

基于Neo4j的人员关系知识图谱构建及应用

陆 枫

(河南警察学院, 河南 郑州 450046)

✉279840745@qq.com



摘 要: 近年来, 知识图谱已逐渐从传统的知识分析应用发展到对社会实体及关系的应用。本文针对知识图谱的人员关系应用研究, 以疫情病例人员关系为例获取数据源, 通过数据预处理、知识抽取等方式构建病例人员关系三元组, 采用Protégé工具构建本体, 并导入Neo4j图形数据库中进行存储, 最后构建病例人员关系知识图谱, 实现可视化查询病例人员及关系信息。同时, 运用Neo4j的中心度算法和度中心性算法分析所有病例人员的感染力和接触人数, 从而发现病例中感染力较强、接触人数较多的人员, 为未来疾病防控体系的建立提供了参考, 对病例人员关系的研究分析也具有一定的实用价值。

关键词: 知识图谱; 图数据库; Neo4j; 关系分析

中图分类号: TP391 **文献标识码:** A

Construction and Application of Personnel Relationship Knowledge Graph based on Neo4j

LU Feng

(Henan Police College, Zhengzhou 450046, China)

✉279840745@qq.com

Abstract: In recent years, knowledge graph has been gradually applied from traditional knowledge analysis to the application of social entities and their relationships. This paper proposes to study the application of knowledge graph to personnel relationship, and the data source is obtained by taking the epidemic case-person relationship as an example. The case-person relationship triplet is constructed by data preprocessing and knowledge extraction, etc., and the Protégé tool is used to construct the ontology, which is imported into Neo4j graphic database for storage. Finally, a knowledge graph of case-person relationship is built, realizing the visual query of case personnel and their relationship information. Meanwhile, the centrality algorithm and the degree centrality algorithm of Neo4j are used to analyze the infectivity and contact number of all case persons, so as to find out among the cases, persons with strong infectivity and more contacts. It provides a reference for the establishment of the future disease prevention and control system, and also has certain practical value for the research and analysis of the case-person relationship.

Keywords: knowledge graph; graph database; Neo4j; relational analysis

1 引言(Introduction)

“知识图谱”这一概念在2012年被Google正式提出^[1]。2013年之后, 随着智能信息服务和应用的不断发展, 知识图谱、大数据和深度学习一起, 成为推动互联网和人工智能发展的核心驱动力。知识图谱用图形显示知识发展进程与结构关系, 以可视化技术来描述知识资源, 构建、分析、挖掘和显示知识资源及它们之间的相互联系^[2]。从目前大数据的数据

分析结果来看, 追踪人员活动轨迹、构建关系知识图谱和定位传播路径等是人们关注的重点内容。

本文提出一种基于病例人员关系图数据库的知识图谱, 该知识图谱以图的形式展示了病例人员的信息和接触关系, 系统采用Neo4j图数据库存储, 对碎片化的病例关系数据进行知识建模与关联聚合, 实现知识层面的数据融合与集成, 构建可视化的病例人员关系知识图谱, 形成病例人员关系图,

同时专业人员也可以分析其接触关系及病例人员的感染力。

2 知识图谱和Neo4j(Knowledge graph and Neo4j)

2.1 知识图谱和Neo4j的定义

知识图谱又称为科学知识图谱,是一种叫作语义网络(semantic network)的知识库。知识图谱用于以符号形式描述物理世界中的概念及其相互关系,它的基本组成单位是“实体、关系、实体”三元组,实体间通过关系相互联结,构成网状的知识结构^[1]。

Neo4j是一种NoSQL的图数据库,也是目前应用较多的图数据库,它基于Java的高性能、高可靠性、可扩展性强的开源图数据库^[2]。Neo4j的创始人们选择用图模型来存储关系,并在此系统中实现了变长的遍历运算。Neo4j的数据存储形式主要是用节点(Node)和边(Edge)来组织数据。节点代表知识图谱中的实体,边代表实体间的关系,关系可以有方向,两端对应开始节点和结束节点。

2.2 知识图谱的构建

目前,知识图谱的构建方式主要有自顶向下和自底向上两种方式。自顶向下构建方式需要先定义好本体(Ontology,或称为Schema),再基于输入数据完成知识抽取到图谱构建的过程。该方式更适用于专业知识方面图谱的构建,比如企业知识图谱,面向专业领域用户使用。自底向上构建方式则是借助一些技术手段,从公开收集的数据信息中抽取置信度较高的信息,加入知识库中。该方式适用于常识性知识图谱的构建^[2]。

本文构建的病例关系知识图谱采用自顶向下的构建方式。下面侧重介绍自顶向下构建方式的相关流程和技术^[3],其知识图谱的关键技术架构如图1所示。

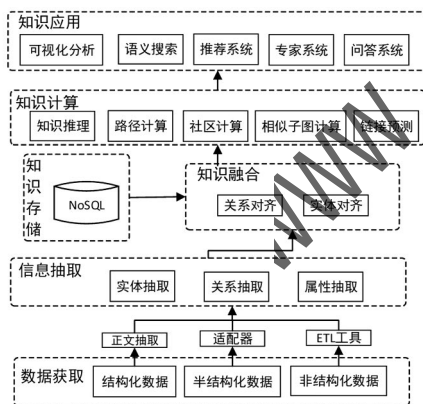


图1 知识图谱的关键技术架构图

Fig.1 Key technology architecture diagram of knowledge graph

3 基于Neo4j的病例关系知识图谱构建(Construction of case relationship knowledge graph based on Neo4j)

病例关系知识图谱的构建主要包括数据获取、信息抽取、知识存储、知识融合和知识应用等部分。首先,获取某地病例人员的数据信息;其次,进行信息抽取,完成对获取数据的识别并抽取相关实体、关系和属性;第三,采用Neo4j图数据库实现知识存储;然后,将获取来源不同的等价或相关实体进行整合和知识融合,以消除矛盾和歧义,比如某些

实体可能有多种表达,某个特定称谓也许对应多个不同的实体等^[4];最后,知识应用主要采用Neo4j的中心度算法和度中心性算法对病例人员进行关系分析。

3.1 数据获取

数据获取主要完成对知识图谱所在领域的数据源信息收集,这些数据主要包括结构化数据、半结构化数据与非结构化数据^[5]。本文的数据主要来自某地病例人员的信息,信息的数据内容包括确诊病例的姓名、性别、职业、居住地址、所在区域、工作地址、确诊日期及接触关系,这些数据多为文本内容和半结构化信息。因此,本文采用分词方式将关键信息提取出来形成结构化数据库,为后续抽取实体和关系形成图数据库做准备^[6]。

3.2 信息抽取

信息抽取也叫知识抽取,是一种通过自动化的方式从半结构化或非结构化数据中抽取实体、关系、属性信息的技术。在早期,知识抽取主要通过启发式算法与规则相结合的方式来实现,现在常用的方法有借助本体与词汇集、借助多层神经网络和自然语言处理等^[7]。

病例关系的知识抽取是从病例中抽取所需要的内容,包含实体、关系和属性。首先,对病例实体的数据属性和关系属性进行梳理和抽取。病例实体主要包括“确诊病例”“无症状感染”“外地确诊病例”三种;实体的属性有姓名、性别、职业、居住地址、所在区域、工作地址和确诊日期等;实体关系的属性主要有“密接”“同空间”“亲戚”“同住”“同事”等几种关系。其次,依据确定的病例本体核心类,利用Protégé工具对病例本体类与关系属性进行创建,设置“are responsible for”“SubClass of”“cooperate”等五种关系^[8]。第三,结合所构建的病例关系知识模型,抽取病例中的病例姓名、居住地点、工作地点、感染时间等关键实例数据并导入,形成病例关系应用本体,其构建本体的部分截图如图2所示。最后,导出为RDF数据,为病例关系知识图谱的构建奠定基础。

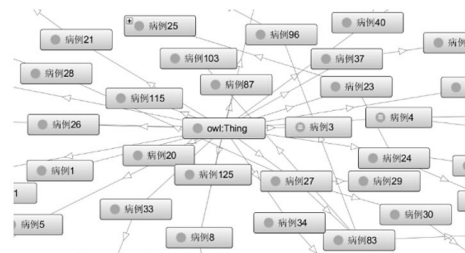


图2 Protégé构建本体的部分截图

Fig.2 Partial screenshot of building ontology with Protégé

3.3 知识存储

知识存储主要研究知识图谱采用何种方式进行存储。目前知识图谱主要有两种存储方式:一种是基于RDF结构的存储方式;另一种是基于图数据库的存储方式^[9]。病例关系知识图谱采用基于图数据库Neo4j的存储方式。Neo4j存储数据可以采用多种导入方式,本系统存储采用将RDF数据导入Neo4j中进行存储,然后利用Neo4j的Cypher语言对实体、关系以及属性进行知识图谱的设计,实现图形化交互式查询和关联化推理^[10],其病例关系知识图谱的部分截图如图3所示。

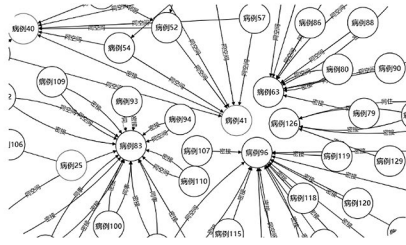


图3 病例关系知识图谱的部分截图

Fig.3 Partial screenshot of case relationship knowledge graph

由病例关系知识图谱可以看出, 病例关系以层次化的图结构形式将其病例接触关系呈现出来, 实现了Neo4j高效的图查询和分析应用。

4 病例关系知识图谱的应用(Application of case relationship knowledge graph)

病例关系知识图谱通过Neo4j中的Cypher语言实现对病例数据进行更新、查询和分析, 主要利用Neo4j的度中心性算法和中心度算法分析病例中感染力强和接触人数较多的人员, 从而更好地了解病例人员及其接触关系, 提高流行性疾病的预防能力, 进一步完善疾病防控体系。

4.1 数据更新

为了使本知识图谱能充分映射现实世界中的真实情况, 应该对模型中的数据记录不断更新, 对实体和关系进行增删操作, 即修改模型中的实体记录和关系记录。

4.1.1 实体和关系增加

CREATE(病例a:Person1{name: '病例a', 状态: '确诊病例', 职业: '职员', 居住区域: 'B区', 居住地址: 'B区阳光大道', 工作地址: 'B区阳光大道', data: '2022-4-28'}) // 增加一个实体

MATCH(n:Person1),(m:Person1) WHERE n.name="病例83" AND m.name="病例103" CREATE(n)-[r:密接]->(m) RETURN r // 在病例83和病例103中直接增加一个“密接”关系

4.1.2 实体和关系删除

删除单个实体有两种方法: 通过name属性或者id属性删除实体; 如果两个实体之间存在关系, 应该先删除关系, 然后再删除实体。

MATCH (n:Person1{name: 'temp'}) delete n // 通过name属性, 删除name为'temp'的实体

MATCH (n:Person1) WHERE id(n)=23 delete n // 通过id属性, 删除id为23的实体

MATCH (n) DETACH delete n // 删除所有实体

MATCH (a:Person1)-[r:密接]-(b:Person1) where ID(a)=47 and ID(b)=48 DETACH delete a,b,r // 删除实体a、实体b及实体a和实体b的“密接”关系

4.2 信息查询

信息查询是对病例人员及关系的查询, 主要了解和病例人员接触的相关人员关系。信息查询主要根据实体属性和关系属性进行查询, 首先提取查询条件, 如实体的属性: 姓名、感染时间等, 或实体的关系: “密接”“同空间”等; 然后利用Neo4j的Cypher语言进行查询, 查询结果可以以图的

形式显示符合条件的病例关系图, 也可以以表的形式显示病例人员信息表。本文以实体“病例83”为例, 主要以实体的属性、实体的关系或实体的属性和关系相结合作为条件进行信息查询。

4.2.1 以病例的感染时间(data)属性为条件查询

查找某时间感染的所有病例人员, 例如查找2022年4月21日的病例人员, 运行结果如图4所示, 其Cypher执行语句如下:

```
MATCH (n:Person1{data:"2022-4-21"}) RETURN n
```

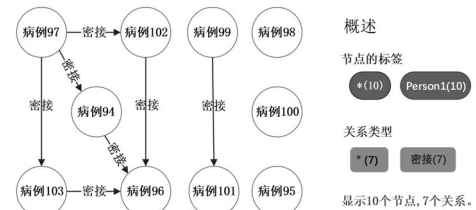


图4 某时间感染的所有病例人员图

Fig.4 Diagram of all infected persons at a given time

运行结果图4的左边显示了2022年4月21日所有病例人员及其关系图; 右边显示了共有病例10个, 其中通过“密接”感染的有7人。

4.2.2 以病例的“密接”关系为条件查询

查找与病例人员有“密接”关系的所有病例人员信息, 运行结果如图5所示, 其Cypher执行语句如下:

```
MATCH (p:Person1)-[a:密接]->()
RETURN p.name AS 病例姓名, p.data AS 感染日期, p.'居住地址' AS 居住地址, COUNT(a) AS 感染人数
ORDER BY 感染人数 DESC
```

病例姓名	感染日期	居住地址	感染人数
“病例130”	“2022-4-28”	“A区阳光大道”	3
“病例131”	“2022-4-28”	“A区阳光大道”	3
“病例15”	“2022-4-2”	“E区公园路”	2

图5 按“密接”关系查找所有病例图

Fig.5 Diagram of finding all cases by close contact relationship

4.2.3 以病例的name属性和“密接”关系为条件查询

根据病例的name属性查找与“病例83”“密接”的人员, 运行结果如图6所示, 其Cypher执行语句如下:

```
MATCH n=(.Person1{name:"病例83"})-[:'密接']-(.)
RETURN n
```

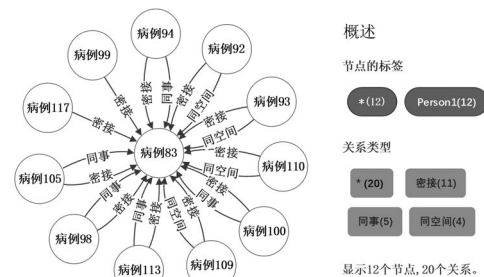


图6 与“病例83”“密接”的人员图

Fig.6 Diagram of close contacts with case 83

运行结果图6的左边显示了与“病例83”“密接”的人员关系图；右边显示了与“病例83”“密接”的人员有11个，其中有4人与“病例83”同空间，有5人与“病例83”是同事关系。

4.3 利用Neo4j算法对病例关系进行分析

本文对病例人员关系的分析主要利用Neo4j的中心度算法和度中心性算法。其中，中心度算法用来分析病例人员的感染力，度中心性算法用来分析和病例人员接触的人数，通过这两个算法分析和发现病例中出现的感染力较强、接触人数较多的情况，及时掌握目前病例人员的状况。

4.3.1 中心度算法PageRank

PageRank算法是一种用于测量实体的传递性或定向性影响的算法。它已在Google中进行推广，被广泛认为是一种检测任何网络中有影响力实体的方法。PageRank已用于对公共场所或街道进行排名，预测这些区域的交通流量和人员流动等^[11]。

在本文的病例关系研究中，PageRank算法的分数(score)可以反映出每个病例的感染力大小，分数值越高越有影响力，因此可以分析出病例中最有影响力的感染者，运行结果如图7所示，其Cypher执行语句如下：

```
CALL algo.pageRank.stream("Person1","密接",
{iterations:20})
```

```
YIELD nodeId,score
```

```
MATCH (node) WHERE id(node)=nodeId
```

```
RETURN node.name AS 病例姓名,score
```

```
ORDER BY score DESC
```

