

DOI:10.3876/j.issn.1000-1980.2021.01.005

领域知识图谱研究进展及其在水利领域的应用

冯 钧¹, 杭婷婷^{1,2}, 陈 菊¹, 王云峰¹, 王秉发¹, 张 涛¹

(1. 河海大学计算机与信息学院, 江苏 南京 211100;
2. 无人机开发及数据应用安徽高校联合重点实验室, 安徽 马鞍山 243031)

摘要: 首先总结现有领域知识图谱的研究现状。其次, 介绍领域知识图谱的发展趋势。然后, 梳理水利领域知识图谱的构建难点, 提出包含知识表示、抽取、融合、推理和存储等关键模块的水利领域知识图谱研究框架, 并简要概括上述各模块的研究内容。最后, 指出领域知识图谱构建存在的表示形式单一、抽取样本稀少、多源知识冲突、规则表示困难和数据管理低效等问题, 认为合理化表示、准确全面抽取、实时性融合、可解释推理和高性能存储是下一步水利知识图谱的研究方向。
关键词: 领域知识图谱; 水利领域; 大数据; 知识表示; 知识抽取; 知识融合; 知识推理; 知识存储
中图分类号: TP391.1 **文献标志码:** A **文章编号:** 1000-1980(2021)01-0026-09

Research status of domain knowledge graph and its application in water conservancy

FENG Jun¹, HANG Tinting^{1,2}, CHEN Ju¹, WANG Yunfeng¹, WANG Bingfa¹, ZHANG Tao¹

(1. College of Computer and Information, Hohai University, Nanjing 211100, China;
2. Key Laboratory of Unmanned Aerial Vehicle Development and Data Application of
Anhui Higher Education Institutes, Maanshan 243031, China)

Abstract: Firstly, this study summarized the current research status of the domain knowledge graph. Secondly, the development trend of the domain knowledge graph was introduced. Then, this study sorted out some difficulties in the construction of water conservancy knowledge graph, proposed a research framework including main modules such as knowledge representation, extraction, fusion, reasoning, and storage, and briefly summarized the research content of each module. