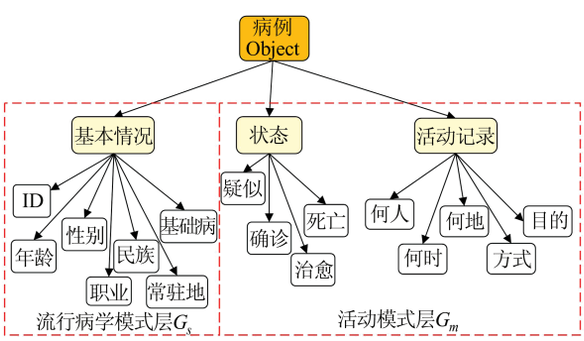


图2 知识图谱模式层概念体系

Fig. 2 Conceptual System of Knowledge Graph Pattern Layer

COVID-19流行病学知识图谱属于领域知识图谱，采用自顶向下的构建方式，即从模式层开始构建。构建流程如图3所示，模式层构建是知识图谱中最核心的部分，为知识图谱定义数据的模式(Schema，即为其定义本体)。在定义本体的过程中，从顶层概念体系逐步细化，进而形成具有良好结构的分类学层次体系。本文采用有人工参与的本体模型构建、分类分层和概念梳理，提高知识图谱数据的完整性和准确性。定义模式层后，从COVID-19传染病相关数据源中进行实体抽取、实体链接和存储，实现数据层填充。在完成模式层和数据层构建之后，就初步完成COVID-19流行病学知识图谱的构建过程，在此基础上可以进行语义检索和决策支持。

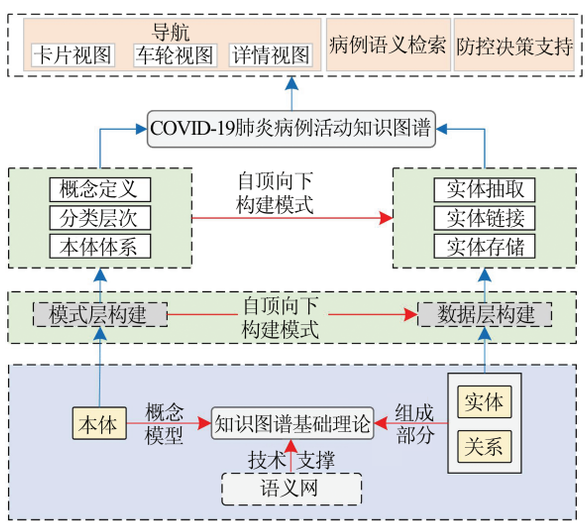


图3 知识图谱构建思路

Fig. 3 Design Ideas of Terrorism Data Model

2 基于简单事件模型的本体规则设计

本节采用有人工参与的本体模型构建、分类分层和概念梳理方式。本体规则是构建COVID-19知识图谱模式层的核心工作，研究人员通过研究传染病本体体系不断扩展传染病领域的知识表达，定义了传染病学的基本概念及概念之间的关系，但仅从医学角度描述传染病知识，较难满足COVID-19病例的动态时空特征展示和预测。

本体是知识图谱的重要组成部分，可形式化表示病例活动相关概念的层次结构关系。面向大数据时代病例活动的多样化描述方式，本文以模式层的事件活动为核心，以不改变现有本体结构即可添加概念与关系为目标，采用本体描述语言(web ontology language，OWL)实现病例的概念层次语义关系和时空关系表达。

简单事件模型(simple event model,SEM)[16]是一种不依赖于领域词汇的通用事件表示模型,可以针对不同领域的事件进行建模。以事件的核心概念、类别系统以及属性约束描述不同领域事件,综合利用时间、地点、对象和事件4个概念来描述事件的组成要素;通过设置对应于核心概念的类别系统,可以在不改变模式层的情况下,通过其具体实例描述事件要素的类别信息;属性约束用于描述知识图谱中的属性,通过向现有属性添加信息可对其进行约束或拓展现有属性的描述信息。采用SEM模型对病例活动建模符合病例活动的特点,构建时空事件模型描述活动记录中的子事件概念模型。

2.1　概念层次关系设计

概念层次关系需包括传染病学本体模型以及拓展的活动概念和概念层次关系,且涉及"5W1H"对应的时间、地点、对象、方式、原因等要素。由于现有模式层结构难以描述多样化信息,因此采用文献[17]研究的事件模型描述活动记录中的子事件概念模型。其中,核心概念(Core Classes)包含时间、参与者、地点和事件4个概念,描述事件的组成要素;类别系统(type system)与核心概念一一对应;属性约束(property constraints)作用于知识图谱的属性。本文首先结合活动事件组成要素扩展何故(why)与如何(how)两个事件要素的概念及其对应的类别系统,即cg:How、cg:Why与cg:HowType、cg:WhyType。此外,结合属性约束展示事件的时序关系、空间关系和语义关系。

