doi:10.3772/j.issn.2095-915x.2023.01.008

基于多源数据融合的新冠肺炎病例活动知识图谱构建与知识发现研究

开放科学（资源服务）标识码（OSID）

刘桂锋 郭科远 包翔

江苏大学科技信息研究所 镇江 212013

摘要：[目的 / 意义] 面对复杂多变的疫情状况，为利用好多源异构的互联网数据资源，保证政府精准施策，保障人民生命安全。[方法 / 过程] 本文利用 Neo4j 图形数据库针对疫情期间病例活动轨迹数据实现了相关的知识图谱的构建与应用，通过 louvain 算法实现对确诊病例的社区划分，分析各个社区内部关系，深入挖掘时空关系，融合政策文本知识图谱用于辅助决策，再融合零散的医疗机构相关信息，生成 XY 市医疗机构知识图谱。[结果 / 结论] 本文借助知识图谱解决与疫情相关的多源异构数据融合问题，通过构建 COVID-19 病例活动知识图谱分析疫情形势及传播特点，并结合政策文本知识图谱实现了政府辅助决策服务，设计一种快速发现病例的方法。

关键词：COVID-19；知识图谱；多源异构数据；数据融合；知识组织；数据管理；数据科学

中图分类号：G35

Research on the Construction and Knowledge Discovery of COVID-19 Case Activity Knowledge Map Based on Multi source Data Fusion

LIU Guifeng GUO Keyuan BAO Xiang

Institute of Science and Technology Information, Jiangsu University, Zhenjiang 212013, China

Abstract: [Purpose/Significance] In the face of the complex and changeable epidemic situation, we should do a good job of tracing the source of flow in order to use many heterogeneous Internet data resources to ensure that the government accurately implements policies and ensures the safety of people's lives. [Methods/process] This paper uses neo4j graph database to construct and apply relevant knowledge graphs for case activity trajectory data during the epidemic period, realizes the community division of confirmed cases through louvain algorithm, analyzes the internal relationships of various communities, digs deep into the relationship between time and space, integrates the policy text knowledge graph, assists in decision-making, and then integrates scattered medical institution related information to generate a knowledge graph of medical institutions in city of XY. [Results/Conclusions] This paper solves the problem of multi-source heterogeneous data fusion related to the epidemic situation by

基金项目 国家社会科学基金一般项目“科学数据融合模式设计与体系建构研究”（21BTQ080）。

作者简介 刘桂锋 (1980-)，科技信息研究所所长 / 研究馆员，研究方向为科研数据管理、大数据分析、专利分析；郭科远 (1997-)，硕士研究生，研究方向为科研数据管理、大数据分析；包翔 (1991-)，馆员，研究方向为人工智能、数据挖掘、知识产权。

引用格式 刘桂锋 , 郭科远 , 包翔 . 基于多源数据融合的新冠肺炎病例活动知识图谱构建与知识发现研究 [J]. 情报工程 , 2023, 9(1): 102-117.

means of knowledge graph, analyzes the epidemic situation and transmission characteristics by constructing a knowledge graph of COVID-19 case activities, and realizes the government-assisted decision-making service in combination with the policy text knowledge graph, and designs a method to quickly find cases.

Keywords: Covid-19; Knowledge Graph; Multi-source heterogeneous data; data fusion; Knowledge organization; data management; data science

引言

新冠肺炎病毒不断变异，导致各地方疫情状况存在不确定性，对人民群众健康与社会经济发展构成严重威胁，面对当前依然复杂严峻的疫情防控形势，疫情防控工作成为各地方政府需要长期关注的重点。尤其是在对疫情期间病例活动轨迹数据以及对接触人群的流调数据的收集是支持疫情防控工作的重点，快速发现并隔离与病例间存在接触的人群已经成为一种有效的隔断疫情传播的方法。疫情也促使了各级政府加强数字政府建设的进程，数字技术在新冠肺炎疫情防控中发挥重要支撑作用[1]，疫情期间通过各类网络平台公布许多政策文本，而这些文本通常是大篇幅的文字，需要人工阅读筛选出针对各地不同疫情状况的相关政策，尤其在疫情期间，不利于工作效率的提高以及政策快速实施部署。同时，互联网还存在海量的有关疫情的文本数据，这些数据由于数据发布者的不同，分布在不同的网站，也造成数据的类型与特征存在差异，面对零散分布的数据，普通的数据整合方法很难解决数据融合问题，难以利用数据创造价值，也造成数据资源的浪费。基于以上背景，本文利用知识图谱技术分析病例活动数据展现规律，融合国家发布的有关疫情的政策文本，为政府疫情防控部门辅助决策提供一种多维可视化方法，利用 Neo4j 包含的 APOC 插件结合 intersection() 函数与 count() 函数设计一种快速发现可能存在感染概率的人群方法，把控病例间关联情况以及发现病例间传播路径，辅助政府疫情防控部门更科学准确的划分风险区域，减少封控对居民的影响，并利用数据融合技术实现医疗机构数据的融合，为居民就医提供帮助。

1 研究现状

为满足人们认知需求，Jeffrey Zeldman[2] 提出 Web 3.0，其中最重要的就是通过语义网等知识互联的方式将互联网本身转化为存储知识的数据库，从而提供更为智能化的服务。2006 年，BERNERS-LEE[3] 提出了关联数据（linked data）用 Web 技术以创建语义关联解决不同数据源的问题；2012 年 Google 公司首次提出知识图谱（Knowledge Graph）[4] 优化搜索引擎。利用知识图谱构建知识库的方式成为实现人工智能领域知识工程的重要手段。

当下国内针对 COVID-19 病例活动的研究主要分以下三种情况：一是利用知识图谱技术探究疫情传播规律，如陈晓慧等 [5] 通过构建 COVID-19 病例活动知识图谱，在个体层面分析具体病例之间的传播关系；李佳等 [6] 通过六

doi:10.3772/j.issn.2095-915x.2023.01.008

元组表示构建新冠肺炎病例活动事件图谱，进行可视化并加以分析，结合文本语义关系和时空特征挖掘新冠肺炎传播规律与发展趋势。二是结合地理信息系统分析疫情分布和传播情况，如蒋秉川[7]等利用 COVID - 19 确诊患者数据，通过地图分布可视化、图谱可视化和轨迹可视化等多视图协同交互分析 COVID - 19 疫情态势；应申等[8]利用五元组模型将病例数据结构化，并结合 GIS 空间分析技术以 COVID-19 流行病学调查数据为研究对象探究疫情分布和传播情况；郑良婷等[9]利用 ArcGIS 从时间、空间和扩散比三个维度对云南省 COVID-19 病例活动数据进行时空特征分析。三是从统计学角度出发，如曹莉等[10]通过绘制确诊病例时序传播图和活动轨迹表描述地区新型冠状病毒肺炎传播模式；王梓涵等[11]以云南省旅游输入型 COVID-19 病例为例探究病例空间分异及行为模式；刘勇等[12]从病例总量、输入性扩散性病例数量以及扩散比三个维度探究了河南省新冠肺炎疫情的时空扩散过程；张新等[13]对确诊和疑似病例诊疗时间记录数据采用分区统计等分析手段，研究了 COVID-19 疫情早期在诊断时间的时空分布、空间分异和动态过程；冯明翔等[14]利用手机用户空间交互数据提出 COVID-19 时空扩散推估方法。国外学者 Lipsitch[15]、Eubank 等[16]构建传播模型，研究特定传染病的传播规律；康大云等[17]探究空间与 COVID-19 的传播关系。此外，在利用知识图谱技术应用方面，张瑞[18]通过研究 COVID-19 相关研究文献，借助知识图谱技术从中识别出候选药物；Daniel Domingo-Fernández 等[19]学者则利用知识图谱技术从生理学角度探究 COVID-19 病毒病理。

