

气象记录档案知识图谱构建^①



江双五, 刘惠兰, 温华洋, 谢 伟

(安徽省气象信息中心, 合肥 230031)

通信作者: 江双五, E-mail: jiangjiang0621@sina.com

摘 要: 为提升气象档案管理和社会化利用能力, 提出了一种气象记录档案知识图谱构建方法. 以安徽省馆藏气象记录档案资源为基础, 通过自顶向下的方式设计概念层及层次关系来构建本体库, 基于本体库对接实体构建知识模型, 分别采取 D2RQ、图像识别和定制包装器对不同存储格式的档案进行知识抽取构建数据层, 通过建立同义词库等来完成知识融合, 利用 MySQL 和 Neo4j 数据库分别完成本体库和实体库的存储, 并实例展示了构建的气象记录档案知识图谱. 本次知识图谱的构建丰富了气象记录档案中蕴含知识单元及其语义关系, 主要包括 28 个核心概念、280 个一般概念, 6 个本体关系模型, 1 万卷以上的档案实体, 为气象和档案跨行业知识图谱的构建提供一定的借鉴.

关键词: 气象记录档案; 知识图谱; 本体; 知识模型; 构建

引用格式: 江双五, 刘惠兰, 温华洋, 谢伟. 气象记录档案知识图谱构建. 计算机系统应用, 2022, 31(1): 73-82. <http://www.c-s-a.org.cn/1003-3254/8315.html>

Construction of Knowledge Graph of Meteorological Record Archives

JIANG Shuang-Wu, LIU Hui-Lan, WEN Hua-Yang, XIE Wei

(Anhui Meteorological Information Center, Hefei 230031, China)

Abstract: To improve the abilities of meteorological archives management and social utilization, this paper proposes a method to construct the knowledge graph of meteorological records archives. With the meteorological records archives resources in Anhui Province, the ontology database is constructed by the design of conceptual layer and hierarchical relationship in a top-down way. A knowledge model is constructed based on the ontology database docking entity. Knowledge is extracted from archives in different storage formats with D2RQ, image recognition and custom wrapper for data layer construction. A synonym database is established to facilitate knowledge fusion. MySQL and Neo4j databases are employed for the storage of ontology database and entity database respectively, and an example demonstrates the knowledge graph construction of meteorological records archives. The knowledge graph construction in this research enriches the knowledge units and semantic relations contained in meteorological records archives, mainly including 28 core concepts, 280 general concepts, 6 ontological relationship models, and more than 10 000 volumes of archives entities. It can provide a reference for knowledge graph construction across industries of meteorology and archives.

Key words: meteorological record archives; knowledge graph; ontology; knowledge model; construction

1 引言

气象记录档案是气象观测和探测中形成的气象要素、天气现象及其变化过程的真实记录, 是基础性的

气象信息资源. 目前全国各类气象档案馆存储并管理着始于 19 世纪中期以来的各类气象记录档案, 这些档案形成的“资源仓库”是开展气象领域及相关行业科学

^① 基金项目: 安徽省气象局新技术集成项目 (AHXJ202003)

收稿时间: 2021-03-29; 修改时间: 2021-04-29; 采用时间: 2021-06-02; csa 在线出版时间: 2021-12-17

技术研究的重要资源,是研究我国近代气候变化不可替代的基础性资料和国家基础性战略资源。

2018年,中国气象局组织安徽、河北、山东和湖北4省开展数字气象档案馆试点建设工作,以馆藏档案社会化利用为目的,期望各试点省积极利用大数据、语义网等信息技术推动气象档案管理和利用方法的创新改变,其中一个重要的试点建设内容是面向全国气象档案馆建立一个语义化、智能化的气象档案管理系统。而数字气象档案管理系统后台的挖掘算法和相关统计查询的分析效果直接依赖于背景知识图谱的构建质量。高质量的气象档案知识图谱的建立不仅为我国气象档案的知识组织提供理论框架,同时也为推动气象档案的社会化利用提供重要的技术支撑。

2 相关研究

2.1 本体

1991年, Neches 等人^[1]将本体定义为构成相关领域词汇的基本术语和关系;1993年, Gruber^[2]将本体定义为“一种概念化的精确的规格说明”;1998年, Studer 等人^[3]将本体定义为“共享概念模型的明确形式化规范说明”。随着专家学者们对本体知识认识的不断深入,目前科学研究工作者普遍将本体定义为特定领域内的概念、属性以及之间关系。

我国在领域本体构建方面已经取得了一些初步成果,如医疗领域本体知识库^[4]、多民族语言本体知识库^[5]和中文语言本体知识库。在气象领域方面少数的气象工作者也开展了本领域本体构建的探索,如陈梅对大气污染和其相关因素进行了本体建模^[6];何险峰开展了气象灾害本体的设计^[7];罗炜基于气象数据集开展了本体的构建^[8]等。但查询公开发表的文献,气象档案领域本体构建方面国内还鲜有高质量的知识库。

2.2 知识图谱

知识图谱概念是从早期的语义知识网演变而来的。目前,知识图谱研究方向分为通用知识图谱和特定行业知识图谱。通用知识图谱主要利用现实生活中的大量常识性知识来构建,为互联网查询能力的智能化提供支持;特定行业知识图谱主要利用行业数据来构建,为行业数据智能利用提供支持。当前,医药、军事和产业等不同行业都在探索构建适合本领域的知识图谱^[9-22]。国内外研究显示档案领域的知识图谱构建研究尚处于起步阶段,雷杰等利用 EAD、DCMI、VIVO、

SWRC、Schema.org 等现有较为通用的本体构建科研档案管理知识图谱^[23],气象领域知识图谱的研究同样处于起步阶段,大多以公开发表的文献为资源开展知识图谱构建研究,王建勋等人^[24]对1995–2017年中国知网的数据库中有关干旱遥感监测研究的642篇文献进行可视化分析得到知识图谱,王萍等人^[25]利用文献分析工具对2000–2017年中国知识资源总库中关于灾害风险研究的2890篇中文文献进行分析,绘制出灾害风险研究的热点、关键词共现矩阵、核心作者群及研究机构的知识结构图谱。李莉等人^[26]针对分散的网络气象文本数据,编写爬虫从网络上爬取相关文本数据,在此基础上构建了气象文献知识图谱。本研究以气象记录档案资源智能化利用为目的,研究对象涵盖气象观测资料、观测规范和气象档案元数据等在内的主要气象档案资源,通过挖掘气象档案资源之间的语义关系来构建知识库,提升档案管理系统查询和检索能力。

2.3 知识图谱与本体的关系

知识图谱从逻辑结构上一般分为概念层与实体层。概念层是构建目标领域的概念模型,以概念、属性与关系来描述目标领域,通常通过建立本体模型来建立概念层;实体层则以实体为构建对象,是以“实体-关系-实体”与“实体-属性-属性值”三元组的形式对真实数据描述和存储,并基于大量的三元组来形成关联语义网络,实体层是通过知识获取技术来实现的。基于本体的知识图谱构建方法中,知识获取是建立在本体模型的基础上,本体模型的构建质量是知识图谱构建的关键。

3 气象记录档案知识图谱构建

3.1 总体构建思路

以用户对气象档案的利用需求为导向,以馆藏档案资源为基础,首先在分析气象档案特点基础上构建气象档案本体库模型,再应用知识图谱构建技术完成气象记录档案知识图谱的构建。气象记录档案知识图谱的构建流程主要分为气象记录档案资源的收集整理、知识图谱模式层的构建以及基于模型层的知识图谱的构建。基于气象记录档案资源及该领域知识理解完成该领域概念、属性、实体以及之间的关系构建,形成知识模型,依据模型从档案资源中抽取实体、关系、属性值数据,将抽取的数据转化成符合气象记录档案知识模型图关系数据,再结合气象记录档案专业

知识对数据进行实体对齐、指代消解等知识融合处理,最后采用知识推理技术挖掘气象档案数据中的隐含关系及知识,形成气象记录档案的知识图谱.从架构上

(图 1)知识图谱的构建自下而上包括 6 大步骤,分别是气象记录档案数据源的收集整理、知识建模、知识抽取、知识融合、知识存储和知识应用.

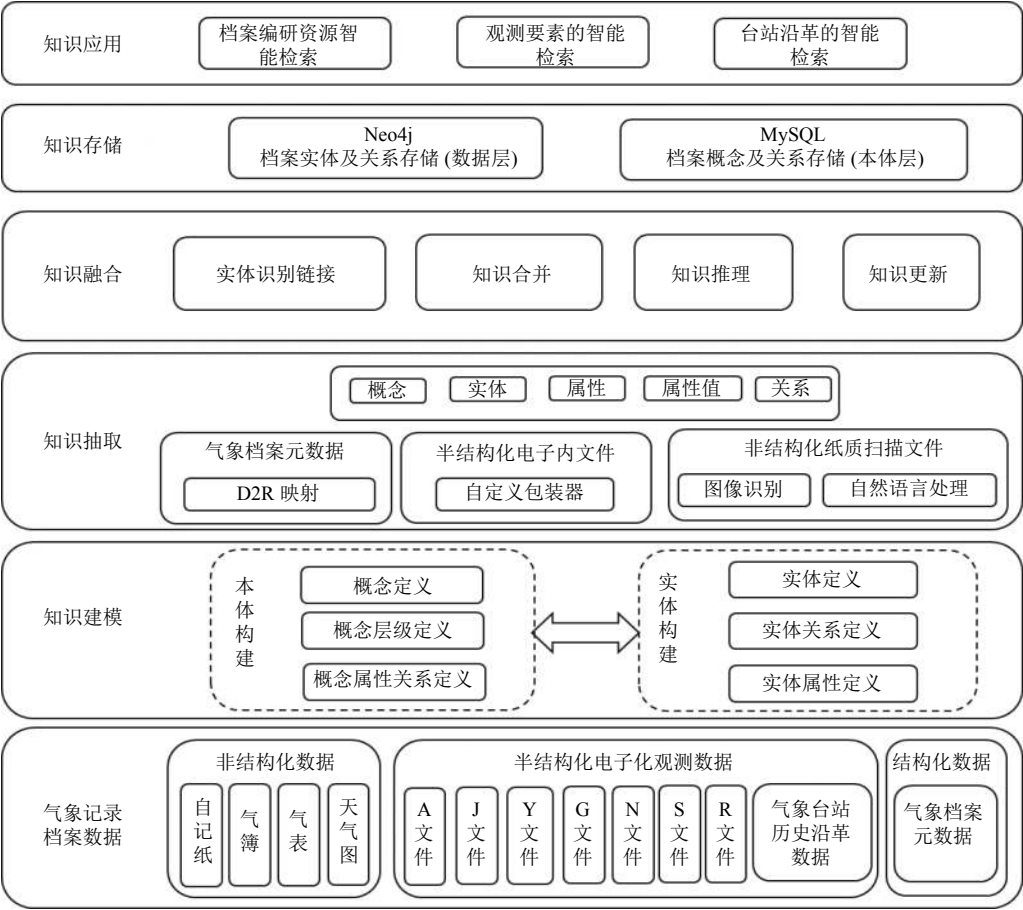


图 1 气象记录档案知识图谱构建架构

3.2 数据源

气象档案资源是知识图谱本体库和实体库构建的基础,主要包括气象档案元数据、馆藏气象档案、气象业务技术规范档案和气象文献相关知识以及基于用户反馈信息等档案资源.

3.2.1 气象档案元数据

基于气象行业标准《气象档案元数据》,主要包括气象档案馆元数据、气象档案类别元数据、气象档案案卷元数据、气象档案卷内文件元数据、气象档案管理元数据等 5 大类 105 项元数据的收集和整理. 主要包含档案的类别、时间、来源、案卷数、存放位置、保存介质等内容.

3.2.2 馆藏气象档案

以安徽省气象档案馆馆藏 1 万卷档案为样例数据,

主要包含全省 81 个台站建站以来的地面、高空、辐射、农业、酸雨、天气图、数据加工文件和台站历史沿革文件等.

3.2.3 气象业务技术规范

收集中国气象局公开发布的气象业务规定和技术规范方面的档案,利用自然语言处理技术,进行气象档案各类词表的扩充并基于数据挖掘技术进行知识图谱中概念和实体关系的发现.

