

DOI:10.3969/j.issn.1671-0673.2021.05.014

基于知识图谱的新冠肺炎病例传播关系可视分析

李 佳, 刘海砚, 刘俊楠, 刘建湘

(信息工程大学, 河南 郑州 450001)

摘要:目前,新冠肺炎传播迅速,影响广泛,对全球的人类生存和经济都造成了重大影响。已有的流行病学分析方法侧重于统计分析,忽视了病例间的时空传播关系和语义关联关系。通过构建新冠肺炎病例知识图谱进行可视化并加以分析,可以结合语义和时空特征挖掘新冠肺炎传播过程和发展趋势。以郑州市疾病预防控制中心发布的病例通报数据为基础,针对人群活动模型组成要素,构建了新冠肺炎病例知识图谱本体层和数据层。在构建知识图谱后,综合应用甘特图、平行坐标图、关联关系图等可视化方法,设计了一个基于新冠肺炎病例知识图谱的交互式可视分析原型系统,发现新冠肺炎病例的多维度特征、病例活动和传播过程。

关键词:新冠肺炎;知识图谱;可视分析;时空关系;语义关系

中图分类号:TP391.7

文献标识码:A

文章编号:1671-0673(2021)05-0606-07

Visual Analysis of Transmission of COVID-19 Cases Based on Knowledge Graph

LI Jia, LIU Haiyan, LIU Junnan, LIU Jianxiang

(Information Engineering University, Zhengzhou 450001)

Abstract: Covid-19 has spread rapidly and affected a wide range of people, with significant impacts of human health and socio-economic. The epidemiological analysis focus on statistical analysis, however it ignore the temporal relationship and semantic relationship between cases. This paper proposes a visual analysis method based on Covid-19 patients knowledge graph, to explore the spread and development trend of Covid-19. Firstly, we provide the ontology-layer and data-layer of the knowledge graph based on the features of crowd activity. Then, we design and implement a visual analysis system which also provides a set of interactions to involve users in the entire process of the epidemiological data analysis. To validate our approach, we utilize the data sets from Zhengzhou and show how our method can find multi-dimensional features, crowd activity and transmission process.

Key words: Covid-19; knowledge graph; visual analysis; spatiotemporal relationship; semantic relationship

新型冠状病毒肺炎(简称新冠肺炎)自 2019 年 12 月爆发以来传播迅速,对人类健康和社会经济都造成了极大影响。在 2020 年 1 月 30 日,该疫情被世界卫生组织认定为国际关注的突发公共卫生紧急事件。新冠肺炎具有强烈的传染性,对武汉市早期确诊病例的接触人群进行分析,确定了新冠

肺炎可在人与人之间传播^[1]。利用新冠肺炎病例活动数据构建知识图谱^[2],可以分析病毒传播的规律。因此,利用新冠肺炎病例数据和流行病学调查数据,运用流行病传播分析方法,挖掘患者的发病规律、传播关系路径和趋势,对临床治疗和疫情防控具有重要意义。

收稿日期:2021-03-14;修回日期:2021-04-06

基金项目:国家自然科学基金资助项目(41801313,41901397)

作者简介:李 佳(1996-),女,硕士生,主要研究方向为地理信息可视化。

随着新冠肺炎病患数量不断增加,疫情时空传播路径会变得越来越复杂。已有的流行病学分析方法侧重统计分析,无法兼顾时空传播关系和语义关联关系,因此需要一种同时兼顾时空和语义特征的数据组织形式来对病例数据进行建模。以新冠肺炎流调数据为中心,结合知识图谱前沿技术,通过构建适应多样化描述方式的流行病病例知识图谱,进而实现整体上概览病例关联关系,细节上展示人群传播过程。

随着可视化技术的发展,可视分析在通信^[3]、智能制造^[4]、工业^[5]等领域得到了广泛的应用。在公共健康领域,可视化与可视分析也发挥了重要的作用。文献[6]应用 Google Earth 三维地理环境可视化技术将传染病早期预警结果进行展现。文献[7]将流行病传染模型进行可视化,可以交互式评估疫情情景下的应对措施,以辅助决策。文献[8]通过可视分析系统,分析不同传染病的时空规律,挖掘传染病与地区之间的相关性。知识图谱由于其易扩展的结构,也经常用于流行病网络分析,文献[9]利用大规模网络以揭示患者之间的传播关系。文献[10]构建病例活动异质网络以寻找影响流行病传染的核心节点。