综上所述，目前研究 COVID-19 病例活动数据主要是着眼于揭示病例间的时空分布规律，追溯疫情传播路径，或是利用研究文献数据，研究 COVID-19 病毒或药物，研究通常集中于疫情的某一方面的数据。本文不仅通过轨迹数据分析传播路径，还同时融合政策文本的数据，根据地区疫情发展状况，选取因地制宜的政策文本作为当地制订详细防疫措施的蓝本。

2 数据来源与理论方法

2.1 数据来源

流调数据是指流行病学调查，是疫情控制的关键，通过流调数据可以追踪到传染源，发现疾病传播途径，挖掘潜在的密切接触者，达到加强疫情防控的目的，由于流调数据涉及隐私，本文以 XY 市 2 月 15 日—2 月 22 日期间公开的 14 名病例的活动轨迹为例，通过“XY 发布”微信公众号获取 XY 市发布的第 61-72 号疫情防控通告，挖掘病例的社会关系以及行动轨迹，发现当地的疫情特征与传播路径，探究简单快速有效的方法，挖掘潜在的密切接触者，减轻流调人员工作负担。针对当前疫情防控形势和人群流行特征，为进一步指导重点场所、重点单位、重点人群做好防护，国务院应对新型冠状病毒感染的肺炎疫情联防联控工作机制颁布了《重点场所重点单位重点人群新冠肺炎疫情常态化防控相关防护指南（2021 年 8 月版）》[20]指导防控疫情工作。本文从中国政府网上获取该文献文本，以此为例，展示如何对政策文本深入挖掘与配合知识图谱技术实现与其他信息的融合应用。

2.2 研究思路与方法

由于来源于不同平台，轨迹数据、政策文本数据以及医疗机构数据相互孤立且存在异构问题，而通过知识图谱的知识表示流程解决轨迹与政策文本这两种非结构化异构数据的融合问题。比如轨迹与政策文本这两种都是非结构化文本，有用的信息都隐藏在文本中，文档式的连接，无法有效挖掘文本中所包含的语义信息，通过知识图谱统一的知识表示，利用 RDF 资源描述框架对轨迹与政策文本所涉及的实体与关系进行抽取与表示，可以将二者统一为三元组的知识表示形式，既可以解决数据异构问题，保证数据融合的进行，同时利用三元组这种小的单元，也加大融合程度，保证了知识图谱应用的灵活性。通过上述统一的知识表示，再通过知识融合与知识推理解决多源数据融合问题，轨迹数据中包含病例实体间大量的社交关系，通过知识推理可以补全实体之间隐含的关系，利用知识图谱处理轨迹数据获取到的现有显性知识以及社交关系关联规则补全社交图谱，比如，某小区超市，结合已知的病例居住地信息以及地理信息，推测超市所处具体位置，此外，利用已知社交关系发现与其他病例间关系等。通过预测图谱尚未存储的隐性知识，逐步完善现有病例活动轨迹知识图谱，尽可能补全语义与逻辑关系，保证新知识的发现。而知识融合技术，则应用到结合流调数据中包含的核酸检测点与就诊机构与医疗机构图谱所包含的采样点与医疗机构实体，从模式层与数据层进行融合，将相关知识图谱通过实体对齐与消歧的方式，通过将图谱间外部联系转化为图谱内部联系，使合并后的新图谱在保证消除冗余的同时，具有更大的价值。轨迹与政策文本本身来源于不同时期不同机构不同发布平台，二者本身是孤立存在的，二者也不存在同一实体，无法直接运用知识融合的方式达到合并两个图谱的目的，所以通过社群发现算法获得社群内关系特征，从而达到与政策文本图谱中的对应实体融合，解决了二者无法联系的问题。

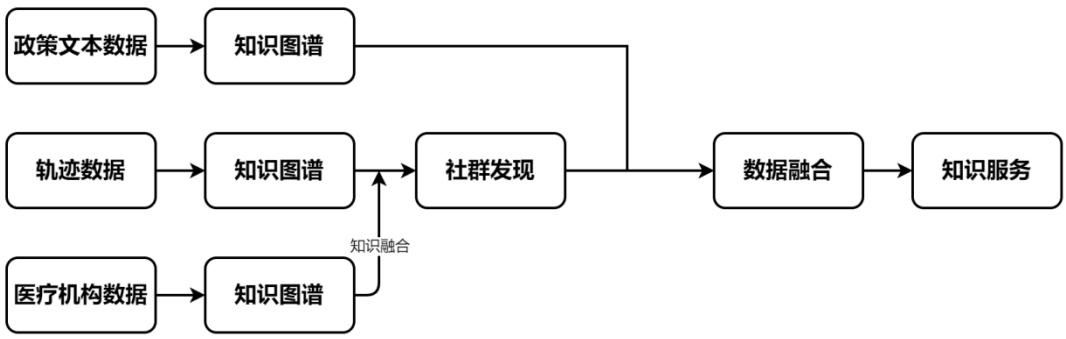

图 1　研究思路与方法

从应用层面上，借助知识图谱图结构特点，利用符号形式结构化优点挖掘文本中包含的隐形语义知识，提供一种高效挖掘快速发现病例间时空交叉关系方法，也为进一步针对政策文

doi:10.3772/j.issn.2095-915x.2023.01.008

本细粒度文本分析提供可能。

3 知识图谱的构建

首先是对数据的采集，由流程图 2 可见结构化数据主要来源于关系数据库，半结构化与非结构化数据主要来源于网络数据的采集，同时查询相关的知识库。以下几种途径获取有关此次疫情的结构化与非结构化数据：从“XY 发布”微信公众号获取政府公开发布的 XY 市疫情防控通告；国务院应对新型冠状病毒感染的肺炎疫情联防联控工作机制发布的《重点场所重点单位重点人群新冠肺炎疫情常态化防控相关防护指南》。其次是抽取收集到的文本中有用的信息，针对政府公开的有关患者轨迹的非结构化文本数据的知识抽取，主要包含实体识别、关系抽取、事件抽取等关键步骤。实体抽取主要抽取文本中的实体信息，比如患者经过的地名、接收的医院等；关系抽取主要是抽取实体间语义关系；事件抽取是从数据中抽取事件信息，并以结构化和语义化形式展现，比如发生时间、地点等。根据疫情传播特点，本文主要是从病患间的社会关系以及时空轨迹重叠两方面出发；一方面通过政府发布的病例轨迹文本抽取实体以及关系构建病患间的社会关系图谱，另一方面，根据时空轨迹重叠特点，利用事件的时间与地点作为事件抽取中事件的命名。根据政府公开发布的 XY 市疫情防控通告，同时也可以获取疫情期间医疗服务机构的有关信息。

