·论　　著·

基于图数据库的新冠感染图谱建模与分析*

聂　鑫，蹇奕苹△，王显科，张盛杰
（重庆市卫生健康统计信息中心，重庆 401120）

［摘　要］ 2021－2022 年，新型冠状病毒感染（COVID-19）疫情仍在全球蔓延，接触者追踪是当时疫情应急管理的重要策略。然而，传统的接触者追踪在实践中面临许多限制。数字技术的应用为地方政府提供了一个更全面、更高效、更准确地追踪个人与 COVID-19 接触者的机会。该文建立了基于 Node4j 图数据库的疫情传播预测系统，通过收集确诊和疑似病患者时空位置数据，综合运用图模型和相关数据库，构建适应多样化描述方式的病例活动轨迹知识图谱，快速定位高危人群群体，给予政府和主管部门决策辅助支持，对于提高疫情防控效率具有重要意义。

［关键词］ 疫情防控；　Node4j 图数据库；　传播过程可视化；　建模
DOI：10.3969/j.issn.1009-5519.2023.11.004　　**中图法分类号**：R181.8
文章编号：1009-5519(2023)11-1818-04　　**文献标识码**：A

Mapping modeling and analysis of COVID-19 outbreak based on graph database*

NIE Xin, *JIAN Yiping*△, *WANG Xianke*, *ZHANG Shengjie*
(*Chongqing Health Statistics Information Center*, *Chongqing* 401120, *China*)

［**Abstract**］ 2021－2022, as the Corona Virus Disease 2019 (COVID-19) continues to spread globally, contact tracing is an important strategy for emergency management at that time. However, traditional contact tracing faces many limitations in practice. The adoption of digital technologies provides local governments with an opportunity to track individuals' COVID-19 contacts more comprehensively, efficiently and accurately. In this paper, a prediction system of epidemic transmission based on Node4j graph database was established. By collecting the spatio-temporal location data of confirmed and suspected patients, the graph model and related databases were integrated to construct the knowledge map of case activity trajectory adapted to the diversified description methods, so as to rapidly locate high-risk population groups and provide decision-making support to the government and competent authorities. It is of great significance to improve the efficiency of epidemic prevention and control.

［**Key words**］ Epidemic prevention and control;　Node4j graph database;　Visualization of propagation process;　Modeling

2021－2022 年，新型冠状病毒感染（COVID-19）仍在全球范围内迅速蔓延，数以千万计的人被感染，而且感染人数还在快速增长，同时，疫情大流行给全球经济带来严重影响[1]。当时，大多数国家或地区仍处于公共卫生紧急状态，尽管中国、美国等国家已经成功研制出 COVID-19 疫苗，但疫苗的生产和应用仍存在较大差距。此外，当时，尚未成功开发出治疗 COVID-19 的有效药物[2]。一些医学专家认为，COVID-19 可能会发展为季节性流行病，类似于流感，这对传统的接触者追踪提出了巨大挑战[3]。因此，快速识别和追踪感染 COVID-19 的个人及其接触者，并采取积极的应急管理措施，如旅行限制、健康监测和居家隔离，是当时所有国家和地区战胜 COVID-19 疫情所必需的策略[4]。

数字接触者追踪已在全球范围内推出，作为遏制 COVID-19 大流行的工具[5]，国内外部分研究从时空演变的角度对传染病的传播进行了分析[6-8]，据历史传染病防控经验表明，通过研究传染病的传播规律和传播途径，可有效地对其进行控制[9]。传染病医疗卫生信息系统大多采用传统的关系模型，且大量的调查数据一般使用手工录入，不仅效率低，更不利于关联性的快速检索和分析[10]。图数据库是一种基于图论和算法的新型数据库系统，能够有效地处理复杂的关系网络。为此，作者设计了一种能高效处理复杂关系

*　**基金项目**：重庆市新冠肺炎防控应急科研专项（2020NCPZX02）。
作者简介：聂鑫（1978－），硕士研究生，大数据智能化工程师，主要从事卫生信息化建设、管理；智慧医院建设、评审工作。　△　**通信作者**，E-mail：95687340@qq.com。