3.2.4 文献知识

查询公开发表气象档案的文献资料,研究并实现基于模板的气象档案知识抽取,对知识图谱的知识进行补充,并研究知识图谱驱动下的气象档案知识自动抽取技术.

3.2.5 用户反馈

广泛开展用户咨询,基于用户的反馈对知识图谱

的知识进行修正和补充. 随着应用服务的不断加强, 这一部分将是未来气象记录档案知识图谱迭代更新的重要知识来源.

3.3 气象记录档案知识模型构建

气象记录档案知识模型的构建以本体模型的构建为核心, 以收集的档案资源为资料源, 在结合《气象学词典》和《中国档案主题词表》(第2版)等领域资料的基础上, 选取气象领域的基本术语、分类信息、主题词来建立概念, 同时结合应用需求和分类方法建立概念层次. 气象记录档案本体模型构建的基本流程包括5个步骤(图2): 资源梳理、概念及层级构建、属性定义、概念关系定义、本体语言表示和本体概念与实体映射.

3.3.1 气象记录档案本体概念及层次构建

从数据源中归纳整理出气象记录档案领域核心概念, 核心概念的归纳整理注意本体中类的设计秉承独立性、共享性原则^[27]以及类的数目最小化原则^[28]. 根据上述原则, 从气象档案应用场景出发, 从档案的生命周期出发选择“气象记录档案”作为最顶层核心概念, 次核心概念为“气象档案文件”“气象记录档案形成”和“气象记录档案管理”3大类, 最终从顶向下定义了气象记录档案领域的28个核心概念(图3).

通过核心概念的建立, 逐步建立各层次概念280个, 核心子概念“气象记录档案文件”概念层次模型如图4.

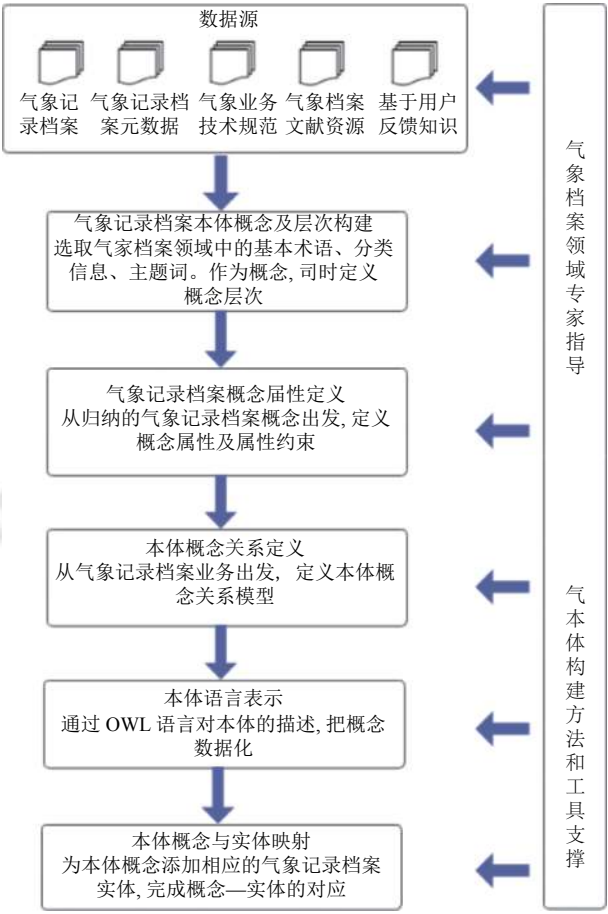


图2 气象记录档案本体模型构建流程

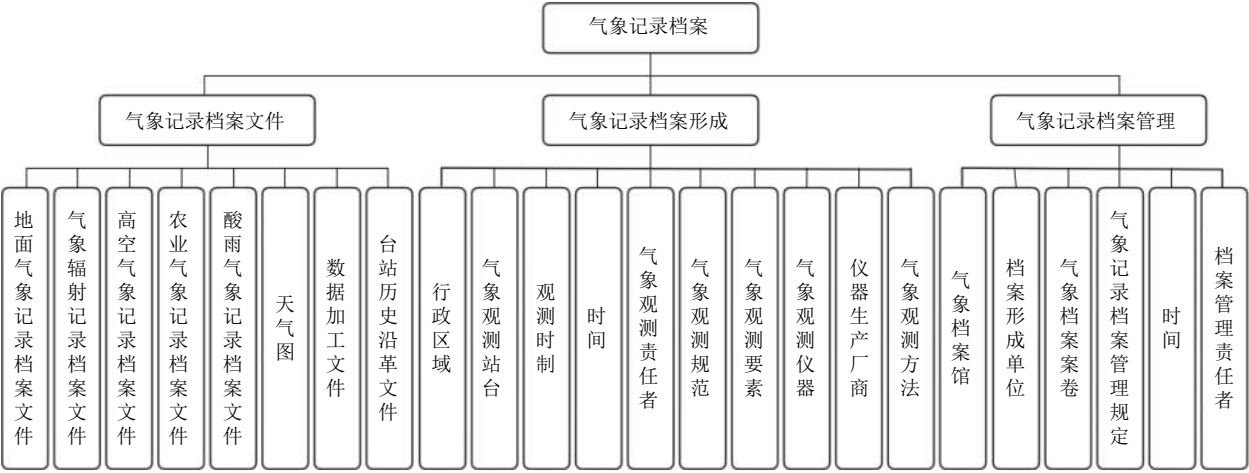


图3 气象记录档案核心概念及层次

3.3.2 气象记录档案本体属性构建

从归纳整理出的气象记录档案概念出发, 对各气象记录档案本体概念定义属性以及属性约束, 其中属

性应结合用户查找需求来定义. 以“气象记录档案文件”概念的属性定义来说明: 首先通过调研用户对气象记录档案的查找需求, 查看分析气象记录档案纸质文

件、电子文件以及气象记录档案元数据,选取能代表气象观测文件特性和能满足用户快速查找定位档案的词作为气象记录档案文件属性(图5),分别为台站编号、文件名称、文件档号、形成单位、地域号、形成日期、记录类型、保管期限、关键词和密级等。