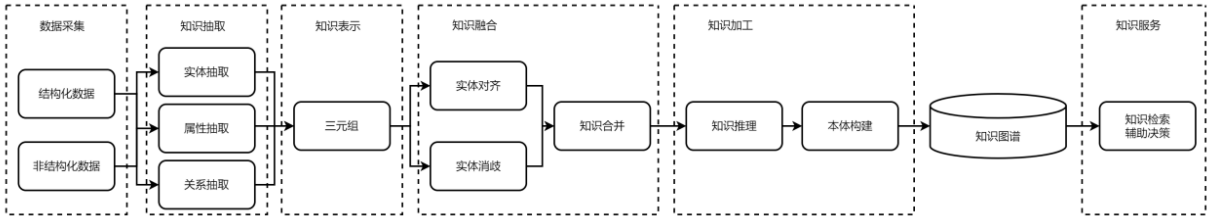

图 2 知识图谱构建流程

本文研究疫情的相关数据，疫情数据来源于不同平台，数据发布时可能存在差异，造成不同知识图谱间存在的异构问题，本文采用 RDF 三元组进行的知识表示，利用关系三元组与属性三元组，即实体—关系—实体，实体—属性—属性值的表述形式对知识进行描述，通过三元组的方式可以将非结构化文本转换为图模型，实现异构数据统一的知识表示，通过知识三元组，统一知识表示，缩小融合粒度以加大融合程度。通过知识图谱这种表达能力极强和应用灵活的语义网络，更容易实现对文本的深入挖掘。本文利用疫情防控通告中包含的语义关联关系，比如文本中描述的病例间的关系以及确诊前活动记录中所包含的病例时空信息，构建 COVID19 病例活动知识图谱。

接着进行知识融合，本文利用统一的实体命名方式，避免出现歧义问题，这种实体对齐方式符合知识图谱间进行融合的条件，保证了知识融合的进行。而知识融合的过程体现在多个知识图谱的融合对知识领域的扩展，扩大了知识应用服务范围，提升知识价值。本文主要从两方面考虑，一方面是从知识融合本身入手，通过多个图谱的融合，达到扩大知识覆盖的范围；另一方面，利用知识图谱图模型结构展示

数据以及关联关系，利用图算法挖掘已有规律，根据规律结果结合其他图谱通过图查询的方式实现新知识的发现。基本符合科研过程中对科研数据的使用过程。

Neo4j 是目前流行的一种图数据库，其使用的存储后端专门为图结构数据的存储和管理进行定制和优化，在图上互相关联的节点在数据库中的物理地址也指向彼此，因此更能发挥出图结构形式数据的优势 [21]，借助 Python 的 Neo4j 库 py2neo 结合 Neo4j 图数据库工具构建并存储国内医疗机构知识图谱，通过知识推理的方式补全知识图谱关系，完成知识加工，为知识服务提供数据支撑。

3.1 病例活动轨迹知识图谱

从“XY 发布”微信公众号获取 XY 市疫情防控通告文本，利用正则表达式结合词性与句法分析方式抽取病例类型实体、地址类型实体，同时结合事件的发生时间与地点，抽取病例的活动轨迹事件，考虑同一天同一地点就存在感染的风险，利用时间加地点的形式指代事件，再利用文本中的语义关联判断实体间存在的关系类型，确定病例 - 关系 - 病例，病例 - 住址 - 具体地址，病例 - 行程轨迹 - 活动轨迹事件三种关系，构建病例活动轨迹知识图谱（如图 3 所示）。由于知识图谱中的同一实体可能存在不同名称，为了保证知识融合质量，在针对地名的命名实体识别时，利用人工检查的方式，对相关地名统一标准化地址，保证后续针对活动轨迹时空交叉规律的探索，通过知识推理可以补全病例间关系，并发现病例潜在可能活动轨迹，后期结合与相关病例的接触关系与 GIS 地理信息系统，可以确定病例活动范围以及预测可能到访过的地区。

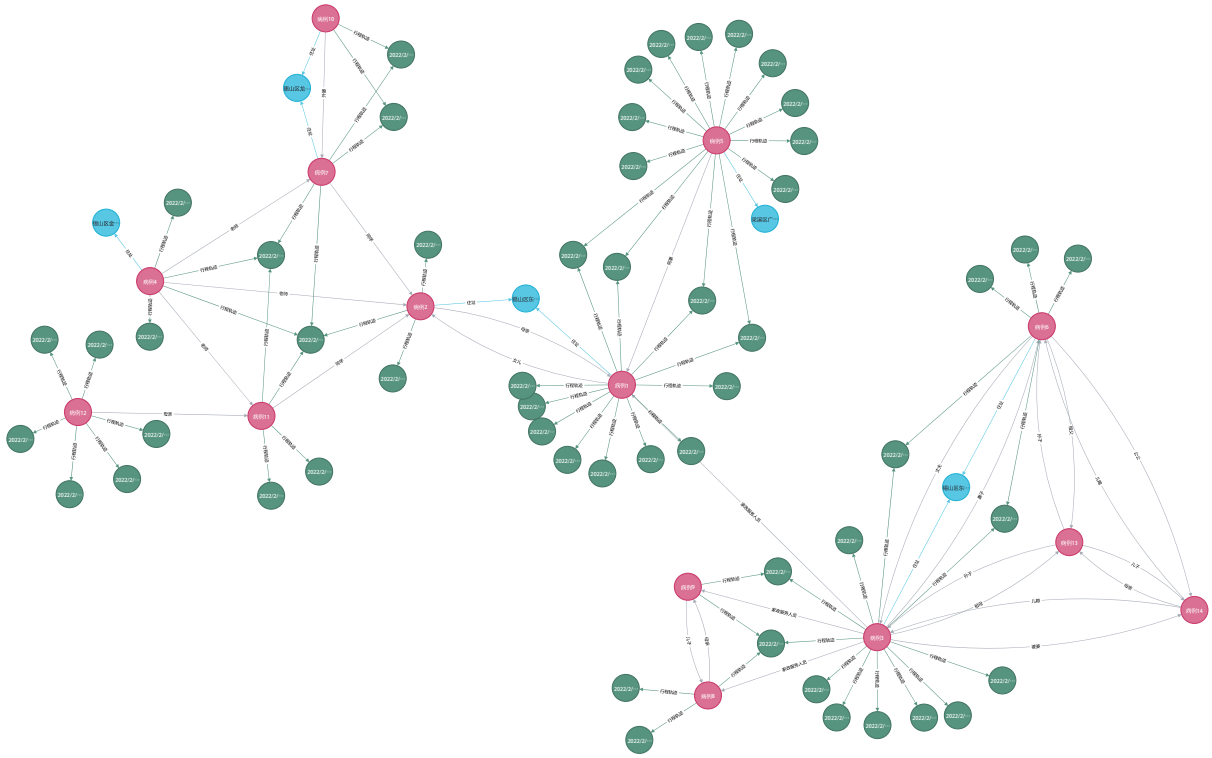

图 3 XY 市病例活动轨迹知识图谱

doi:10.3772/j.issn.2095-915x.2023.01.008

3.2 政策文本知识图谱

当前疫情形势依旧严峻，在抗击新冠肺炎疫情的防控工作中，也促使政府加强数字化建设，建设许多政策数据库，但存储仍以文档文本的方式，不利于文本的深度挖掘。非结构化文本数据的方式进行知识表示，也大大阻碍了政府工作人员对政策文本具体内容的查询检索，通常需要阅读大篇幅文本才能获取相应的政策建议，给人们对政策理解与执行造成了难度，不利于行政效率的提高。通过获取《重点场所重点单位重点人群新冠肺炎疫情常态化防控相关防护指南》的文本，针对文中对场所、单位、人群的政策建议，分成 Guide 指南、Publicplace 重点场所、Workplace 重点单位、Person 重点人群、Suggestion 建议几种实体类型，并依照 consist_of 与 suggest 两种关系，将 17 个单位、38 个场所、30 种重点人群共计 686 条建议做成的《重点场所重点单位重点人群新冠肺炎疫情常态化防控相关防护指南》政策建议知识图谱（如图 4 所示），构建对具体对象的政策语义网络，方便了政府工作人员检索针对具体实施对象的相关政策，提供相应的政策建议知识库辅助决策，也为深入挖掘政策文本的语义关系提供了可能。

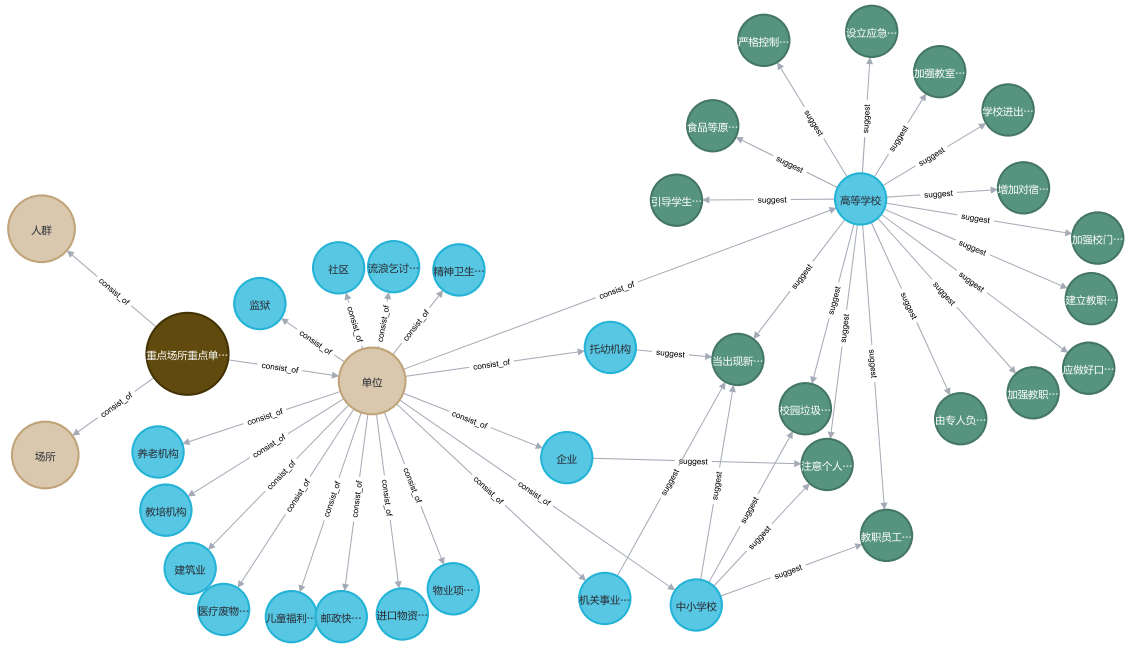

图 4 政策文本知识图谱（以高等学校为例）

4 知识发现与应用服务

4.1 辅助决策服务

社区（community）是由内部连接比较紧密的节点子集合构成的子图。社区结构作为网络的普遍特征之一，利用图网络拓扑方式更容易进行社区发现特点 [22]，本文采用社区发现（community detection）的方法，通过对病例与病例间关系对病例进行分类，通过社区内部关系与社区间关系，把握地区疫情特点与病例间传播路径。疫情防控通告 COVID19 病例活动数据不仅展示出了描述了病例间的亲属关系、

师生关系等社会接触关系，同时也记录确诊前活动事件所包含事件间关系。通过知识图谱技术手段将文本语义信息包含的社交关系转化为图模型结构，针对当前图结构模型利用社区发现算法探究病例间的传播关系。通过将图中病例实体节点划分到不同社区，保证社区内部关系紧密，社区外部关系稀疏。由于边上的权重默认为 1，模块度 Q 也可以简单理解为社区内部边减去所有与社区节点相连的边。而本文采用的 louvain 算法是一种基于模块度的社区发现算法。首先假设社区不重叠，将每一个病例节点视作一个独立社区，则存在与节点个数相等的社区；其次将每个节点尽可能分配到邻居节点的社区，计算分配前与分配后模块化指数增量 **ΔQ**，判断 **ΔQ** 是否大于 0，如果大于 0，选择 **ΔQ** 最大的邻居节点所在社区加入节点，否则，则保持不变；最后将划出的社区看作节点，循环以上步骤直到算法达到局部最优。依照 louvain 算法思路可以将病例间的社会网络关系划分为多个社区。通过观察图 5 可以看出，louvain 算法将病例划分为三个社区，分别涉及学校、工作单位以及家政服务。0 号社区的核心社会关系是师生关系；1 号社区的核心社会关系是企业中的同事关系；2 号社区比较复杂，主要涉及病例 3 的家庭关系以及雇佣关系。通过针对不同社区人群内部关系的分析，基本可以确定疫情防控工作的重点机构与重点人群，从社区 0 中可以得知重点机构中小学校、重点人群教师与学生，社区 1 中可以得知重点机构企业，社区 2 包含的重点人群保洁员。

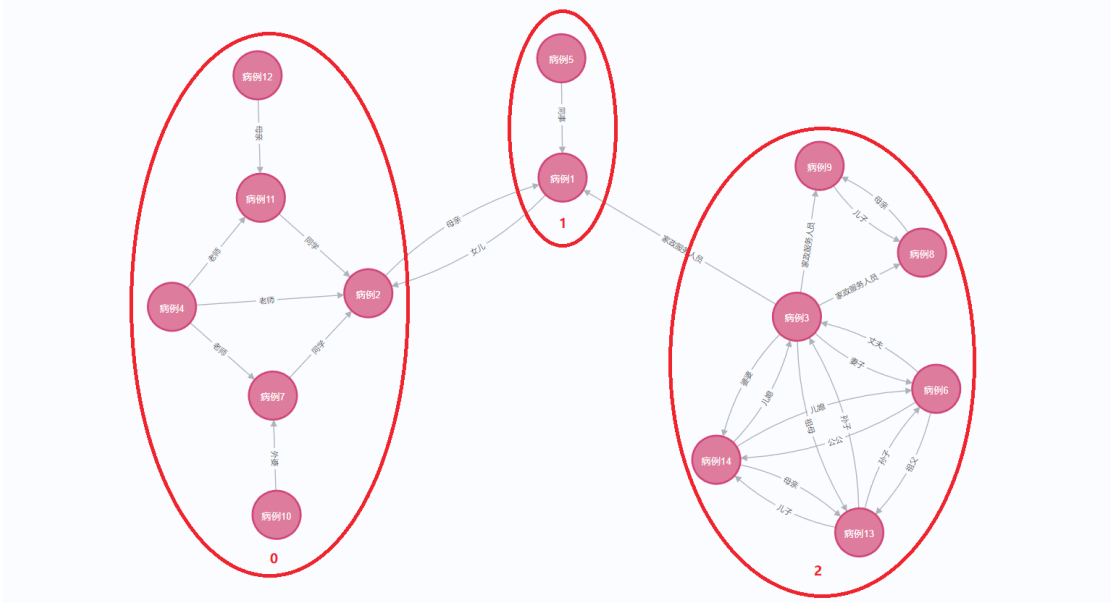

图 5 针对社交关系的社区发现

根据上述社区发现确定的疫情防控工作的重点人群，本文以《重点场所重点单位重点人群新冠肺炎疫情常态化防控相关防护指南》为例，确定政府适合状况的决策。利用《重点场所重点单位重点人群新冠肺炎疫情常态化防控相关防护指南》文本构建政策知识图谱，通过