病例姓名	score
“病例83”	1.968303
“病例96”	1.7155384999999999
“病例7”	0.6313124999999999

图7 PageRank算法分析结果图

Fig.7 Diagram of PageRank algorithm analysis result

从运行结果图7中可以分析出“病例83”的分数最高，感染力也较大；“病例96”的感染力次之。

4.3.2 度中心性算法

度中心性算法是由Linton C. Freeman于1979年提出的，该算法经常用于找出单个实体的受欢迎程度，常作为对全图进行最小度、最大度、平均度及标准差整体分析的一部分。它是任何通过查看入度和出度关系数量进行影响力分析的重要组成部分^[11]。

在本文的病例关系研究中，病例关系图的入度用于记录感染该病例人员的人数(即感染人数)，出度用于记录被该病例人员感染的人数(即被感染人数)，运行结果如图8所示，其Cypher执行语句如下：

```
MATCH (u:Person1)
```

```
RETURN u.name AS 病例姓名,
```

```
size((u)-[:密接]-()) AS 感染人数,
```

```
size((u)-[:密接]->()) AS 被感染人数
```

```
ORDER BY 感染人数 DESC
```

病例姓名	感染人数	被感染人数
“病例83”	11	0
“病例96”	8	2
“病例103”	4	1

图8 度中心性算法分析结果图

Fig.8 Diagram of degree centrality algorithm analysis result

从运行结果图8中可以分析出每个病例的接触和被接触人数，动态掌握病例接触情况。例如，接触“病例83”的有11人；接触“病例96”的有8人，“病例96”接触的病例者为2人。

5 结论(Conclusion)

本文首先介绍知识图谱和图数据库Neo4j的定义以及构建知识图谱的关键技术，然后根据病例关系知识图谱的构建流程和构建技术，对获取的数据源进行数据提取和对实体、关系和属性的信息抽取，再用Protégé工具构建本体RDF数据导入图数据库Neo4j中进行存储，实现多种查询检索和关系分析。病例关系知识图谱的构建也为病例关系分析提供了参考，进一步完善了传染疾病防控体系，从而有针对性地提供较为完善的防控措施。

参考文献(References)

- [1] 刘屹,李杨,段宏,等.知识图谱构建技术综述[J].计算机研究与发展,2020(3):582-600.
- [2] [法]杰罗姆·巴同,[英]里克·范·布鲁根.Neo4j 3.x入门经典[M].北京:清华大学出版社,2017:170-193.
- [3] HAO X, JI Z, LI X, et al. Construction and application of a knowledge graph[J]. Remote Sens, 2021(13):2511-2529.
- [4] 曹皓伟,徐建良,窦方坤.基于Neo4j生物医药知识图谱的构建[J].计算机时代,2020(6):35-38.
- [5] 赵雪芹,杨一凡,于文静.基于Neo4j图数据库的工程档案知识图谱构建及应用[J].档案与建设,2022(5):48-51.
- [6] 肖飞龙,张爽,胡志凌.基于Neo4j的疾病预防与控制知识图谱研究[J].电子技术与软件工程,2021(2):180-182.
- [7] 张莉,王玉廷.基于Protege的糖尿病本体构建[J].科学咨询(教育科研),2020(11):9-11.
- [8] 沈阳,李洪磊,陈杰.图数据模型及其在疫情追溯领域的应用研究[J].软件导刊,2021(2):13-17.
- [9] 丁洪丽.基于Neo4j图数据库的人员关系挖掘[J].电讯技术,2018(6):25-32.
- [10] 张帆.Neo4j权威指南[M].北京:清华大学出版社,2017:449-457.
- [11] 俞方桦.Neo4j图数据库扩展指南:APOC和ALGO[M].北京:清华大学出版社,2017:189-268.

作者简介:

陆 枫(1976-),女,硕士,讲师.研究领域:数据库应用,计算机技术.