3.3.3 气象记录档案本体概念关系定义

在气象记录档案本体中, 根据本体关系的层级和结构, 通过归纳总结得到本体关系模型主要见表 1, 共分为物理、空间、管理、观测、时间和事件相关性 6 大类。

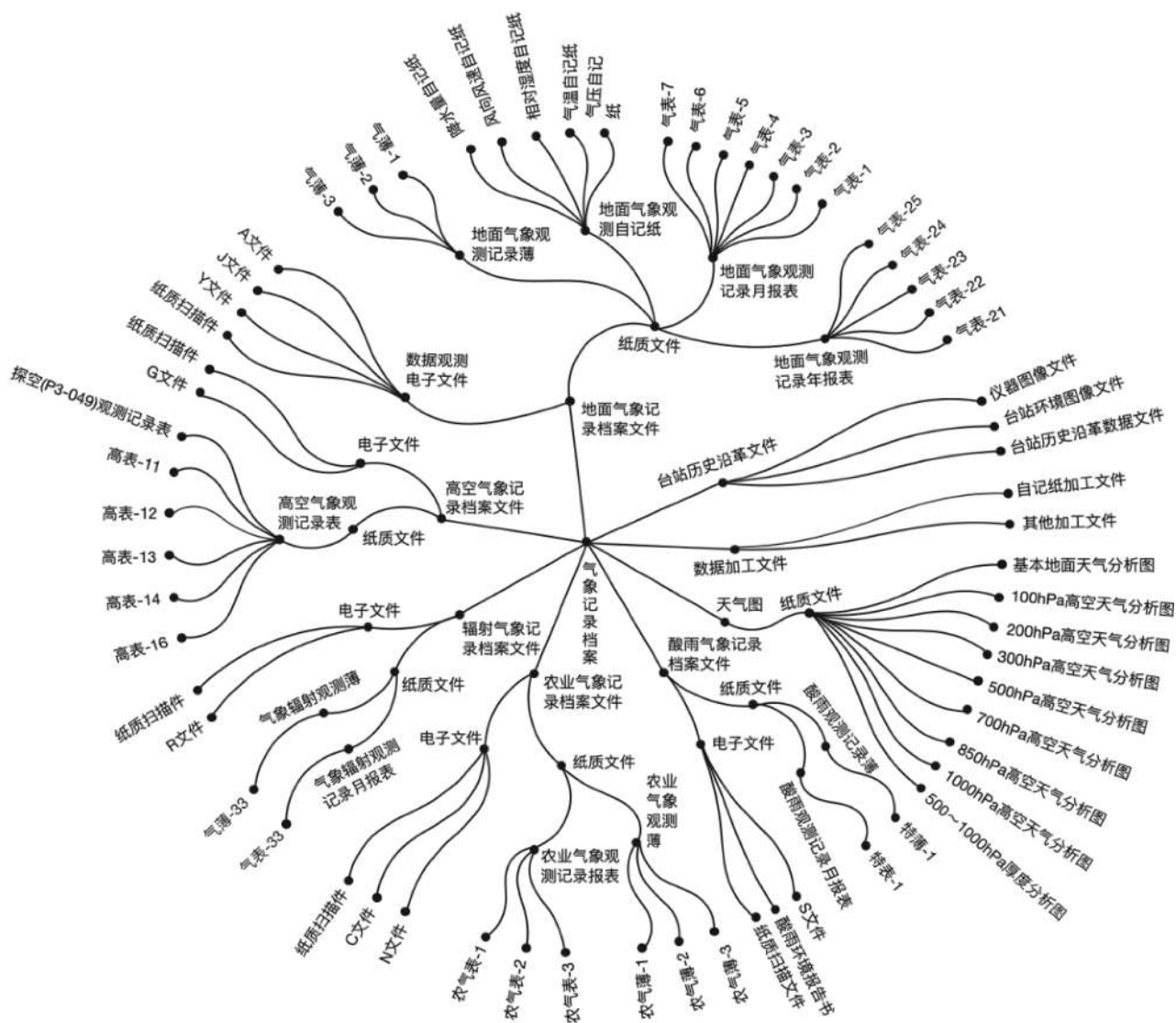


图4 气象记录档案文件概念层次

气象记录档案本体关系的定义是根据气象记录档案业务分析, 围绕气象记录档案的产生、管理和利用, 以气象记录档案自身即“气象观测文件”为核心, 定义各概念之间的关系, 在此基础上建立气象记录档案数据之间的关联. 总的概念关系图如图 6 所示.

3.3.4 本体语言表示

本体中的概念、属性以及层级结构需要采用

OWL 语言来描述,把概念数据化,使计算机能够理解与处理.

3.3.5 本体概念与实体映射

在知识图谱中,本体概念是对实体的一个抽象描述,实体是最基本的元素,它们之间存在对应关系.只有实体对接到概念模型中才能完成知识图谱的构建,如“气象观测台站”是一个概念,“砀山县国家一般气象

站”是一个实体,它们分别是知识图谱中“图”的一个节点,通过“图”中节点之间的连线来建立概念和实体间的关系.知识图谱模型实例见图7.

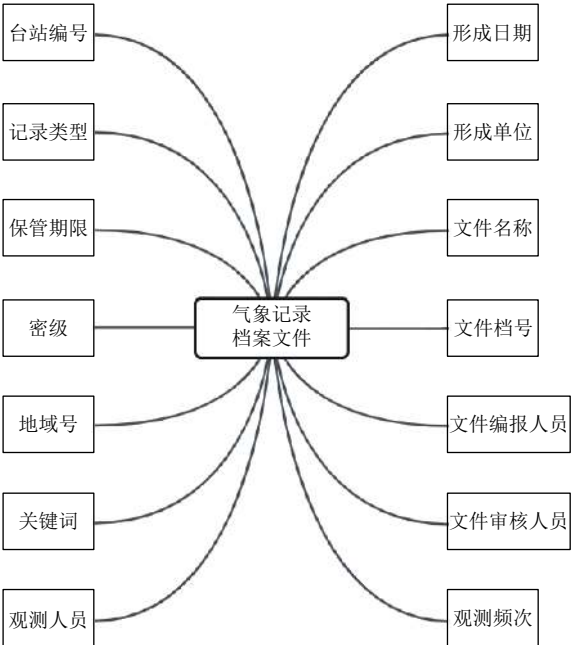


图5 气象记录档案文件属性

3.4 气象记录档案知识抽取

依据已构建的气象记录档案知识模型,抽取安徽省馆藏气象记录档案实体、属性及关系.目前安徽省气象档案馆馆藏气象记录档案资源从存储格式上分为结构化、半结构化和非结构化3种结构.要根据不同的存储结构定义不同的知识抽取策略,具体来说结构

化的馆藏档案著录元数据采用D2RQ数据处理工具进行处理;非结构化的纸质气象记录档案数字化档案图像、文本、PDF等数据采用图像识别、文本抽取以及自然语言处理等技术进行处理;半结构化的电子类气象观测数据文件及台站历史沿革数据文件则根据文件的存储规则,定制其对应的包装器来完成气象观测数据的抽取.将抽取的数据转化成符合知识图谱模型结构的数据,存入图数据库中.

3.5 气象记录档案知识融合

3.5.1 气象记录档案实体识别链接

实体识别链接是将气象档案实体实例数据与气象档案知识模型中的概念进行对应识别链接的操作.在气象档案数据中,气象台站的台站名称、观测位置、观测仪器等会发生改变,可能导致存在同一个观测位置出现不同的名称,同一个观测要素出现不同的要素名称,同一个观测仪器出现不同的仪器名称.如“合肥国家基本气象站”与“合肥国家基本气象站;安徽省合肥国家基本气象站;合肥站;合肥;合肥气象站;合肥基本站;合肥国家站;合肥国家基本站”等在系统中为同一实体,类似这样的数据在实体识别链接中需要解决同名实体含义不同和不同名实体含义相同的问题,构建出符合业务实际的气象记录档案知识图谱.

3.5.2 气象记录档案知识合并

气象档案观测数据形式多样,同一地点、同一时间气簿、气表、自记纸中的观测数据可能产生重复记录的情况,数据抽取到图谱库中产生了冗余,这就需要对气象档案实例及关系数据进行知识合并.

表1 气象记录档案本体关系模型

关系类别	自定义概念关系	关系注释
物理	所在案卷	组成/由组成的关系
	馆藏	包含/被包含的关系
空间	案卷位置、所在区域、区域相邻、地理相邻、台站位	位置
观测	台站周围障碍物	遮挡/被遮挡的关系
	台站合并	合并/被合并的关系
	观测台站、观测文件、形成单位、加工提取	产生/被产生的关系
	所用观测仪器、所用观测方法	使用/被使用的关系
	依据观测规范	依据/被依据的关系
时间	要素高度相关、要素计算产生	导致、引起、被引起的关系
	观测时间、移交时间、接收时间、借阅时间、事件发生时间、规范制定时间	事物/工作时间的关系
事件	台站变动事件、仪器变动事件、其他观测过程事件	发生的关系
管理	档案管理者、管理档案	管理/被管理的关系
	档案移交者、移交档案	移交/被移交的关系
	档案接收者、接收档案	接收/被接收的关系
	档案借阅者、借阅档案	借阅/被借阅的关系