doi:10.3772/j.issn.2095-915x.2023.01.008

上文分析的已知规律，例如第 0 社区中，包含的关系除普遍存在的家庭关系以外，主要是涉及师生关系，而师生关系是学校等教育机构关系的一种，根据文本中包含的地址信息，可以确认该社区涉及的政策文本主要针对的是机构实体中小学校、人物实体教师与学生，根据涉及的机构实体与人物实体，利用 Neo4j 的 Cypher 图查询语言在现有的《重点场所重点单位重点人群新冠肺炎疫情常态化防控相关防护指南》图谱中检索中小学校、教师以及学生实体，便可以得到针对该社区人群的政策文本（如图 5 所示）。利用 Neo4j 的 Cypher 图查询语言：

```
match(n)-[r]->(m)
where n.name=”中小学校” or n.name=”学生” or n.name=”教师”
return n,r,m
```

从图 6 中可以看出社区 0 针对当地疫情中包含学校以及师生关系方面的建议总计 31 条，涉及重点机构中小学校防控的政策建议 15 条，重点人群教师防控政策建议 8 条，学生防控政策建议 8 条，教师与学生存在相同防控建议条数。一方面体现了两者间的相关性，另一方面也展现了利用知识图谱可以有效解决数据冗余，减少数据存储上的空间资源浪费。图 7 展示了 neo4j 返回政策文本具体内容。

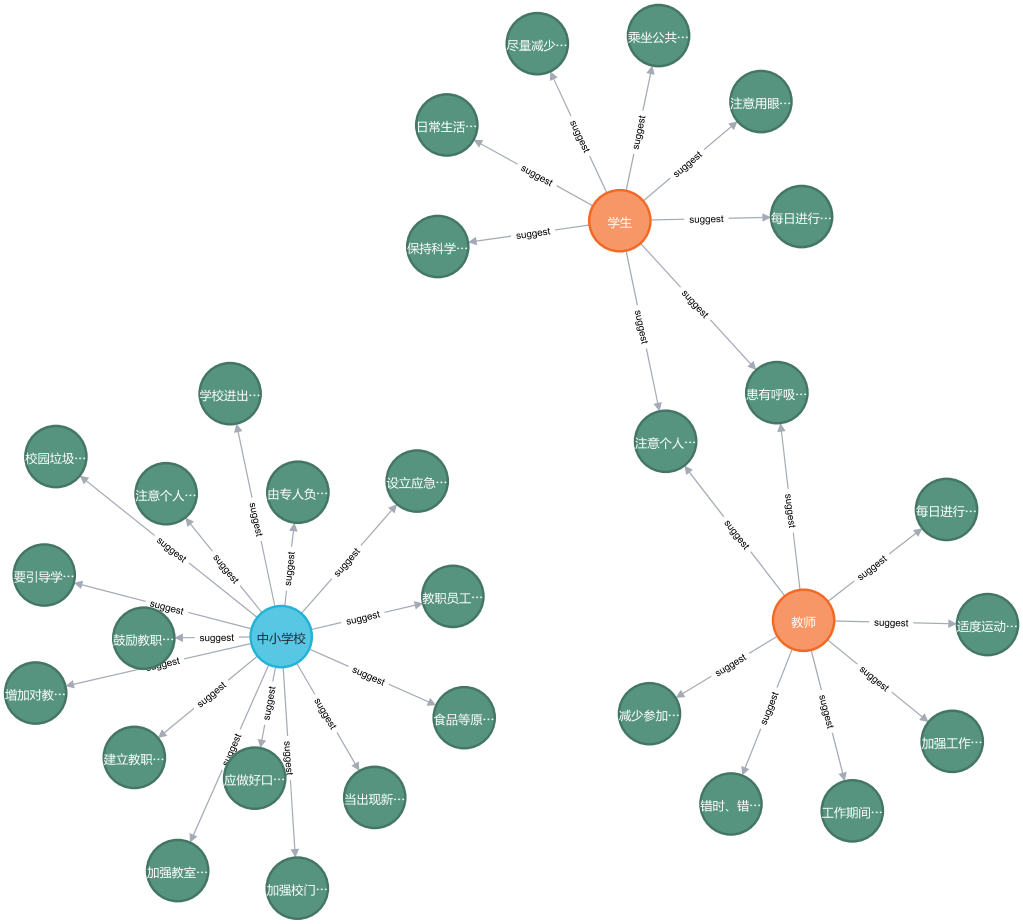

图 6　社区 0 中所涉及的政策文本检索结果图

图 7　社区 0 中所涉及的政策文本内容

同样针对社区 1 涉及的同事关系，属于企业关系的一种，可以对机构实体企业以及重点人群企业职工获取相应政策，利用 Neo4j 的 Cypher 图查询语言：“match(n)-[r]->(m)where n.name=” 企业” or n.name=” 企业职工” return n,r,m”进行查询，查询结果例如图 8 所示。