2.2　实体关联关系设计

知识图谱模式层的关联关系包括传染关系、社会关系、病例活动事件关系以及时序关系。社会关系属于语义关系,包括亲属关系、同事关系等(图4中ⓐ);传染关系与病例之间的接触有关(图4中ⓑ),即与病例活动相关,具有丰富的语义和时空关系。

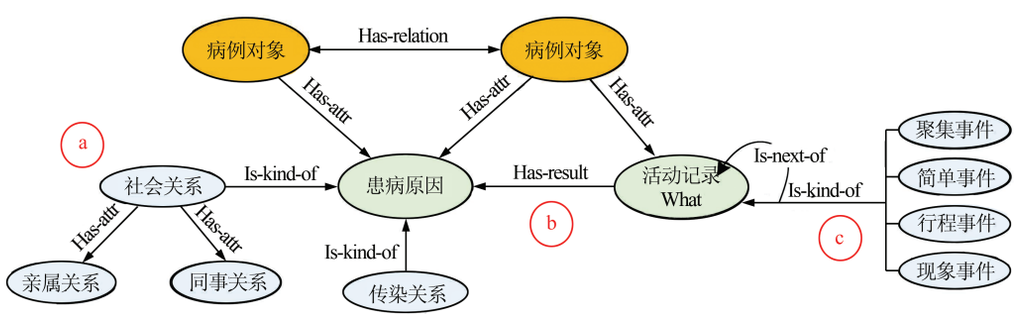


图4　实体关联关系设计

Fig. 4　Entity Correlation Design

病例活动事件包括聚集事件(如聚餐事件)、简单事件(如接触事件、居家隔离事件、就医事件)、行程事件(如旅行事件、购物事件)、现象事件(如发热事件)(图4中ⓒ)。此外,事件的语义关系包含组成、因果、跟随、并发关系,主要以非层次关系存在,可以刻画病例之间的传染关系。如表1所示,活动记录M_1由E_1、E_2、E_3、E_4和E_5组成,可描述为$R_{\text{composedof}}(M_1,E_1)$等,指活动与事件构成组成关系;因果关系指事件导致其他事件发生,如接触事件E_1导致发热事件E_4,可表示为$R_{\text{cause}}(E_1,E_4)$;跟随关系指在一定时间区间,事件跟随其他事件发生,例如"发热事件"(E_2)跟随"行程事件"(E_3),可表示为$R_{\text{followed}}(E_2,E_3)$;并发关系指在一定时间范围内,两个事件同时发生但不具有因果关系,如就医事件发生时,发热事件同时进行,可以描述为$R_{\text{concur}}(E_2,E_5)$。

时序关系可通过时间点和时间段描述传染病活动事件的时间特性,也能以此为基础提取事件的时序关系,进而发现事件在邻近时间域内的相互依赖关系和作用机制。不同类型的事件之间的时序关系如表1所示。根据时序关系的传递性,若E_1事件先于E_2事件,E_2事件先于E_3事件,可推导出E_1事件先于E_3事件。其中,称E_1事件与E_2事件存在直接的时序关系,E_1事件与E_3事件存在间接的时序关系。为了避免构建过程中时序关系的冗余性,仅建立事件之间直接的时序关系。

表1　传染事件语义关系实例

Tab. 1　Semantic Relationship of Epidemic Event

语义关系	逻辑描述	说明
组成关系	$R_{\text{composedof}}(M_1,E_1)$	活动是由事件E_1组成
因果关系	$R_{\text{cause}}(E_1,E_4)$	E_1事件导致事件E_4发生
跟随关系	$R_{\text{followed}}(E_2,E_3)$	事件E_2跟随事件E_3发生
并发关系	$R_{\text{concur}}(E_2,E_5)$	事件E_2和事件E_5同时进行

2.3　层次关系与事件关系表示

病例的活动记录由多个子事件构成,事件要素由时间、地点、方式、参与人、目的等构成。利用sem:Actor和sem:Object表示事件的参与者,sem:How

和sem:Why描述事件的发生方式和原因,sem:Place和sem:Time表示事件的时间、地点信息,时间和地点可通过属性值和实体进行表示。本节在不拓展模式层现有概念情况下,通过添加概念实例,表示活动事件的类别信息,同时通过hasSubType表示不同概念实例的父子关系,进而表示概念的层次关系(图5),如<交通工具,hasSubType,飞机>三元组。

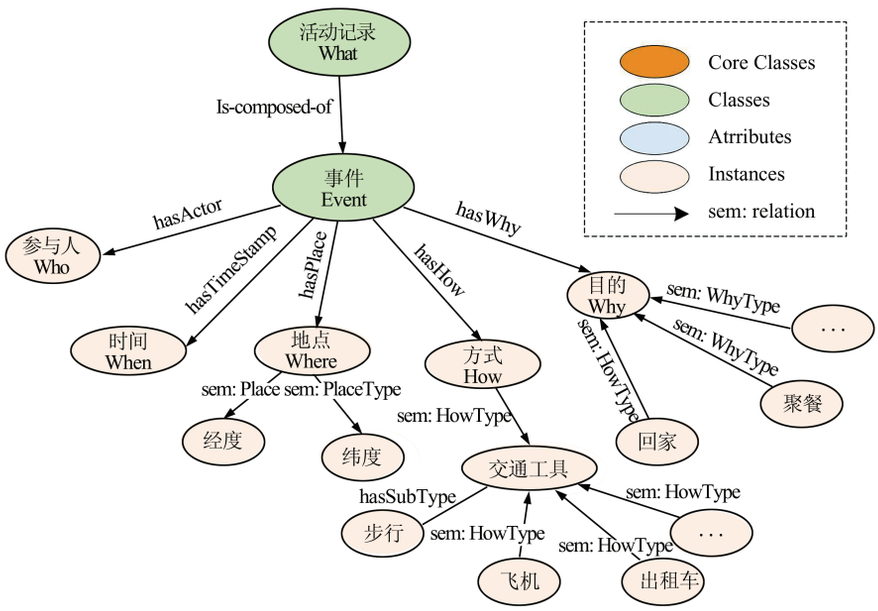


图5 活动事件的关联关系

Fig. 5 The Structure and Relationship of Movement

利用传统知识图谱以模式层的概念关系为约束添加实例关系,用以表达传染病学知识图谱模式层,通过模型属性约束(sem:Role)的实例以及主语(rdf:subject)和谓语(rdf:object)两个关系连接核心概念(sem:Event)的实例,并结合sem:RoleType表示事件关系,如三元组< E_1, cg:after, E_2>可通过<cg:after, rdf:subject, E_1>、<cg:after, rdf:object, E_2>和<cg:after, rdf:type, sem:RoleType>3个三元组表示;通过类别系统的具体实例描述事件类型,如< E_1, sem:EventType, 出行事件>和< E_1, sem:EventType, 确诊事件>两个三元组分别表示E_1和E_2为出行事件和确诊事件(图6)。其中,实心箭头、空心箭头和虚线箭头分别表示对象属性、概念父子关系和概念间关系,椭圆代表概念,矩形表示概念的属性值。本小节以现有本体结构为基础,无需扩充模式层关系类型,通过属性约束和主谓语关系等现有关系和概念类型,实现事件的非层次关系表示,提升了模型的向后兼容性,可更好地适应COVID-19的多样化描述方式。

3 COVID-19活动知识图谱数据层构建

数据层构建的主要任务是构建COVID-19病例流行病学调查数据中实体、关系、属性等知识的三元组。首先,对流调数据进行特征分析,解析其数据的组成类型;然后,对流调数据中出现的事件要素进行识别并抽取,将识别并抽取的实体链接到知识库中,通过D2R与结构化数据进行知识融合;最后,采用图数据库进行知识的管理和存储。通过上述步骤,可以完成知识图谱的数据层构建,进而形成数据丰富、语义完整的知识体系。

3.1 病例数据特征分析

本文收集了2020-01—2020-03期间的河南省郑州市COVID-19确诊患者的流行病学调查数据(下文简称为病例数据),这些数据来自河南省及郑州市卫健委公布的个案流调信息,部分原始数据如表2所示。从表2可以看出,病例数据是一种非结构化文本数据,主要内容包括患者基本情况、患者活动记录(含时间和地点、活动目的等信息)、患者社会关系、当前状态和疾病传播路径等,涉及以下几种数据类型。

1)时序数据

时间是一个非常重要的维度和属性。随时间变化、带有时间属性的数据称为时序数据,是一种有序型数据,主要包括两类:以时间轴排列的时间序列数据和不以时间为变量,但具有内在