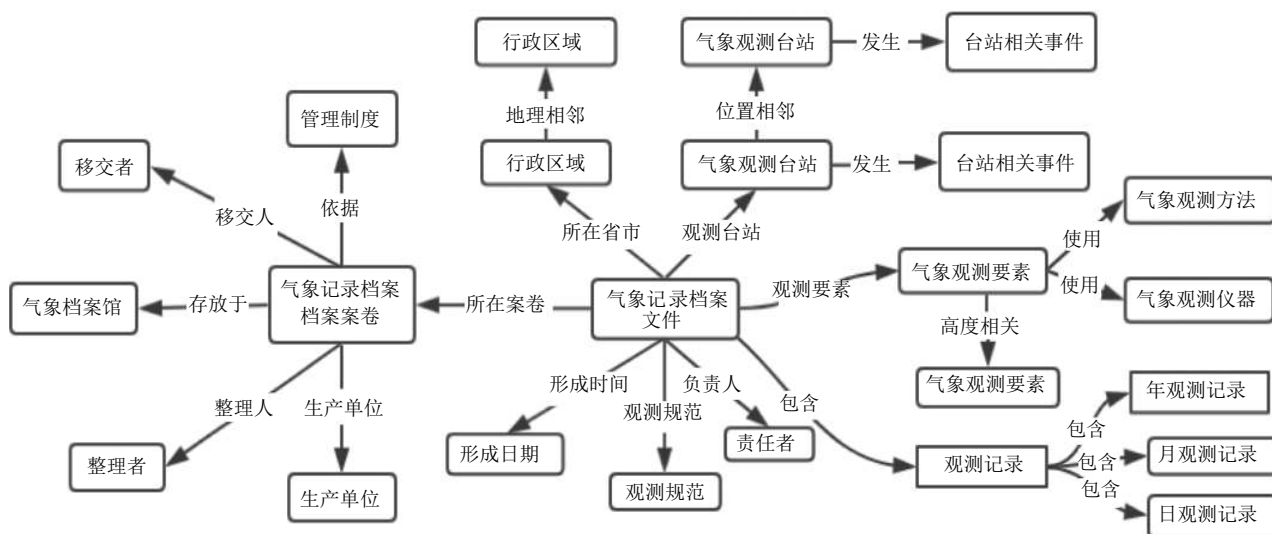


图 6 概念关系图

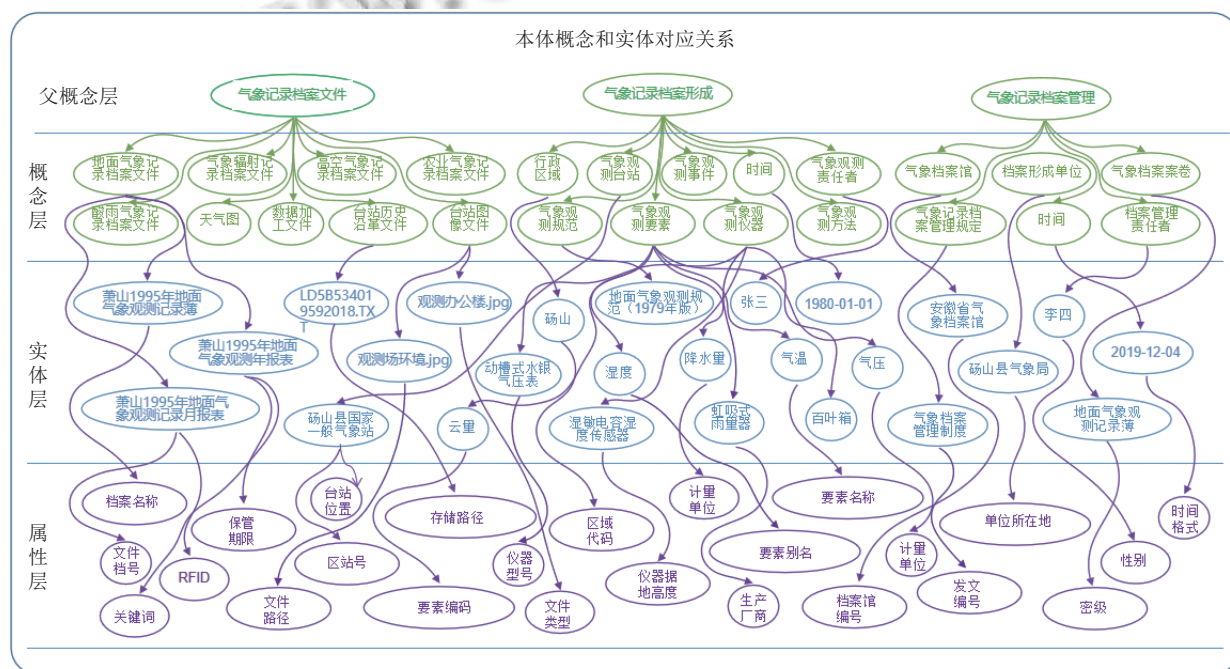


图 7 知识模型图

3.5.3 气象记录档案知识推理

知识推理是指从已构建的气象记录档案关系数据出发,进行合理推理,发现和建立新关联,从而丰富和完善气象记录档案知识图谱,提升气象记录档案智能化利用能力。

3.5.4 气象记录档案知识更新

气象档案知识更新可分为概念层更新和实体层更新。随着新的观测仪器、观测方式和观测方法的不断

发展,新的气象记录档案概念将产生,老概念将变化.每年新的档案资源不断地被收集到档案馆,档案实体必将逐年增加,这些概念和实体的变化必将推进气象记录档案知识图谱不断更新.概念层更新是将新的概念通过知识融合加入到概念层中,实体层更新是将每年新接收的档案实体、关系以及属性值加到实体库中.

3.6 气象记录档案知识图谱存储

气象档案知识图谱分为本体库存储和实体库存储。

本体库存储一般采用 MySQL 来存储, 实体库存储一般采用 Neo4j 来存储. MySQL 作为通用的一种关系型数据库, 它以表的形式存储气象记录档案本体概念. Neo4j 是一种广泛应用于知识图谱领域的图形数据库^[29], 它以节点和边的形式来存储实体库, 一个节点表示一个具体的实体, 边则表示实体与实体之间的关系, 这种实体和实体之间的关系以图的形式展示出来. 以下是查询节点关系的代码.

```
@AutoLog(value="根据节点 ID 查询节点的关系及目标节点")
@ApiOperation(value="根据节点 ID 查询节点的关系及目标节点",
notes="根据节点 ID 查询节点的关系及目标节点")
@GetMapping("/getById")
```

```
public Result getDataById(Integer modelId, Integer conceptId, String
conceptCode, String nodeId) {
    Neo4jNode dataById = kgDataService.getDataById(modelId,
conceptId, conceptCode, nodeId);
    Neo4jNode neo4jNode = dataById;
    return Result.ok(neo4jNode);
}
```

3.7 气象记录档案知识图谱应用

3.7.1 档案本身的智能检索

基于以档案为核心的关联关系网, 沿着边呈发散状关联大量信息, 为用户提供从档案的形成、内容到管理的相关信息(图8), 展示出档案馆目前馆藏的各种类型的档案及包含的内容, 方便用户快速查找档案.