图 8　社区 1 中所涉及的政策文本检索结果图

获取到 11 条有关企业防控的政策建议，以及有关企业职工防控建议 7 条，政府就可以依照以上政策，有针对的根据企业目前疫情状况制订详细防疫措施。

doi:10.3772/j.issn.2095-915x.2023.01.008

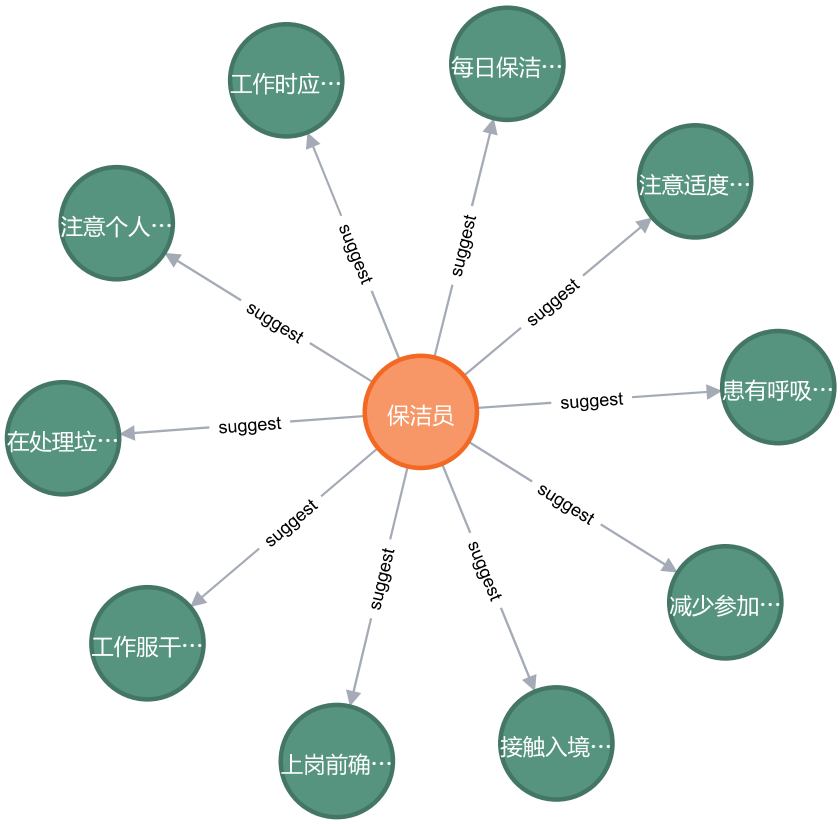

图 9 社区 2 中所涉及的政策文本检索结果图

同理如图 9，通过“match(n)-[r]->(m) where n.name=”保洁员” return n,r,m”有关社区 2 的涉及的重点人群保洁员 9 条相关建议。通过汇总去重，获取 50 条建议。政府可以根据不同人群的不同建议，有针对性的制定相关防控措施意见，同样也可以考虑汇总意见，针对本市疫情传播特点，制定适应本市疫情的防控政策。

4.2 时空数据挖掘

疫情期间存在大量的人员流动，而人员流动也是造成疫情传播的原因之一。因此，我国加强了对疫情人员流动数据收集，也就是各地方采集的流调数据。这些海量流调数据中包含了可能存在感染风险的人群，如何快速利用这些海量的记录人们活动的轨迹数据，筛选出与感染病人存在时空接触的人群成为分析流行病传播路线与溯源，有效遏制传播继续发展的重要手段。本文以利用病例的活动数据为例，提供一种快速发现病例间时空交叉关系的方法。目前，针对公共场所的消毒，基本是一天一次，时间跨度为 24 小时，而根据中国工程院院士李兰娟表示：新型冠状病毒在干燥的环境当中，存活时间也只有 48 小时。在公共场所正常消毒的条件下，那么表明在同一天同一地点就存在感染的风险，海量的流调数据不利于工作人员对存在感染风险人群的确认。由于无法获取流调数据，本文以病例的活动数据为例，展示如何快速发现病例间时空交叉关系方法。基于同一天同一地点就存在感染风险的思考，利用病

例活动事件图谱，便可以直观看到病例间的时空交叉关系，可随着人数的增加，会使病例活动事件图谱变得愈发复杂，很难从错综复杂的图谱中快速发现人群中的时空交叉关系（如图 10 所示）。Neo4j 中 APOC 插件中的 intersection() 函数与 count() 函数结合可以解决以上问题，intersection() 函数是求 A 与 B 两集合间的交集，再利用 count() 函数统计交集大小，而放在探究时空交叉关系的情景下，则考虑 A 病例的活动轨迹集合与 B 病例的活动轨迹集合的交集大小，当然这里 A、B 可以不仅仅是病例，还可以是密接人群，这样就可以快速定位存在感染风险的人群。而这种存在感染风险人群里也可以通过接触次数划分出需要重点关注的人群。

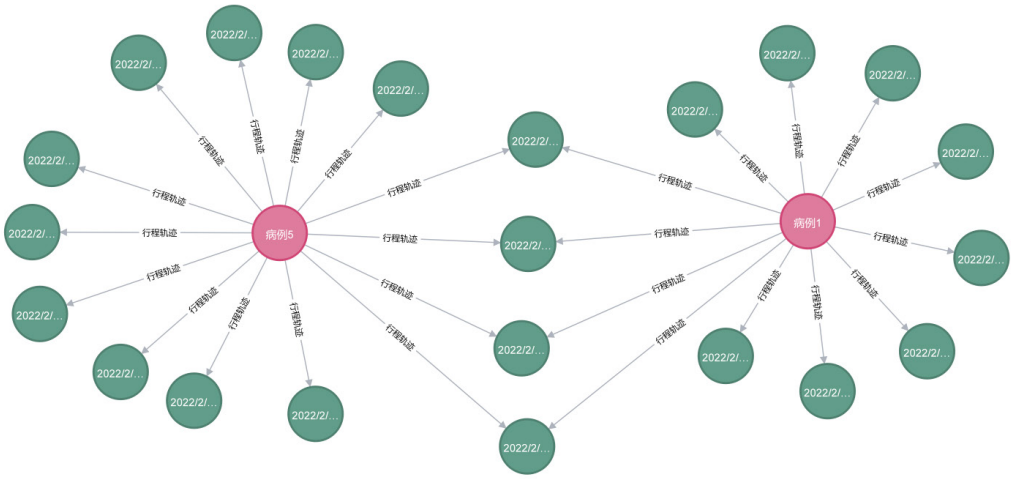

图 10 病例间时空交叉关系

表 1 快速发现病例间轨迹交集

from	to	intersection
病例7	病例10	2
病例4	病例7	2
病例4	病例11	2
病例7	病例11	2
病例8	病例9	1
病例3	病例9	2
病例1	病例5	4
病例2	病例4	1
病例2	病例7	1
病例2	病例11	1
病例3	病例6	2
病例3	病例8	1

如表 1 所示，利用 intersection() 函数快速发现病例间轨迹交集，再用 count() 函数统计病例间轨迹交集的元素个数，表中的 intersection 就是指二者轨迹事件的交叉次数，intersection 次数可以反应本文要挖掘时空交叉关系，如果可以采用流调数据的话，一般无同一时间同一地点上接触，则 intersection 为 0，否则 intersection 越大，表明接触次数越多，感染风险就越大。比如如果要查询病例 1 与病例 5 病例间交叉轨迹：

```
match (a:person)-[:`行程轨迹`]->(b1:track),(c:person)-[:`行程轨迹`]->(b2:track)
where a.name=”病例 1” and c.name=”病例 5”
RETURN apoc.coll.intersection([b1],[b2]) AS output
```