网络的图数据库算法，该算法依靠政府大数据平台分析 COVID-19 流行病数据，并在高危感染者、普通人群、车辆和公共场所之间建立关系网络，以识别和追踪接触者。本研究使用 Neo4j3.4.15 版本的 Cypher 算法构建了高危人群、普通人群、车辆、公共场所等关键信息之间的关联图，揭示隐藏的关系网络，并识别关系中的风险，以识别和追踪 COVID-19 病例的接触者，将数据分析结果可视化，使数据更直观。

1　资料与方法

1.1　资料　快速实现逐行访问是关系型数据库的设计原则之一。当数据之间出现复杂的关联时，跨表的关联查询增多，就会出现问题。图数据库是一种基于图论和算法的新型数据库系统，能够有效地处理复杂的关系网络[11]。与关系数据库相比，图数据库还将关系映射到数据结构中，这对于查询高相关性的数据集来说速度更快，特别适合那些面向对象的应用工具[12]。同时，图数据库可以更自然地扩展到大数据应用场景。因为图数据库模式更灵活，更适合管理临时或变化的数据。关系数据库和图数据库主要区别如下：(1)数据规模问题，图数据库的实现可以基于 KV 存储，可以高效方便地存储亿或十亿级别的数据；关系型数据库在此规模下，必须进行复杂的分库分表设计，否则根本无法胜任。(2)复杂关系查询能力，图数据库以实体和关系为基本单位，特别适合查询和分析多层次、多样的复杂关系；关系数据库则在复杂关系查询方面不堪重负，尤其是涉及多表关联或者递归查询时。(3)查询语言不同，图数据均有配套的查询语言，比如 Gremlin、Cypher，以更为贴合自然语言的方式限定查询条件，易表达查询需求；关系型数据库使用 SQL 查询语言，同样抽象层次较高，尤其对于多层关系的查询(需要 join 操作)时，语句非常复杂且效率低。(4)建模方式不同，图基于现实世界的实体和关系建模，更直接易懂；传统关系数据库建模需要的抽象层次更高，也更复杂。

1.2　方法

1.2.1　效率分析　以关系数据库 MySQL、图数据 Neo4j 分别对 150 万个数据节点，350 万条边数据量的不同深度遍历进行了测试比较，在遍历查询深度为 1、2 层时，关系型数据库和图数据库均能在很短时间内完成遍历任务。当遍历深度提高到 3、4 层，关系型数据库查询时间呈指数级增长，然而图数据库却仍然能在较短的时间内完成查询任务。

1.2.2　患者传播过程研究　通过提取病例特征数据，使用 HanLP 语义识别工具对传播过程进行识别，然后使用 Neo4j 图数据进行病例活动图谱构建。

1.2.3　数据提取　由于患者数据信息复杂多元，属于非结构化文本数据，主要包括患者的基本情况、患者活动记录、患者社会关系、当前状态和疾病传播路径等，涉及 3 种数据类型：时序数据、地点空间数据、关联数据。

1.2.3.1　时序数据　时序数据是按时间组织的一组值，使用时序数据来回顾并度量变化，或者展望并预测未来的变化，通常按时间顺序到达插入数据存储，较少更新。相比之下，标准联机事务处理 (OLTP) 数据管道就可接受任何顺序的数据，可随时更新。本研究将病例经过的地点在时间轴上按照顺序进行排列，就可生成实体的时序数据。

1.2.3.2　地点空间数据　地点空间数据记录患者活动的地点名称及地点坐标信息，主要以 POI 形式存储，包含名称、地址、坐标、类别 4 个属性。

1.2.3.3　关联数据　患者数据包含详细的信息，如传染病史、病历数据、亲属关系、婚姻情况等，看似不相关的数据可以通过语义分析进行关联。本研究的样本数据来源包括 12320 热线信息、医疗机构预约挂号信息、人口家庭关系信息、核酸检测数据等。样本数据获取来源。