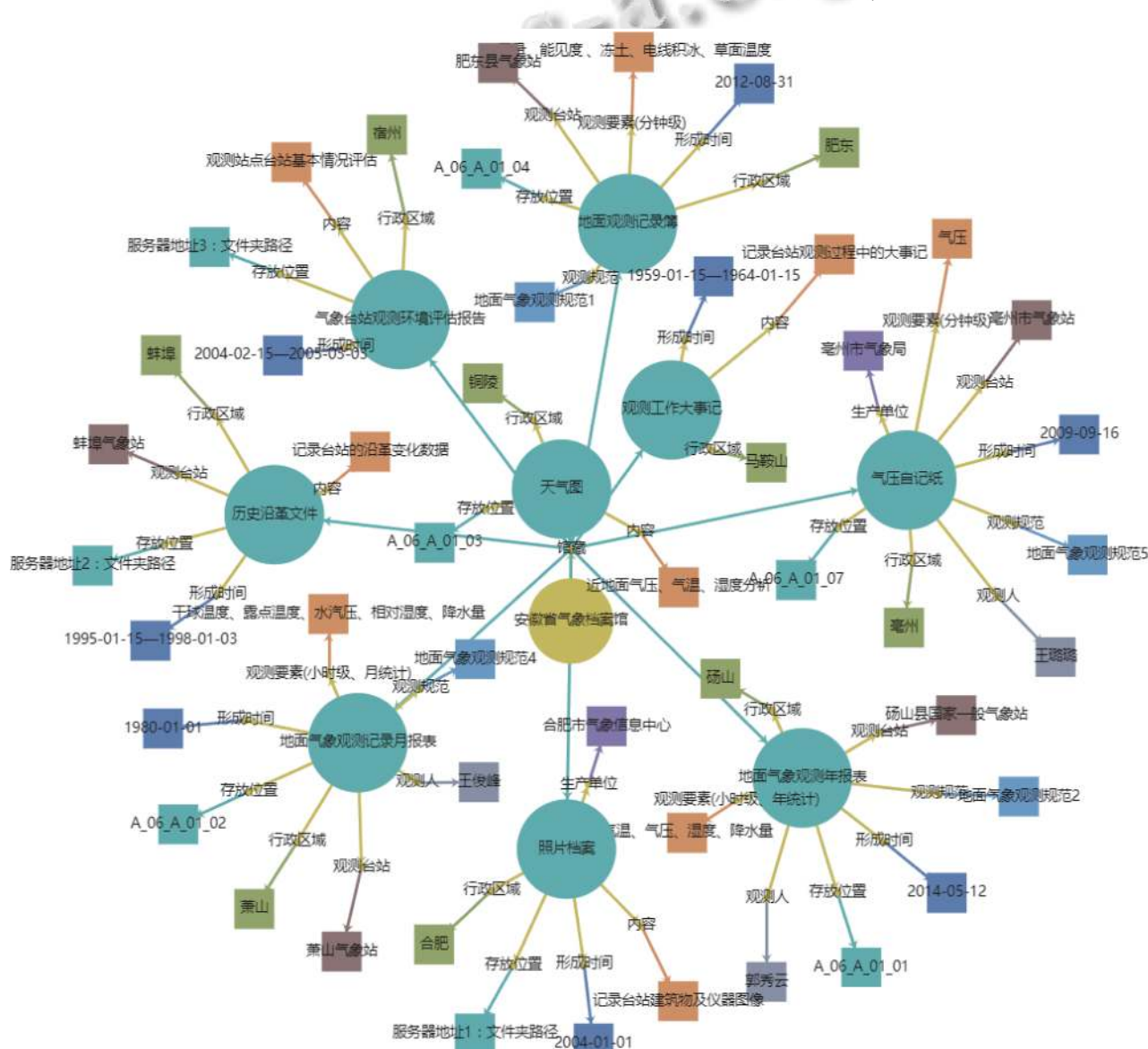


图8 档案检索实例图

3.7.2 观测要素的智能检索

基于观测要素与观测要素, 观测要素与档案、观测台站之间的关联关系, 以获得包含同一个观测要

素、不同观测频次、不同观测时间的档案(图9). 满足用户以某地某时某观测要素的值为查询档案的条件, 的逆向查找需求.

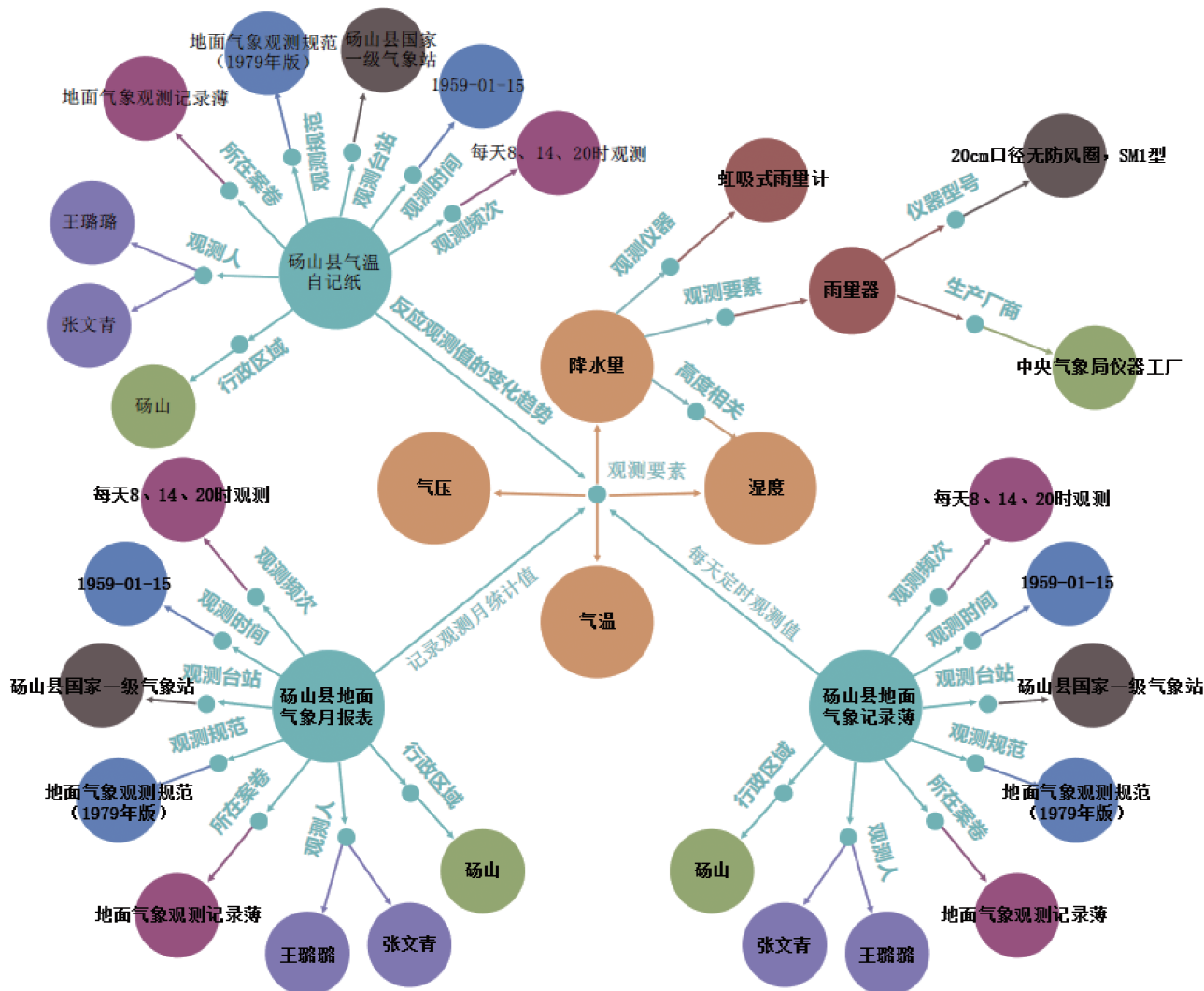


图9 档案要素检索实例图

4 结论

目前, 各类气象档案资源之间关联关系还未被深度挖掘出来. 气象档案具有专业性强、结构复杂且数据量大等特点, 如何面向社会化利用需求建立气象档案知识模型, 从多源异构的气象档案观测数据中定义气象档案概念、实体、属性以及相互之间的关系是构建气象档案知识图谱的难点.