排列顺序的顺序型数据集。这类数据的变化顺序可以映射为时间轴进行处理。将病例经过的地点在时间轴上进行有序串联,就可以生成实体的时序数据。

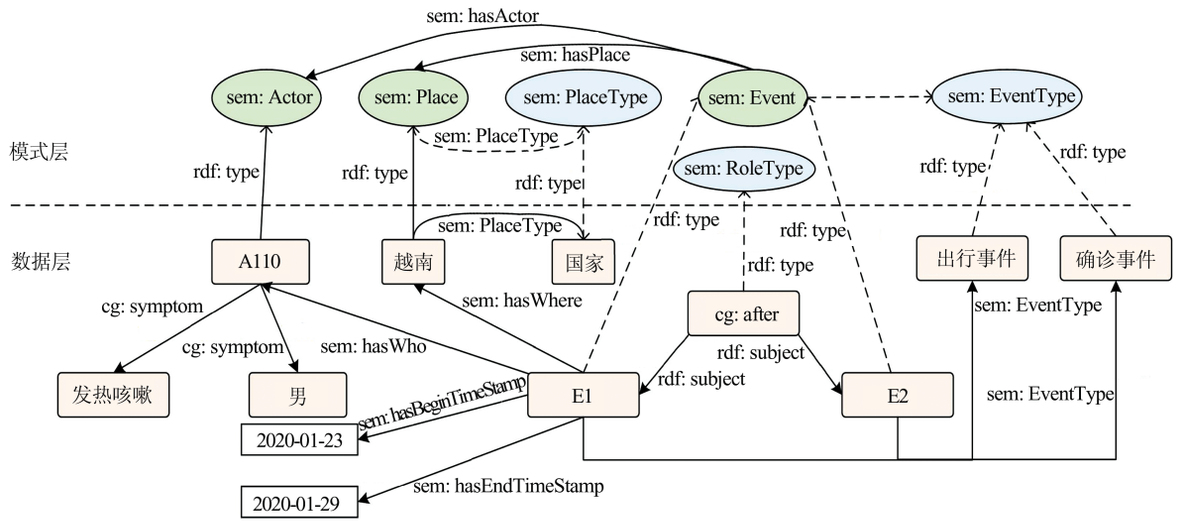


图 6　COVID-19事件关系与类型表示实例

Fig. 6　COVID-19 Event Relationships and Types Represent Instances

表 2　郑州市 COVID-19确诊病例流行病学调查数据(部分)

Tab. 2　Epidemiological Investigation Data of COVID-19 Cases in Zhengzhou

ID	基本情况	活动事件描述
病例 1	男,65岁,周口市太康县清集镇人	1月 7日由武汉乘私家车返回太康县,1月 8日前往太康县人民医院就诊,1月 10日经 120急救车转运至郑州颐和医院就诊,1月 20日由负压救护车转诊至郑州市第六人民医院,1月 21日确诊。
病例 2	男,55岁,信阳罗山人	1月 8日从武汉乘大巴车回到信阳罗山县,1月 11日乘 T3040次列车从信阳转至新乡,先后在原阳县人民医院和原阳县妇幼保健院就诊,1月 16日驾车至郑州大学第一附属医院(郑东院区)就诊,1月 22日确诊。
病例 4	女,汉族,30岁现住新乡市原阳县	1月 11日起陪护其家人(病例 2),1月 16日到郑州大学第一附属医院(郑东院区)就诊,1月 23日确诊。

2)地名数据

地名数据是一种包含地点名称和坐标的数据。患者活动记录数据含有地点名称,如“湖北省武汉市”、“郑州市”等。以上形式的地理空间数据均具有精确的地理属性,可以根据地点名称在地名数据库和 POI(point of interest)数据库中查询地点坐标。全国地名数据库存储了全国范围内从省一级到村一级行政区域的地名信息。POI数据是一种兴趣点数据,如“颐和医院”、“河南省武警医院”、“永安街 35号院”等,包含了地点名称、类型、坐标的信息,记录对象是与生活、工作、教育、娱乐等相关的建筑设施,参考这两种数据可以对地点进行精确定位。

3)语义关联数据

病例数据包含详细的语义信息,如患者户籍信息和患病原因,前者为社会关系信息,包含了病例的潜在社会关系;后者为传染关系信息,虽然部分病例没有明确描述,但是可以根据多个病例进行语义关联分析,如可以根据描述“病例 14为病例 6的家属”推断“病例 14与病例 6可能存在传染关系”。

3.2　传播事件实体识别

流行病学调查数据是非结构化文本数据,但是有着规范清晰的描述标准,一般一句话代表一个事件,可以通过自然语言处理,对病例数据进行语料分析处理。

本文使用 HanLP 中文处理包对文本数据进行分词,并通过分词系统得到相关元素分词后的词性、依存语法等,根据事件要素将所有词的定义分为 5个类型,分别是时间、地点、触发词、参与者、对象。其中,触发词是对事件类别进行识别的重要特征,不同事件具有不同的触发词,如“确诊事件”有“确诊”,“出行事件”有“乘坐”、“自驾”、“步行”等触发词;事件类型的识别由事件触发词和事件要素共同决定,在分词后与语料库中标注触发词进行比对,若相同,则标记该词为触发词。参与者、对象、时间、地点等要素也通过该方式设置类别,处理的结果将作为知识表示的基

础，并标识相应的标签，最后保存为.csv文件并导入图数据库中。

3.3 实体数据存储

采用图像数据库Neo4j来存储抽取的事件要素，Neo4j不同于结构化数据库，不限定存储内容，其数据模型隐含在所要存储的数据中。因此，在对传染病病例实体进行存储时，采用节点和边作为描述形式，以知识三元组（实体，关系，实体）和（实体，属性，属性值）实现病例实体的存储。