表 1　样本数据获取来源清单

数据库	数据来源	收集方法
高危人群关系数据库	重庆市卫生健康统计信息中心	通过核酸数据库中的高危人群与人口信息及家庭关系数据库进行比对建模
高风险地区数据	官方信息发布渠道及市卫生健康统计信息中心	由各个官方发布渠道及 12320 健康热线和预约挂号数据库提供
居民轨迹信息库	重庆市卫生健康委员会	由市卫生健康委数据交换平台提供
手机信令数据库	通信运营商	中国移动、中国联通、中国电信、中国广电提供

1.2.4　患者传染过程识别　COVID-19 患者的传播过程与患者周边的人群活动事件紧密关联，患者及与其有时空接触的人群的活动决定了 COVID-19 的时空扩散范围。人群活动是由一系列事件构成，主要包括是什么人(Who)、什么时候(When)、在哪里以(Where)及干什么(What)的组合。本研究采用简单事件模型来模拟病例的传播过程，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。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。社会关系、传染关系、活动事件关系之间的关联见图 1。

在患者实体关联关系基础上，本研究定义了 4 种传染事件语义关系实例：组成关系、因果关系、跟随关系、并发关系，用于描述病毒在患者间是如何传播的。

流行病学调查数据虽然是一种非结构化文本数据，只要制定规范的数据描述标准，清晰准确描述事件，进而可以通过自然语言处理，对病例数据进行语义分析处理。本研究使用 HanLP 对患者文本数据进行切割分词，分为 5 种类型：时间、地点、事件、参与者和对象，处理结果作为知识表示的基础，并识别出相应的标签，识别出的标签可以以 csv 格式导出，最后导入到图数据库中。

1.2.5 患者活动信息图构建　本研究以确诊、疑似、无症状感染者及有居住史的个人为人口节点；以私家车、火车、航班为交通枢纽；社区和商场作为公共场所的节点。技术人员将关系定义为同时出现在同一公共场所或车辆中。基于这些节点和关系，应用 Neo4j Cypher 算法图数据库构建关联图，分析和展示重点人群、车辆、公共场所等关键信息之间的关联。本研究中使用的图数据库的核心算法见表 2。

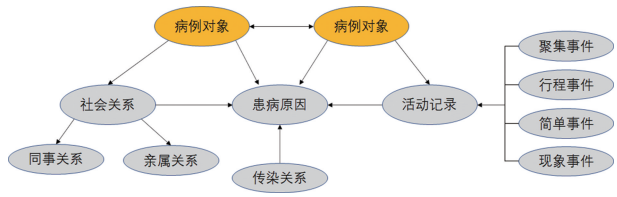

图 1　患者实体关联关系设计

表 2　基于图数据库的数字接触者追踪核心算法

算法功能	具体算法	算法说明
追踪确诊、疑似和无症状感染者的旅伴	MATCH p = (n: Individual)-[r:'sameTransportation']-(n1)-[rr:'sameTransportation']-(n2) where n.name>n2.name and upper(n.ID) in ['ID1','ID2','ID3','ID4','ID...'] 返回不同的 n.name, upper(n.ID), n.phoneNo, n1.name, upper(n1.ID), n1.phoneNo, n2.name, upper(n2.ID), n2.phoneNo;	“个人”是指分析中包含的个人；“sameTransportation”是指同时采取一种运输方式；“名称”是指被分析对象的名称；“ID”为身份证号，其中“ID/ID1/ID2/ID3/ID4/ID...”为确诊、疑似和无症状感染者的身份证号；“phoneNo”为手机号码。
追踪与 2 名确诊感染者有过接触的人	MATCH p=(n:'个人')-[r1]->()-[r2]-(nm)-[r21]->()-[r22]-(m:'个人') 其中 n.is_confirmedindividual 在 ['yes','sameID'] 和 m.is_confirmedindividual 在 ['yes','sameID'] 和 EXISTS(nm.is_confirmedindividual)=false 和 n.ID>m.ID 返回 p;	“已确认的个人”是指已确认的个人；“sameID”表示相同的 ID 号。
通过高感染风险的交通方式追踪接触者	匹配 p=(n:Transportation)-[*..4]-() 其中 n.name=\'CarNo1\' return p;	“交通工具”是指分析中包含的交通工具（私家车、火车、飞机）；“CarNo1”是某辆私家车的车牌号。