结果如下图 11：

doi:10.3772/j.issn.2095-915x.2023.01.008

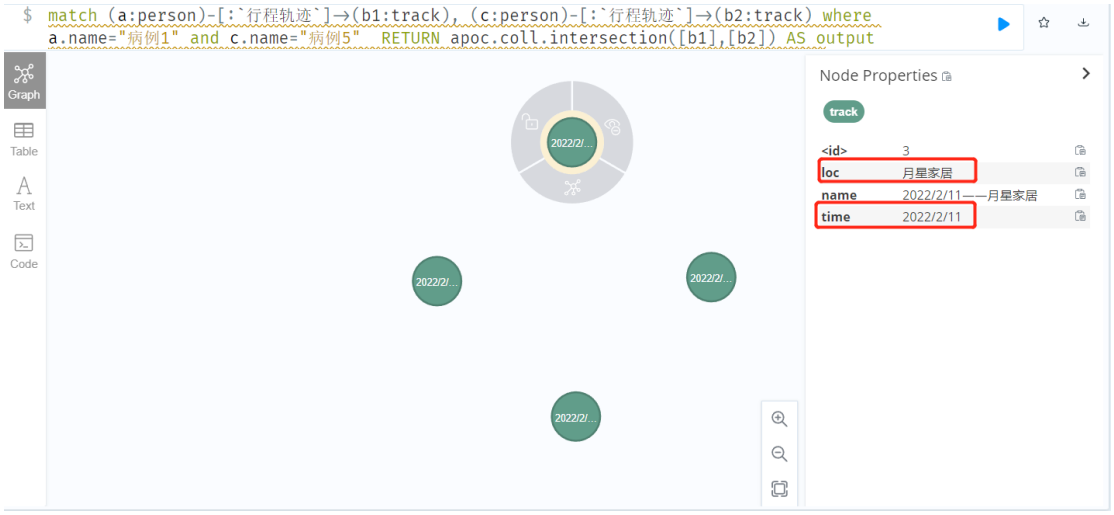

图 11 病例间交叉轨迹查询

如图 11 所示，可以通过 intersection() 函数结合图谱确定交叉轨迹的时间与地点，为政府疫情防控部门准确把控病例间关联情况，发现疫情传播路径，更科学精准划分风险区域，减少了疫情封控对居民出行的影响。

4.3 医疗机构数据融合

从 XY 发布微信在 XY 市疫情防控通告后附的XY市各市县区疫情防控指挥部联系电话、XY 市 24 小时核酸检测医疗机构信息以及 XY 市开诊发热门诊医疗机构名单等信息，可以通过知识图谱实现碎片化信息的融合，构成医疗机构知识图谱（如图 13 所示）。与通告非结构化文本内容不同，通告后附三副图片包含内容是以表格的形式存在的，这就方便了对实体信息的抽取。这种类似二维表结构的结构化数据，一般采用将表对应图谱中的类，列对应图谱中的属性，行对应图谱中的实例，而每一个单元则是属性值，可以直接映射将输入的数据表输出为三元组，依照抽取类，抽取属性，抽取实例以及建立类间关系，将 XY 市各市县区疫情防控指挥部联系电话中包含信息转换为县区 - 联系方式 - 属性值（具体电话）类型的三元组，XY 市 24 小时核酸检测医疗机构信息包含信息转换为县区—核算检测医院—医疗机构、医疗机构 - 地址 - 详细地址、医疗机构 - 采样点 - 采样地点类型的三元组，XY 市开诊发热门诊医疗机构名单中包含的医疗机构 - 发热门诊 - 详细地址类型的三元组，通过以上三元组融合后，构建医疗机构图谱的模式层，如图 12 所示，再将具体数据导入，生成图谱的数据层。通过这种方式完成多个结构化二维表数据的融合，一方