§3.2通过自然语言提取的实体以及彼此之间的关系信息，先用.csv格式文件存储，然后使用Cypher语言将整理好的.csv文件导入到Neo4j数据库中，即完成了病例活动知识图谱数据层的构建。

4 实验验证

以2020-01—2020-03郑州市COVID-19病例的传染病调查数据为数据基础，重点研究面向人群活动的COVID-19病例活动知识图谱构建方法，对COVID-19患者进行调查、溯源，以及分析推理密切接触者，提供数据组织基础。本文将郑州市所有确诊病例的时间、空间和语义关系进行整理，使用关系图、地图和多属性分析图多视图进行交互，实现实体的时空特征和语义特征的交互探索分析，支持对COVID-19传播过程和时空活动规律进行探索式分析。

采用本体构建工具Protégé构建模式层，利用Pandas、开源工具Refine、HanLP进行数据预处理，实现病例文本数据的实体抽取，采用Neo4j实现COVID-19病例知识图谱数据的存储与访问，并采用JavaScript语言搭建B/S模式的COVID-19病例知识图谱可视化平台。下面从COVID-19病例知识图谱概览、病例传播过程和活动轨迹回溯3个层面验证模型的有效性和可行性。

4.1 COVID-19病例知识图谱概览

基于实验平台实现郑州市COVID-19病例知识图谱概览，以病例关联的事件、发生时间和地点对病例关系进行聚类，通过原型系统进行可视化展示，从宏观上概览疫情发展态势，从微观上发现病例传染关系、挖掘时空活动轨迹、跟踪超级传播者。时间从2020-02-21—2020-03-11，图谱共计256实体节点，其中病例节点158个，地点实体7个，时间实体31个，交通工具实体69个，活动实体节点385个。

图7为患者的COVID-19病例知识图谱原型系统。其中，图7(a)为病例概览；图7(b)为病例空间分布；图7(c)为病例活动轨迹回溯；图7(d)为病例传播关系；图7(e)为信息编辑；图7(f)为病例活动图谱，图7(g)和图7(h)为病例流行病学信息概览。在病例传播关系概览中，蓝色表示一代病例，玫红色表示二代病例。从图7中可以看出，大部分患者为一代病例，部分患者通过活动接触感染了其他患者，但是没有大规模的扩散传播。