1.2.6 可视化的交互操作系统设计　为更好地服务于疫情防控，本研究进一步设计了疫情可视化监控系统。结合功能需求，系统设计了四大功能模块，分别是数据获取、数据存储、数据分析、前台页面。通过四个功能模块组成系统，以实现疫情可视化监控，系统总体功能设计见图 2。

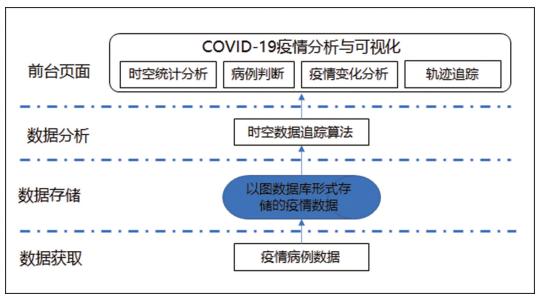

图 2　地理位置关联的 COVID-19 传播时空分析框架

利用 ECharts 将基础数据和分析结果可视化，设计了可通过浏览器网页访问的交互式操作界面，系统界面见图 3，可以展示追踪总人数、重庆市输入人员常住人口统计、参与人员信息图，以及高危感染者在交通工具和公共场所的分布。此外，系统提供图形分析结果查询功能，行政人员和疾病预防控制人员可以通过关键字查询，直接查看图形结果。