本文提出一种基于新冠肺炎病例知识图谱的可视分析方法。该方法根据病例实体的特征和人群活动要素建立新冠肺炎病例知识图谱,包含流行病学知识图谱和基于人群活动模型的活动图谱。在病例图谱的基础上建立交互式可视分析系统,支持病例流行病学特征分析和病例活动分析。针对病例的流行病学特征分析,运用平行坐标图可视化病例多维特征,以甘特图和地理散点图可视化病例的时空信息。针对病例活动分析,采用关联关系图可视化病例关系,支持用户查询病例活动图谱和活动轨迹,以帮助用户发现新冠肺炎病例的多维度特征、病例活动和传播过程。同时,为了验证本文方法的有效性,采用郑州市新冠肺炎病例数据进行案例分析。经实验验证,本文提出的基于新冠肺炎病例知识图谱的可视分析方法可以有效地帮助用户进行病例的多维度特征、病例活动和传播过程分析。

1 数据来源与需求概述

1.1 病例数据

本文采用郑州市疾控中心发布的病例通报数据进行分析,如表1所示。该地区一共报告了158位病例数据,时间跨度从2020年1月21日到2020

年3月11日,主要包括了病例的基本情况、语义关联关系和确诊前活动记录。基本情况包括了病例编号、性别、年龄、民族、住址和病例的初次就诊日期、确诊日期等、确诊医院等信息;语义关联关系描述了病例和已确诊病例之间的关系(如亲属关系、同行关系、接触关系);确诊前活动记录包括了病例的时空信息,主要为确诊前的旅行史,包括在何时与何人乘坐何种交通工具达到某地。

表1 郑州市疾病预防控制中心病例通报数据(部分)

编号	通报数据
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日由负压救护车转运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3	男,68岁,现住巩义市园丁街烟草局家属院。1月10日乘Z6列车(4车厢)从武汉返回郑州,当日乘坐K735次列车从郑州返回巩义,1月20日在巩义市人民医院就诊,1月21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。
.....

1.2 需求分析概述

病例基本信息不仅包含了年龄、性别等病例的基本情况,还包含了包含社会关系信息、时空信息等多维度信息。在面对包含多维度、复杂的病例信息时,分析人员无法直接对病例进行分析以发现患病规律、获取隐含信息。因此,通过研究流行病学特征分析相关文献和人群活动模型,分析病例数据特点和初步实验,本文从病例的总体态势分析和病例的传染链分析这两个方面总结了病例可视分析的需求。

①病例总体态势析。分析人员希望了解该地区疫情的发展情况和该地区病例的总体情况。具体包括了该地区病例的增长数量分析;病例来源分析;病例流行病学特征分析;病例的潜伏期分析。

②病例传播关系分析。为了能够分析病例的传播关系,分析人员希望能够清楚地展示病例间的语义关联关系、病例活动信息和病例分布信息。

2 新冠肺炎病例知识图谱构建

本节结合流行病学特征要素和5w1h活动模

型,从知识图谱结构出发,将病例实体、病例基本信息、病例活动事件抽象为知识图谱的节点,将病例实体—病例基本信息、病例实体—病例活动事件、病例实体—病例实体表示为知识图谱的边,构建面向模式层和数据层的流行病学病例知识图谱表示模型。

2.1 数据要素解析

流行病病例数据中包含了病例基本情况(年龄、性别等),病例的时空活动轨迹以及语义关联关系(包括亲属关系、同行关系、接触关系);根据病例确诊前的活动路径确定病患的暴露史(主要分为武汉,以及湖北地区暴露史);之后结合病例的暴露史和传播关系,确定病例代数。在武汉及其周边城市被感染后转移到当地的病例确定为一代病例;一直在当地活动且由一代病例传染的病例确定为二代病例;由二代病例传染的病例定义为三代病例。

为了清晰地描述病例确诊前的活动以追踪病例感染过程,采用了5w1h法(what、who、when、where、why和how)对病例的活动进行描述。其中,What指事件参与对象以某种方式发生显著时空变化的过程,具体为病例参与活动的目的或活动演化结果;who为参与该活动的对象集合;when为该活动发生的时间;where为活动发生的地点,可展示病例的活动范围;why为活动发生的原因,即为病例活动意图;how为病例完成活动所采用的手段,具体可细化为病例参与活动所采用的交通工具,可推断该活动的波及范围。