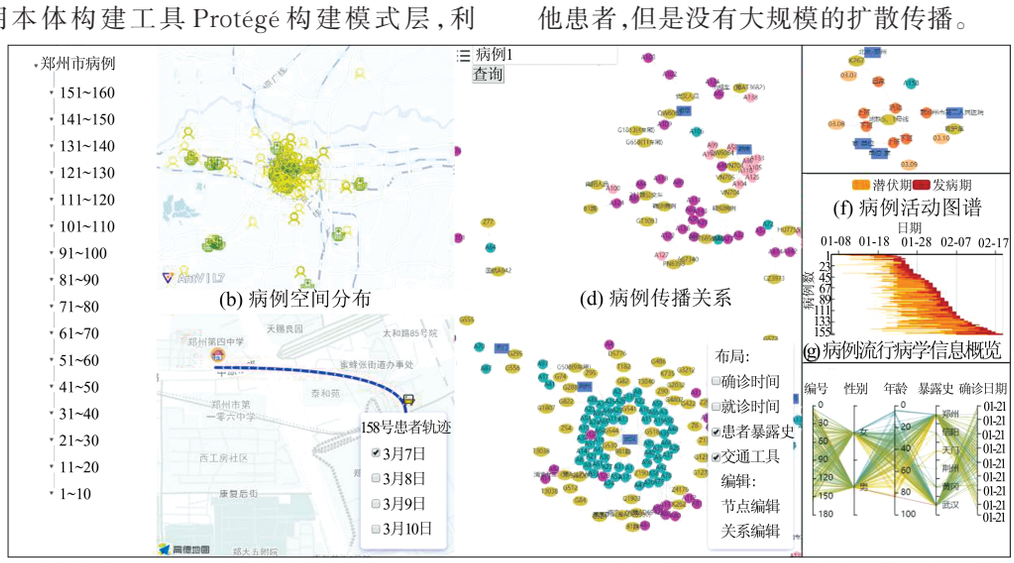


图7 COVID-19病例活动知识图谱可视化原型系统

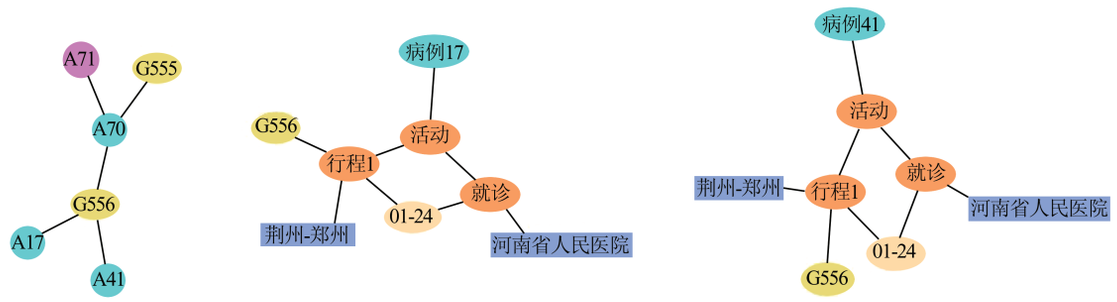
Fig. 8 Visualization of the Coronavirus Disease 2019 Epidemic Cases Activity Knowledge Graph

4.2 COVID-19病例传播过程推理

1）COVID-19病例关系推理

通过语义信息检索病例，并基于原型系统的病例关联关系模块分析病例的语义关系，以病例17（A17）和病例41（A41）的活动图谱为例，如图8(a)所示，发现病例17和病例41在01-24都乘坐

了G556班次高铁,通过图8(b)、图8(c)分别对病例17、病例41的活动事件进行分析,发现01-24他们具有同样的行程,即到同一医院(河南省人民医院)就诊,推测病例17和病例41可能为亲属关系。同时病例17和病例41在到达河南省人民医院就诊时缺少交通工具信息,为了方便传播关系追踪和疫情防控,工作人员应该询问病例增加交通工具信息。



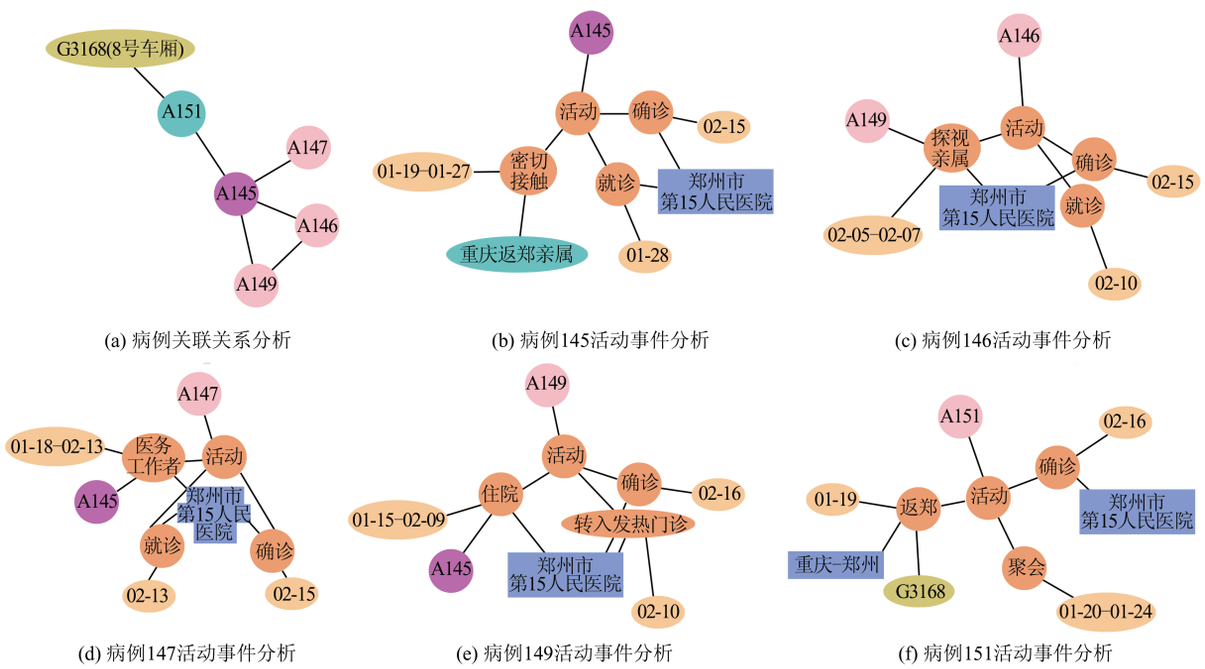
(a) 病例关联关系分析　　(b) 病例17活动事件分析　　(c) 病例41活动事件分析