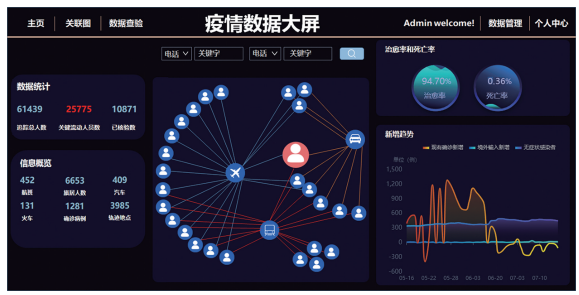

图 3　高风险人群防控系统界面

2　结　　果

2.1 效率分析实验　效率分析实验结果见图 4。

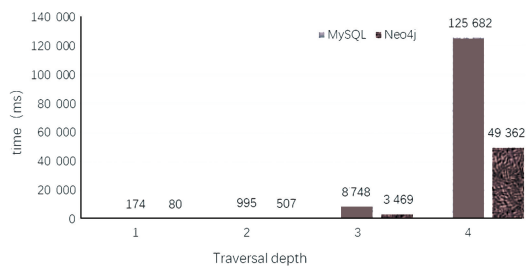

图 4　关系型数据库与图数据库查询效率对比

2.2 图数据库算法代表性分析结果　与确诊病例直接接触过的高危人员分析图见图 5，与确诊病例有过直接或间接接触的分析图见图 6，他们均有感染风险。

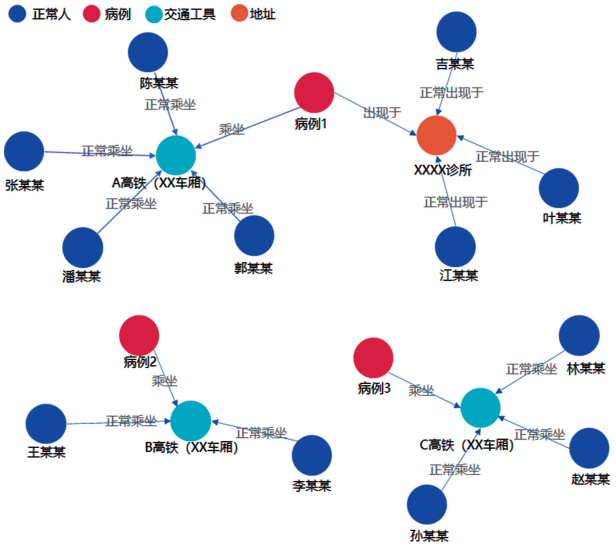

图 5　场景 1：与确诊病例直接接触分析

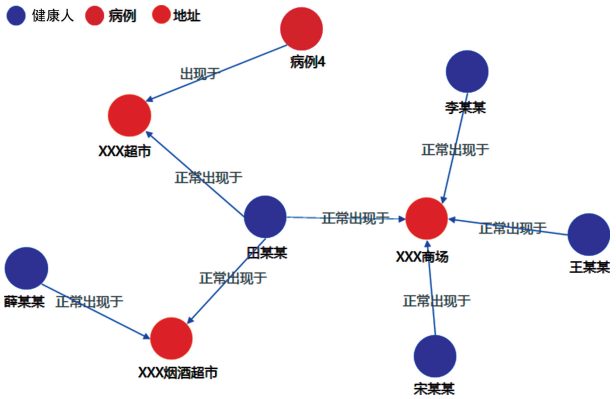

图 6　场景 2：与确诊病例间接接触分析

3　讨　论

当时的 COVID-19 疫情大流行对社会稳定、经济发展及人类的生命健康带来严重影响，快速控制传染病传播的重要性凸显，如何快速发现疫情传播链中的传播关系成为当时疫情防控重点关注对象。本研究首先通过对比关系数据库、图数据库性能差异，得出图数据库在处理复杂关系时性能优于关系数据库，然后基于 Neo4j 图数据库算法和 ECharts 数据可视化工具对多源头 COVID-19 疫情大数据进行挖掘和分析，用于发现并追踪与确诊患者、疑似患者和无症状感染者相关的隐蔽接触者和公共场所，可以快速准确地发现密切接触者、二级或三级接触者，并能识别高感染风险的公共场所。数据可视化系统帮助应急管理专员简单直接地查询高感染风险人群和公共场所，并可以动态监测疫情情况，提供决策支持。本研究样本数据有限，如在真实场景中运用时，因疫情数据涉及个人隐私问题，数据的安全性是要重点思考的问题，系统需要进一步综合运用密码学方法、数据加密技术、系统漏洞监测与修复技术、数据自动删除技术等数据安全手段去加以保障，这也将是未来进一步的研究方向。本研究也为突发公共卫生事件流行防控奠定了基础。

参考文献

[1] RICCI L，MAESA D F，FAVENZA AND E. Ferro，blockchains for COVID-19 contact tracing and vaccine support：A systematic review [J]. In Ieee Access，2021，9(4)：37936-37950.

[2] FERRETTI L，WYMANT C，KENDALL M，et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing[J]. Science，2020，368(6491)：122-134.

[3] JAHMUNAH，VICNESH，et al. Future iot tools for COVID-19 contact tracing and prediction：A review of the state-of-the-science[J]. International J Imaging Systems Technol，2021，31(2)：455-471.

[4] 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.

[5] MARTIN T，KAROPOULOS G，HERNÁNDEZ-RAMOS J，et al. Demystifying covid-19 digital contact tracing：A survey on frameworks and mobile apps[C]. Wirel Commun Mob Computing，2020：1-29.

[6] 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.

[7] 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.

[8] 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)：2924-2927.

[9] MUNZERT S，SELB P，GOHDES A，et al. Tracking and promoting the usage of a COVID-19 contact tracing app[J]. Nature Human Behav，2021，5：247-255.

[10] GIABELLI A，MALANDRI L，MERCORIO F，et al. Graphlmi：A data driven system for exploring labor market information through graph databases[J]. Multimed Tools Appl，2020，29：384-395.

[11] MOON M J. Fighting COVID-19 with agility，transparency，and participation：Wicked policy problems and new governance challenges[M]. PUBLIC ADM REV，2020：651-656.

[12] TANG X，MA C，YU M，et al. A visualization method based on graph database in security logs analysis[C]. 6th. International Conference On Advanced Materials And Computer Science (Icamcs)，2007：82-89.

（收稿日期：2022-11-15　修回日期：2023-01-23）