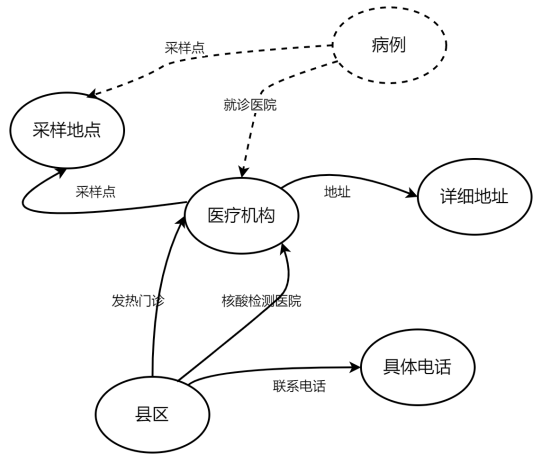

图 12 流调数据与医疗机构知识图谱模式层融合方法

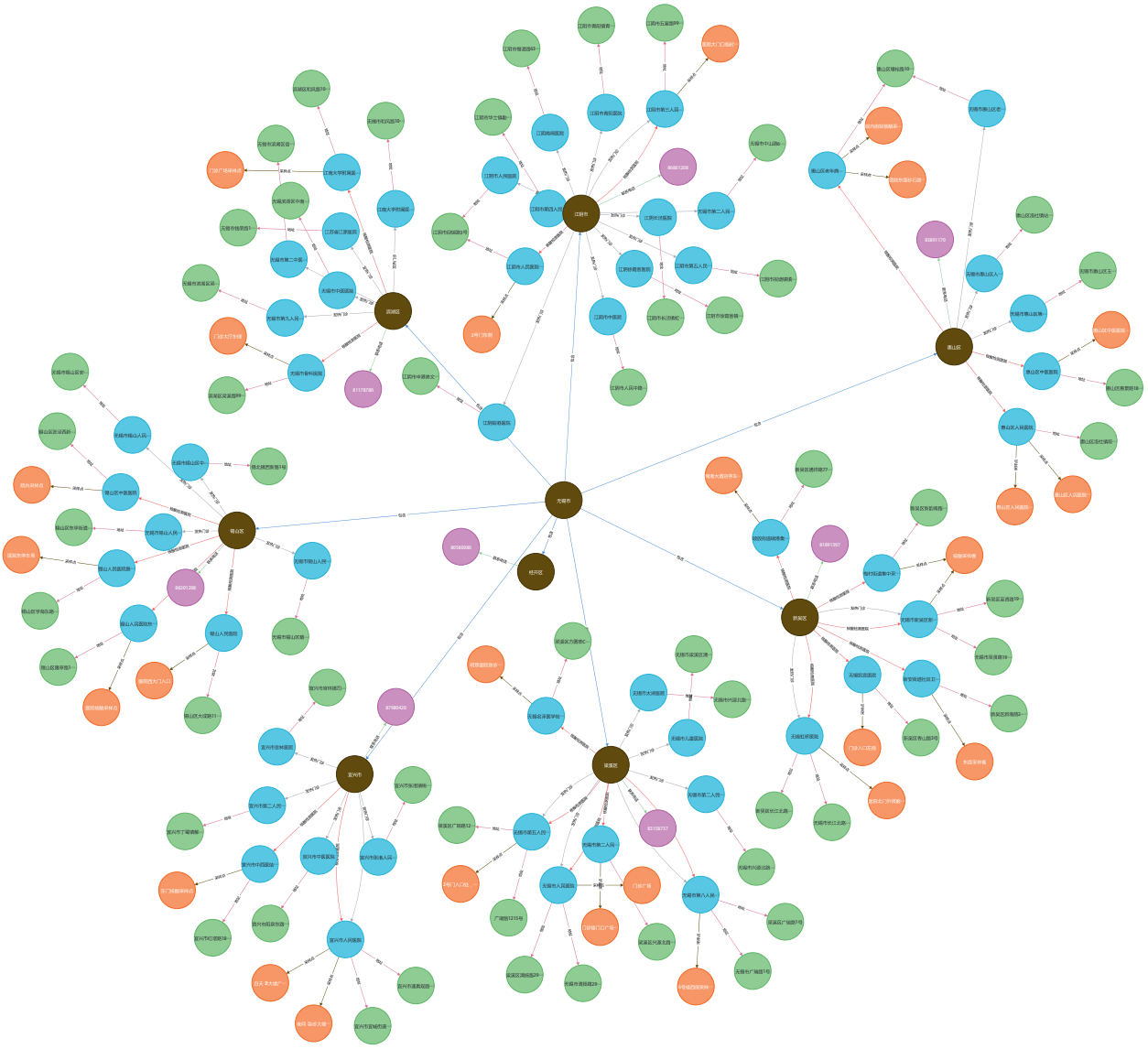

图 13 医疗机构知识图谱

面去除了冗余数据，另一方面也体现了同一领域的知识扩展，为建立医疗机构信息查询服务以及结合流调数据分析就诊医院奠定了数据基础。

4.4 小结

知识图谱作为人工智能底层技术之一，具有图的各项特征，相比于复杂文本，有着更为直观表现和形式，适用于各种图挖掘算法，此外，知识图谱通过对实体关系的抽取，包含原有文本中关联以及语义信息，支撑数据分析与解读。本文利用知识图谱提供针对多源异构数据融合方法，一方面利用知识图谱结合社区发现算法挖掘轨迹数据的社区规律（新知识），再结合政策文本所抽取的知识图谱，提供辅助决策等知识服务。另一方面通过对医疗机构与病患流调数据中存在的就诊记录利用知识图谱的知识融合、知识推理等技术实现两图谱间的融合，进而提供信息查询服务等知识服务。从疫情防

doi:10.3772/j.issn.2095-915x.2023.01.008

控角度而言，通过对相关数据知识图谱构建与融合，向政府部门直观展示了疫情防控措施的关联情况，为政府疫情防控部门全面掌握疫情防控政策，针对特点人群、地点准确运用防控措施提供帮助。后期可以结合本文建立的图谱作为背景知识设计问答系统以提高防控工作效率，同时也可以为居民提供外出就诊等医疗信息服务。

5 结语与展望

本文以 COVID19 相关知识图谱的构建为例，从科研数据使用的两种情况出发，一方面利用知识图谱通过分析病例的活动轨迹，挖掘病例间的时空关系，通过社区发现 louvain 算法实现对确诊病例的社区划分以及社区内部关系规律的发现，通过划分人群以及特点分析，结合政策文本的知识图谱，根据当地疫情防控地重点场所重点单位重点人群，找出对应的防控疫情政策建议。这一过程所利用的活动轨迹数据与政策文本数据做不到具体的关联，但由活动轨迹数据所构建知识图谱揭示出的规律，却可以与政策文本数据关联，实现辅助决策的应用，产生二者单独都无法产生的价值。另一方面，就是存在关联关系，如医疗机构数据融合，可以通过具体存在的相关实体实现多源异构数据融合。为应用知识图谱提高多源异构数据融合的程度提供较好的范例，更好的挖掘文本具体内容。

本文通过研究知识图谱与多源异构数据关系，发现知识图谱中的知识表示、知识融合与知识推理技术在多源异构数据融合过程的主要功能，利用知识表示统一数据描述方式，解决数据异构问题，利用知识融合可以对多源异构数据进行关联和发现，进而得到融合后的新知识或新的解决方案。知识推理利用知识图谱现有的显性知识来预测图谱中尚未存在的隐性知识，并逐步将知识图谱补充完整。

由于针对政策建议的知识表示仅仅停留在建议文本一层，未能从语义层面更好地挖掘政策内容，许多建议的相似度更高，在未来针对建议文本进行深入挖掘。仅仅从理论的角度出发，采用的样本量较小，还未能在大规模使用中检验效果。

参 考 文 献

[1] 中国政府网 . 国务院关于加强数字政府建设的指导意见 [EB/OL]. [2022-07-15]. http://www.gov.cn/zhengce/content/2022-06/23/content_5697299.htm

[2] Sheth A, Thirunarayan K. Semantics Empowered Web 3.0:Managing Enterprise, Social, Sensor, and Cloud-Based Data and Services for Advanced Applications. Vermont: Morgan &Claypool, 2012

[3] Berners-Lee T, Hendler J, Lassila O. The semantic Web[J]. Scientific American Magazine, 2008, 23(1): 1-4.

[4] AMIT S. Introducing the knowledge graph[R]. America: Official Blog of Google, 2012.

[5] 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.

[6] 李佳 , 刘海砚 , 刘俊楠 , 等 . 基于知识图谱的新冠肺炎病例传播关系可视分析 [J]. 信息工程大学学报 , 2021, 22(5): 606-612.

[7] 蒋秉川 , 游雄 , 李科 , 等 . 利用地理知识图谱的 COVID-19 疫情态势交互式可视分析 [J]. 武汉大学学报 (信息科学版), 2020, 45(6): 836-845.

[8] 应申 , 徐雅洁 , 窦小影 , 等 . 地理位置关联的 COVID-19 传播时空分析 [J]. 武汉大学学报 (信息科学版), 2020, 45(6): 798-807.

[9] 郑良婷 , 胡文英 , 胡斌 , 等 . 云南省新型冠状病毒肺炎的时空特征分析 [J]. 中国卫生统计 , 2022, 39(2): 176-179, 185.

[10] 曹莉 , 周永江 , 张帆 , 等 . 确诊病例时序传播图及活动轨迹表在新型冠状病毒肺炎疫情分析中的作用 [J]. 中华流行病学杂志 , 2020, 41(11): 1782-1785.

[11] 王梓涵 , 陈方 , 戢晓峰 , 等 . 旅游输入型传染病病例的空间分异及行为模式——基于云南省 COVID-19 病例的实证分析 [J]. 昆明理工大学学报 (社会科学版), 2020, 20(5): 72-78.

[12] 刘勇 , 杨东阳 , 董冠鹏 , 等 . 河南省新冠肺炎疫情时空扩散特征与人口流动风险评估——基于 1243 例病例报告的分析 [J]. 经济地理 , 2020, 40(3): 24-32.

[13] 张新 , 林晖 , 朱长明 , 等 .COVID-19 疫情早期中国确诊时间的时空特征及动态过程分析 [J]. 武汉大学学报 (信息科学版), 2020, 45(6): 791-797.

[14] 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.

[15] Lipsitch M, Cohen T, Cooper B, et al. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome[J]. Science, 2003, 300(5 627): 1 966-1970

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

[17] Kang D, Choi H, Kim JH, et al.Spatial epidemic dynamics of the COVID-19 outbreak in China. International Journal of Infectious Diseases, 2020, 94(3): 96-102.

[18] Zhang R, Hristovski D, Schutte D, et al. Drug repurposing for COVID-19 via knowledge graph completion[J]. Journal of biomedical informatics, 2021, 115: 103696.

[19] Domingo-Fernández D, Baksi S, Schultz B. COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics, btaa834[J]. 2020.

[20] 中国政府网 . 关于印发重点场所重点单位重点人群新冠肺炎疫情常态化防控相关防护指南（2021 年 8 月版）的通知 .[EB/OL]. [2022-07-15]. http://www.gov.cn/xinwen/2021-08/13/content_5631094.htm

[21] weixin_39621774.neo4j 知识图谱 _ 知识图谱里的知识存储：neo4j 的介绍和使用 .[EB/OL]. [2022-07-15]. https://blog.csdn.net/weixin_39621774/article/details/111281048

[22] Bandyopadhyay S, Vivek S V, Murty M N.Outlier resistant unsupervised deep architectures for attributed network embedding[C]//Proc of the 13th International Conference on Web Search and Data Mining, 2020:25-33.