图8　病例活动图谱分析图

Fig. 8　Patients Activity Analysis Graph

2)COVID-19传播关键节点分析

通过SEM模型的活动事件发生时间和病例,基于原型系统的活动分析模块在微观上对关键节点进行分析。图9(a)展示了一个完整的传播链条,病例151(A151)为潜伏期较长患者,其乘坐G3168车次从重庆返回郑州,之后与病例145(A145)一起生活,之后病例145先出现症状前往郑州市第15人民医院就诊(图9(b))。但是郑州市第15人民医院并没有按照COVID-19的标准收治患者,还安排病例149(A149)同住,使病例145感染了前去探望的病例146(A146,图9(c))以及该医院的一名医护工作者即病例147(A147,图9(d))以及病例149(图9(e))。直到02-15,病例151(A151,图9(f))才得到确诊。可以推断,在该起传播链条中,病例151虽然确诊时间最晚,但却是这条传播链条的原始节点,而病例145是这条传播链条的关键节点,由于对病例145的疫情管控不足,致使3人确诊,多人被医学观察。



(a) 病例关联关系分析　　(b) 病例145活动事件分析　　(c) 病例146活动事件分析

(d) 病例147活动事件分析　　(e) 病例149活动事件分析　　(f) 病例151活动事件分析

图9　病例活动关键节点分析图

Fig. 9　Key Nodes of Patient Activity Analysis Graph

4.3　COVID-19病例时空活动回溯

通过模型的事件发生时间、地点和病例实体检索活动事件,并基于原型系统的病例活动轨迹回溯模块再回溯病例的活动事件。以病例158为例,如图10(a)所示为病例158确诊前的活动图谱,图10(b)、图10(c)分别为该病例在郑州市内

(a) 病例158活动事件分析 (b) 03-07病例158活动轨迹 (c) 03-08病例158活动轨迹

图10 病例时空活动回溯

Fig.10 Retrospective to Patient Activity

轨迹回溯图。03-07,病例158乘坐K267班次列车返回郑州,之后通过步行到家;03-08,病例158乘坐了地铁上下班。由于该病例的活动比较丰富,工作人员可利用该病例的活动图谱和轨迹辅助进行密切接触者的排查工作。

5 结 语

COVID-19疫情的暴发严重影响了社会稳定、经济发展和人类健康,如何基于大数据领域先进技术,综合时空特征和语义信息,分析COVID-19的传播途径成为当前的研究热点。本文首先根据“5W1H”模型对COVID-19病例人群活动要素进行解析;其次,在现有SEM通用事件表示模型的基础上,综合人群活动描述,设计了COVID-19病例知识图谱概念体系和本体规则,完成了模式层的构建;再次,对病例数据进行解析、事件实体识别和实体存储,完成了数据层的构建;最后,通过图数据库和B/S端构建原型系统进行实验验证。结果表明,通过COVID-19病例知识图谱可以整合病例语义关系和时空活动关联关系,提供扩展机制,具有活动事件概念表达清晰、类层次关系明确以及非层次关系丰富的优势。本文研究仅在COVID-19领域进行了实验,未来还需进一步探讨该图谱在其他传染病领域的适用性。

参 考 文 献

[1] Lloyd A L. Realistic Distributions of Infectious Periods in Epidemic Models: Changing Patterns of Persistence and Dynamics[J]. *Theoretical Population Biology*, 2001,60(18): 59-71

[2] Xu X J, Wang W X, Zhou T, et al. Geographical Effects on Epidemic Spreading in Scale-free Networks [J]. *Int J Mod Phys C*, 2006, 17(12):1 815-1 822

[3] Lipsitch M, Cohen T, Cooper B, et al. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome[J]. *Science*, 2003,300(5 627):1 966-1 970

[4] Riley S, Fraser C, Donnelly C A, et al. Transmission Dynamics of the Etiological Agent of SARS in Hong Kong: Impact of Public Health Interventions [J]. *Science*, 2003,300(5 627):1 961-1 966

[5] Leslie W D, Brunham R C. The Dynamics of HIV Spread: A Computer Simulation Model [J]. *Computers and Biomedical Research*, 1990,23(4):380-401

[6] Fuentes M A, Kuperman M N. Cellular Automata and Epidemiological Models with Spatial Dependence[J]. *Physica A*, 1999,267(3):471-486

[7] Eubank S. Scalable Efficient Epidemiological Simulation[C]. The 2002 ACM Symposium on Applied Computing,Madrid,Spanish, 2002

[8] Gong Lu, Liu Xiangnan, Zou Xinyu. Spread of Infectious Disease Risk Assessment Based on the Spatial-temporal Trajectory Data Analysis[J]. *Acta Geodaetica et Cartographica Sinica*, 2015,44(B12):6-12(宫路, 刘湘南, 邹信裕. 基于时空轨迹数据的传染病传播风险评估[J]. 测绘学报, 2015, 44(B12): 6-12)

[9] Feng Mingxiang, Fang Zhixiang, Lu Xiongbo, et al. Traffic Analysis Zone-Based Epidemic Estimation Approach of COVID-19 Based on Mobile Phone Data: An Example of Wuhan[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(5):651-657, 681(冯明翔,方志祥,路雄博,等. 交通分析区尺度上的COVID-19时空扩散推估方法:以武汉市为例[J]. 武汉大学学报·信息科学版, 2020,45(5):651-657, 681)