2.2 模式层构建

为了表达病例的流行病学特征和活动情况,定义流行病学病例知识图谱模式层为流行病学模式层(G_s)、病例活动模式层(G_m)和病例之间的关系(Rel),如图1所示。

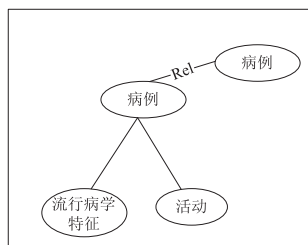
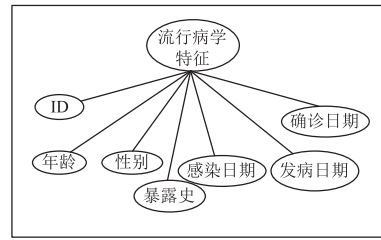


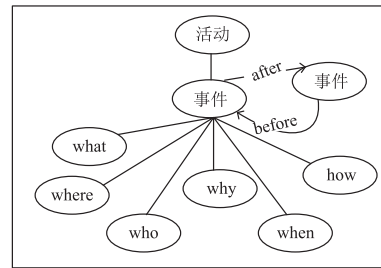
图1 新冠肺炎病例知识图谱模式层概念体系

流行病学模式层 G_s 定义为病例实体的流行病学特征,即 $G_s = \langle ID, 年龄, 性别, 暴露史, 感染日期, 发病日期, 确诊日期 \rangle$,如图2(a)所示;由事件图谱 G_{event} 和事件的时序关系 R_t 构成病例活动模

式层 G_m 即 $G_m = \langle G_{event}, R_t \rangle$,如图2(b)所示。事件图谱 G_{event} 由一个六元组 $\langle when, where, who, why, what, how \rangle$ 组成,表示病例活动状态。



(a) 流行病学模式层



(b) 活动模式层

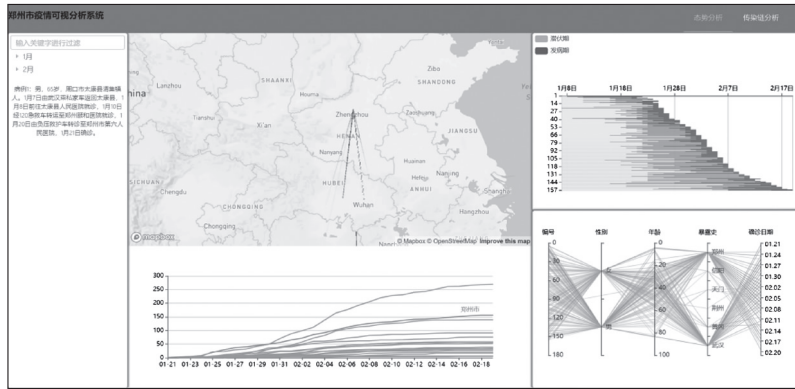
图2 新冠肺炎病例知识图谱模式层

2.3 数据层构建

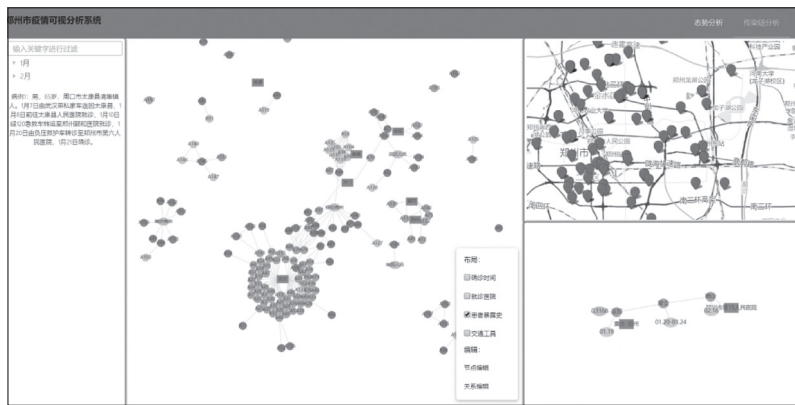
数据层 G_d 包括了流行病学数据层和病例活动数据层,流行病学数据层由流行病模式层实例化而成,病例活动数据层由病例活动模式层实例化而成。由于原始的流行病病例数据为非结构化的文本数据,本文使用Hanlp自然语言处理工具包进行病例基本情况、病例活动和病例语义关系提取,主要包括中文分词、去停用词、词性标注、命名实体识别等步骤。之后,根据数据要素解析形成的数据格式进行要素抽取形成结构化数据。由于Neo4j数据库采用网络结构存储数据,因此把病例对象和病例属性作为节点、把病例的关系和事件时序关系作为边,将整理好的结构化数据以三元组的形式导入到Neo4j数据库中,构建流行病病例知识图谱数据层。

3 病例可视分析与交互操作

本文设计并实现了一个可视分析原型系统,帮助工作人员对新冠肺炎病例进行可视分析。根据分析需求,它主要包括病例态势分析和病例传染链分析块两个模块,每个模块中都具有刷选、点击等简明的交互功能。图3为系统界面,其中图3(a)为态势分析模块,包括了病例来源、病例增长折线图、潜伏期统计图和流行病学特征图;图3(b)为传染链分析模块,包括了病例的关联关系图、病例分布



(a) 态势分析



(b) 传播关系分析

图 3 系统界面

图和病例活动图谱。本节先介绍系统设计目标,然后分别介绍每个模块的设计细节。

3.1 设计目标

为了指导系统界面和交互的设计,本文确定了以下设计目标:

(1)帮助用户进行疫情发展态势分析。在病例数量规模较大的情况下,可视化系统能够直观地呈现出该地区疫情的发展状况、病例来源和病例流行病学信息,帮助用户分析病例的来源特征、发病规律等信息。

(2)帮助用户进行病例传染链分析。利用病例活动知识图谱直观地展示病例间的关联关系,同时还能查询病例确诊前的活动情况和位置信息,帮助用户进行信息推理。

3.2 态势分析模块

疫情态势分析模块旨在帮助用户进行该地区的疫情发展态势分析,并支持用户在界面上通过交互,选择感兴趣的单个病例,进入后续病例传播关系分析。模块包括病例增长折线图、病例来源图、潜伏期统计甘特图和流行病学特征统计图等 4 个部分。

病例增长折线图展示了该地区的病例增长情况以及和其他地区的情况对比,用户可以通过折线图对比了解该地区疫情的发展状况;病例来源图以直观的方式展示了输入型病例的来源,用户可以直观地看到输入型病例的主要来源地;潜伏期统计甘特图展示了病例从感染到就诊、就诊到确诊直接的时间间隔,以供用户观察病例的潜伏期,对于潜伏期特别长的病例,用户应该重点关注病例的活动;流行病学特征统计图展示了病例性别、年龄、暴露史、确诊日期,支持刷选交互操作,以分析不同时期病例的特征以及特征之间的关联,挖掘疫情发生、发展规律。同时,系统还提供了病例信息概览功能,用户可以通过左侧的折叠面板,按日期打开或者通过关键词搜索,查看病例信息。

3.3 传播关系分析模块

病例传播关系分析模块旨在利用病例活动知识图谱帮助用户对病例的传染链进行分析,以挖掘病例隐含的传播关系,具体包括关联关系图、病例分布图和病例活动图谱。

基于病例活动知识图谱的关联关系图展示了病例之间的传染关系。用户可以查看病例之间的

传染链,通过颜色确定病例的病例代数,其中蓝色的为一代病例,红色的为二代病例,粉色为三代病例。之后,可以对关联关系图中的关联条件(确诊日期、就诊医院、暴露史和交通工具)进行增减,以分析病例间的关系以及病例确诊日期、就诊医院、暴露史和出行事件的关系。病例分布图以多图联动的交互方式展示传染链中病例的分布。病例活动图谱以图谱形式展示了病例活动情况。在活动事件图谱中可以看到病例确诊前的活动信息。用户可以根据病例活动图谱进行病例关系推理:如两个病例住址相同,确诊前有相同的路径,乘坐了相同的交通工具,但是却没有上报病例之间关系,可以推断出两个病例可能为亲属关系。

4 案例分析

本文选用河南省郑州市新冠肺炎病例数据集,以验证文中提出的可视化模块的有效性;同时,对郑州市的新冠肺炎病例数据进行案例分析,以检验该交互式的病例实体可分析系统的实用性。

根据该病例数据集的特点,结合系统模块,提出了一种典型的疫情态势分析和病例传播关系分析的可视化分析方法。具体分析过程如下:

(1)疫情态势分析。根据郑州市的疫情发展状况,并分析病例的来源、潜伏期和流行病学特征,挖掘病例特征之间的关联,总结病例发生特点。

(2)病例传播关系分析。根据疫情态势分析中的异常值和病例的关联关系图,结合病例分布图和病例活动图谱对病例进行传播关系分析,挖掘病例间关系。

4.1 态势分析

疫情态势分析主要通过病例来源、病例增长折线图、潜伏期统计图和流行病学特征图分析该地区疫情发展态势。

图4为病例增长折线图,郑州市作为河南省省会城市,从1月21日出现第一例病例起,病例数量不断增加,直到2月16日以后,增长曲线趋于平稳。其病例数量仅次于信阳市,疫情发展情况不容乐观。

通过对郑州市的输入型病例来源地进行分析,发现郑州的输入型病例主要来源为武汉市,如图5所示。疫情防控人员应该对本市中近期有武汉旅行史的人员进行着重关注,筛查可能病例。

病例的潜伏期统计图统计了患者从感染到就

诊、就诊到确诊的时间间隔。图6中,发现一些病例较早就出现了症状,但是一直未确诊,或者病例的潜伏期较长。结合病例传播关系分析模块,发现这些病例都发生了二次传播,有共同居住亲属被感染。

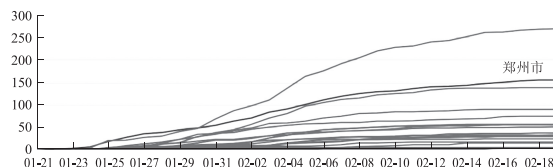


图4 病例增长分析



图5 病例来源分析

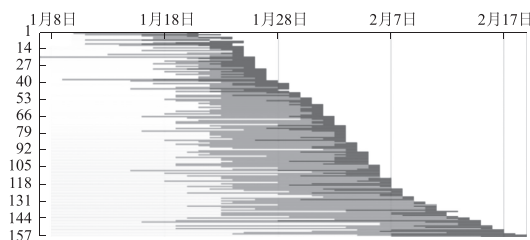


图6 潜伏期统计分析

在病例流行病学特征概览图中,将所有病例流行病学特征信息映射到平行坐标图。图7为病例流行病学特征概览平行坐标图,图7(a)为整体病例信息,坐标从整体病例信息来看,病例没有太大的性别趋势,而年龄主要分布在20~70岁之间,有外地暴露史的病例当中,来自于武汉的病例最多,但也有少部分病例来自于武汉周边城市。图7(b)为对整体的病例进行筛选,选择了前30位病例进行分析,发现前30位病例当中,男性病例多于女性病例,且大部分人都具有武汉暴露史。图7(c)为筛选了后30位病例进行分析,后30位病例当中,男女数量不再有较大差别,输入型病例人数减少,本地病例人数增加,但还是有来自于武汉及其周边城市的病例,通过查询这些病例活动,发现该病例在一月份就到了郑州,但是由于潜伏期比较长,发病时间较晚。