本文基于馆藏气象记录档案资源提出了一种气象记录档案知识图谱的构建方法, 并实例展示了构建的气象记录档案知识图谱, 是气象领域和档案领域知识

图谱构建的一次探索, 对构建整个气象和档案领域知识图谱提供了参考, 期望后续有更多的学者关注和开展气象和档案信息领域的知识图谱建设研究.

参考文献

- 1 Neches R, Fikes RE, Finin T, *et al.* Enabling technology for knowledge sharing. *AI Magazine*, 1991, 12(3): 36-56.
- 2 Gruber TR. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 1995, 43(5-6): 907-928.

- 3 Studer R, Benjamins VR, Fensel D. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 1998, 25(1-2): 161-197.
- 4 宣腾. 区域医疗本体知识库构建及其语义应用 [硕士学位论文]. 成都: 电子科技大学, 2013.
- 5 赵小兵, 邱莉榕, 赵铁军. 多民族语言本体知识库构建技术. *中文信息学报*, 2011, 25(4): 71-74. [doi: 10.3969/j.issn.1003-0077.2011.04.015]
- 6 陈梅, 吕晓晶, 刘杨. 对大气领域的本体建模. *兰州交通大学学报*, 2015, 34(4): 1-5, 17. [doi: 10.3969/j.issn.1001-4373.2015.04.001]
- 7 何险峰, 张祥锋, 郑利娟, 等. 气象灾害本体设计. *气象科技*, 2012, 40(6): 1007-1012, 1017. [doi: 10.3969/j.issn.1671-6345.2012.06.023]
- 8 罗炜, 胡友彬, 孔华武, 等. 气象数据集元数据的本体推理研究与实现. *电脑与信息技术*, 2012, 20(2): 4-7, 55. [doi: 10.3969/j.issn.1005-1228.2012.02.002]
- 9 刘家玮, 刘波, 沈岳. 知识图谱在农业信息服务中的应用进展. *软件*, 2015, 36(3): 26-30. [doi: 10.3969/j.issn.1003-6970.2015.03.006]
- 10 张德政, 谢永红, 李曼, 等. 基于本体的中医知识图谱构建. *情报工程*, 2017, 3(1): 35-42.
- 11 陈亚东, 鲜国建, 寇远涛, 等. 我国苹果产业知识图谱构建研究. *中国农业资源与区划*, 2017, 38(11): 40-45. [doi: 10.7621/cjarrp.1005-9121.20171106]
- 12 王巍巍, 王志刚, 潘亮铭, 等. 双语影视知识图谱的构建研究. *北京大学学报(自然科学版)*, 2016, 52(1): 25-34.
- 13 王艳博. 图书馆学知识图谱分析. *科技情报开发与经济*, 2015, 25(21): 84-86.
- 14 尹亮, 何明利, 谢文波, 等. 装备-标准知识图谱的过程建模研究. *计算机科学*, 2018, 45(6A): 502-505.
- 15 邵丹, 林婉清, 杨惠, 等. 2005-2017年留守儿童心理健康研究热点知识图谱. *西南师范大学学报(自然科学版)*, 2018, 43(6): 142-147.
- 16 周亦, 周明全, 王学松, 等. 大数据环境下历史人物知识图谱构建与实现. *系统仿真学报*, 2016, 28(10): 2560-2566.
- 17 苏松锦, 刘金福, 兰思仁, 等. 黄山松研究综述 (1960-2014) 及其知识图谱分析. *福建农林大学学报(自然科学版)*, 2015, 44(5): 478-486.
- 18 李薇, 肖仰华, 汪卫. 基于中文知识图谱的人物实体识别. *计算机工程*, 2017, 43(3): 225-231, 240. [doi: 10.3969/j.issn.1000-3428.2017.03.038]
- 19 蒋秉川, 万刚, 许剑, 等. 多源异构数据的大规模地理知识图谱构建. *测绘学报*, 2018, 47(8): 1051-1061. [doi: 10.11947/j.AGCS.2018.20180113]
- 20 崔洁, 陈德华, 乐嘉锦. 基于 EMR 的乳腺肿瘤知识图谱构建研究. *计算机应用与软件*, 2017, 34(12): 122-126. [doi: 10.3969/j.issn.1000-386x.2017.12.023]
- 21 王雅晶, 朱化彬, 王雅春, 等. 世界奶牛繁殖研究的知识图谱分析. *畜牧兽医学报*, 2014, 45(3): 494-501. [doi: 10.11843/j.issn.0366-6964.2014.03.022]
- 22 姜赢, 张婧, 朱玲萱. 基于本体的家谱知识图谱模型及检索系统. *电子设计工程*, 2017, 25(12): 161-165. [doi: 10.3969/j.issn.1674-6236.2017.12.038]
- 23 雷洁, 赵瑞雪, 李思经, 等. 科研档案管理知识图谱构建研究. *科技管理研究*, 2020, 40(11): 162-169. [doi: 10.3969/j.issn.1000-7695.2020.11.023]
- 24 王建勋, 华丽, 邓世超, 等. 基于 CiteSpace 国内干旱遥感监测的知识图谱分析. *干旱区地理*, 2019, 42(1): 154-161.
- 25 王萍, 刘涛, 杜萍, 等. 2000-2017年中国灾害风险研究的知识图谱分析. *自然灾害学报*, 2019, 28(4): 169-177.
- 26 李莉, 崔美姬. 气象文献知识图谱构建. *计算机科学与应用*, 2018, 8(3): 242-251.
- 27 刘宇松. 本体构建方法和开发工具研究. *现代情报*, 2009, 29(9): 17-24.
- 28 杨玉基, 许斌, 胡家威, 等. 一种准确而高效的领域知识图谱构建方法. *软件学报*, 2018, 29(10): 2931-2947. [doi: 10.13328/j.cnki.jos.005552]
- 29 潘理虎, 张佳宇, 张英俊, 等. 煤矿领域知识图谱构建. *计算机应用与软件*, 2019, 36(8): 47-54, 59. [doi: 10.3969/j.issn.1000-386x.2019.08.009]