[10] Gao Shan, Wang Wenjun, Du Lei, et al. Research on Shared Ontology Model for Infectious Disease Emergency Case[J]. *Journal of Computer Applications*, 2010, 30(11):2 924-2 927(高珊, 王文俊, 杜磊, 等. 传染病应急案例共享本体模型研究

[J]. 计算机应用, 2010, 30(11):2 924-2 927)

[11] Chen Xiaohui, Wan Gang, Zhang Wei, et al. The Narrative Structure of GeoInt Visual Analysis[J]. *Journal of Geomatics Science and Technology*, 2017, 34(1):85-90(陈晓慧,万刚,张伟,等. 面向叙事结构的地理空间情报可视分析方法[J]. 测绘科学技术学报, 2017, 34(1):85-90)

[12] Peuquet D J. It's About Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems[J]. *Annals of the Association of American Geographers*, 1994,84(3):235-246

[13] Peuquet D J, Kraak M J. Geobrowsing: Creative Thinking and Knowledge Discovery Using Geographic Visualization [J]. *Information Visualization*, 2002, 1(1):80-91

[14] Orellana D, Renso C. Developing an Interactions Ontology for Characterizing Pedestrian Movement Behaviour[J]. *Movement-Aware Applications for Sustainable Mobility: Technologies and Approaches*, 2010,1(3):62-86

[15] Li Z Y, Ding X, Liu T, et al. Constructing Narrative Event Evolutionary Graph for Script Event Prediction[C]. IJCAI,Stockholm,Sweden,2018

[16] Hage W R V, Malaisé V, Segers R, et al. Design and Use of the Simple Event Model (SEM) [J]. *Social Science Electronic Publishing*, 2011, 9(2): 128-136

[17] Liu Junnan, Liu Haiyan, Chen Xiaohui, et al. Terrorism Event Model by Knowledge Graph[J]. *Geomatics and Information Science of Wuhan University*, 2020, DOI: 10.13203/j. whugis20190428 (刘俊楠,刘海砚,陈晓慧,等. 利用知识图谱的恐怖主义事件模型[J]. 武汉大学学报·信息科学版, 2020,DOI:10.13203/j.whugis20190428)

Construction of the COVID-19 Epidemic Cases Activity Knowledge Graph: A Case Study of Zhengzhou City

CHEN Xiaohui[1]　*LIU Junnan*[1]　*XU Li*[2]　*LI Jia*[1]　*ZHANG Wei*[1]　*LIU Haiyan*[1]

1　Institute of Data and Target Engineering, Information Engineering University, Zhengzhou 450001, China

2　Institute of Geospatial Information, Information Engineering University, Zhengzhou 450001, China

Abstract: At present, the number of coronavirus disease 2019 (COVID-19) cases worldwide is increasing, the spatio temporal spread of the epidemic becomes more and more complicated. The traditional researches on the transmission process is mainly focus on transmission trends of infectious diseases at the macro level. It is impossible to analyze the transmission relationship between specific cases at the individual level, to accurately locate the transmission paths of the epidemic, and it is difficult to support the precise prevention of infectious diseases. So, it is an urgent need to study the transmission process of infectious diseases on both of the semantic and spatio-temporal features. Based on the analysis of COVID-19 epidemic cases data, we construct the COVID-19 cases activity knowledge graph, which adapts to various description methods. Then, we design the ontology rules to complete the construction of the pattern layer. The epidemiological survey data has been analyzed to recognize the event entities, and to complete the construction of data layer. Finally, through the graph database and the B/S pattern to build a prototype system for experimental verification. The results show that it is effective and feasible to analyze the transmission process, infection relationship, key nodes and activity trajectory through the COVID-19 cases activity knowledge graph.

Key words: coronavirus disease 2019(COVID-19); epidemiological survey; activity knowledge graph; visualization of knowledge graph

First author: CHEN Xiaohui, PhD, associate professor, specializes in the theories and methods of spatial temporal data mining and knowledge graph. E-mail: cxh_vrlab@163.com

Corresponding author: LIU Haiyan, PhD, professor. E-mail: liu2000@vip.sina.com

Foundation support: The National Natural Science Foundation of China(41801313, 41901397).

引文格式: CHEN Xiaohui, LIU Junnan, XU Li, et al. Construction of the COVID-19 Epidemic Cases Activity Knowledge Graph: A Case Study of Zhengzhou City[J]. Geomatics and Information Science of Wuhan University, 2020, 45(6):816-825. DOI:10.13203/j.whugis20200201 (陈晓慧,刘俊楠,徐立,等. COVID-19病例活动知识图谱构建——以郑州市为例[J]. 武汉大学学报·信息科学版, 2020, 45(6):816-825. DOI:10.13203/j.whugis20200201)