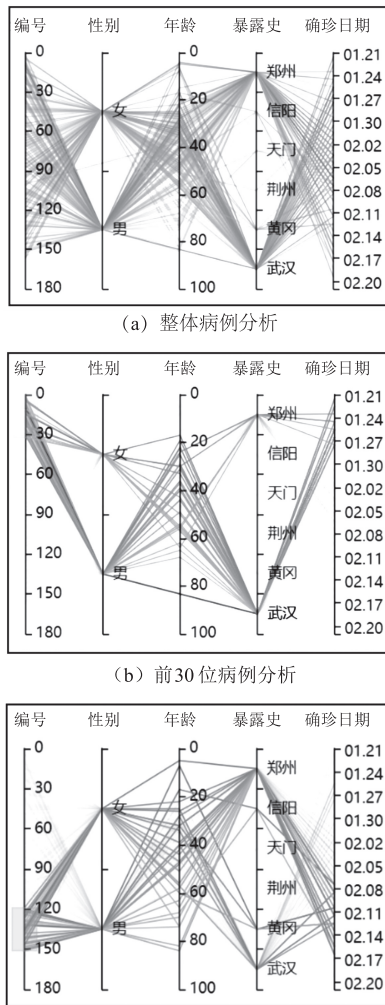


图 7 流行病学特征分析

4.2 传播关系分析

传播关系分析包括了病例关联关系图、病例分布图和病例活动图谱。病例传播关系分析基于病例活动图谱进行 5w1h 要素分析,需要利用病例活动对病例关系进行分析和挖掘。

图 8 为病例的关联关系图,图中大部分为蓝色的一代病例,只有部分病例通过活动接触其他病例,导致二代病例和三代病例出现。通过分析一代病例和二代病例之间的关系,发现大部分的传染发生在亲属或同行人员之间。

图 9 为关联关系图加上病例暴露史布局之后选取的病例异常传染链。病例 49 在接触了一位病例之后被感染,之后同旅行团的 9 名人员先后被感染。由于病例共同旅行,同行时间较长,同旅行团人员被感染风险较大。工作人员在发现该旅行团的第一起病例之后,应该联系旅游公司进行排查,防止继续出现被感染病例。

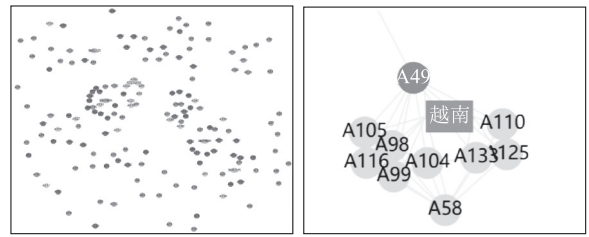


图 8 病例关联关系分析 图 9 同行病例传染分析

图 10(a)为关联关系图加上交通工具布局之后选取的异常关系结点。通报中仅告病例 88 和病例 95、病例 97 分别为亲属关系,没有仔细说明病例之间的同行关系。图中可以发现病例 85、病例 88 和病例 95 乘坐了同一班次高铁,病例 87 和病例 97 也乘坐了同一班次高铁。通过查询 5 位病例的住址,发现 5 位病例的住址均为“郑东新区如意湖办事处绿地老街”。图 10(b)、图 10(c)、图 10(d)、图 10(e)、图 10(f)则分别为病例 85、病例 87、病例 88、病例 95 和病例 97 的活动图谱。对 5 位病例的活动行程进行分析,推测在 1 月 21 日,病例 85、病例 88 和病例 95 一同乘坐 G541 班次高铁到达武汉,之后又一同乘坐私家车到达天门;而病例 87 和病例 97 于 1 月 19 日乘坐 G295 班次高铁到达天门。5 位病例在天门汇合之后,又于 2 月 2 日

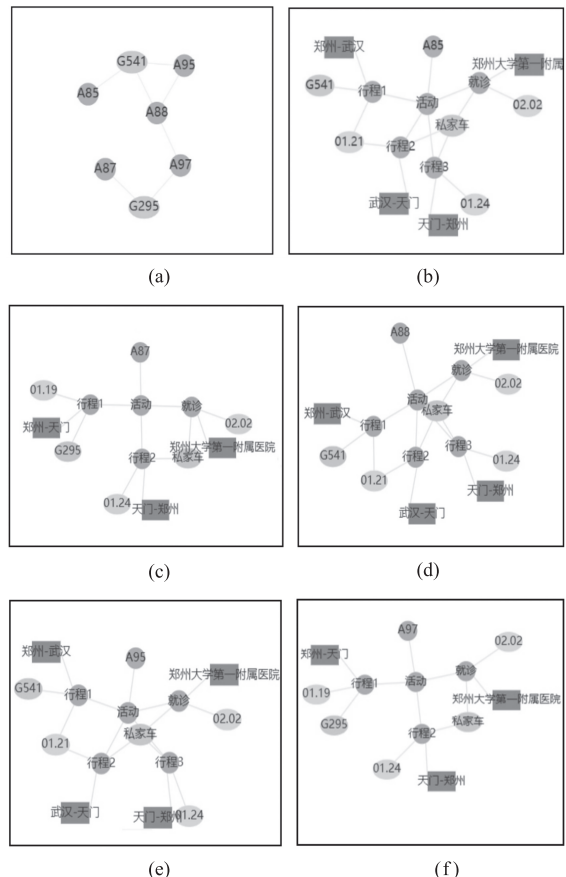


图 10 病例 85、87、88、95、97 活动图谱分析

一同乘坐私家车返郑。从图5中分析可推测该5位病例应该互相为亲属关系,是一起典型的家族集聚型传染案例。

分析人员可以通过高铁班次查找该5位病例的密切接触者,进行潜在感染人员排查。之后,可把挖掘到的同行信息和亲属信息加入到数据库中,方便进行后续分析。

5 结束语

本文提出了一种基于新冠肺炎病例知识图谱的可视分析方法,进行疫情发展态势分析与病例传播关系分析。通过病例基本特征和人群活动模型构建新冠肺炎病例知识图谱,以描述病例的基本信息、关联关系和病例活动导致的时空传播过程。在疫情态势分析部分,通过折线图分析病例的增长情况,通过流行病学特征平行坐标图进行病例刷选,分析不同病例的特性以及特征之间的关联,挖掘病例特征;在病例传播关系分析部分,根据病例的关联关系图,采用人群活动模型描述病例活动,帮助用户从 what、why、who、where、when、how 多个角度去分析病患活动,并根据病例活动实现病例关系挖掘和潜在传染人群预测。最后,通过实验和用户反馈,本文方法利用新冠肺炎病例知识图谱能充分展示病例多维度特征、进行病例活动和传播过程分析,并且可通过简单的交互融合用户体验,挖掘隐含的信息。

但该系统还存在许多缺陷,主要为以下几点:①目前,该系统只支持一个小区域的病例分析。在未来的工作中,需要尝试扩大区域范围,支持多区域病例分析。②本文基于知识图谱构建病例数据,但在表达病例多样化的信息时,只能采用多视图的方法展示病例信息。在未来的工作中可以探索适合的方法展示图谱信息。③该系统只能依靠用户分析提取出异常病例。如何将用户体验更多地直接融入系统中,实现自动化分析值得进一步探索。④当病例增多时,图的节点数量会大幅度增加,后续可尝试使用基于变化的大图可视化方法^[11]和空间聚类方法^[12]。

参考文献:

- [1] WU Z Y, MCGOOGAN J M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China; summary of a report of 72-314 cases from the Chinese center for disease control and prevention [J]. JAMA, 2020, 323(13): 1239-1242.
- [2] 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.
- [3] ZHAO Y, LUO X B, LIN X R, et al. Visual analytics for electromagnetic situation awareness in radio monitoring and management [J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 590-600.
- [4] ZHOU F F, LIN X R, LIU C, et al. A survey of visualization for smart manufacturing [J]. Journal of Visualization, 2019, 22(2): 419-435.
- [5] 黄辉, 陆利忠, 闫镡, 等. 三维可视化技术研究 [J]. 信息工程大学学报, 2010, 11(2): 218-222, 247.
- [6] 殷菲, 冯子健, 李晓松. Google Earth 在传染病早期预警结果三维可视化中的应用 [J]. 中华流行病学杂志, 2011, 32(4): 396-399.
- [7] AFZAL S, MACIEJEWSKI R, EBERT D S. Visual analytics decision support environment for epidemic modeling and response evaluation [C] // 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), 2011: 191-200.
- [8] 金思辰, 陶煜波, 严宇宇, 等. 基于多维时空数据可视化的传染病模式分析 [J]. 计算机辅助设计与图形学学报, 2019, 31(2): 241-255.
- [9] CLÉMENCON S, DE ARAZOZA H, ROSSI F, et al. Visual mining of epidemic networks [C] // International conference on artificial neural networks, 2011: 276-283.
- [10] CASTELLANO C, PASTOR-SATORRAS R. Competing activation mechanisms in epidemics on networks [J]. Scientific Reports, 2012, 2: 371.
- [11] 时磊, 廖琦, 林闯. 基于变换的大图点边可视化综述 [J]. 计算机辅助设计与图形学学报, 2013, 15(3): 304-311.
- [12] 夏佳志, 张亚伟, 张健, 等. 一种基于子空间聚类的局部相关性可视分析方法 [J]. 计算机辅助设计与图形学学报, 2016, 28(11): 1855-1862.

(编辑:李志豪)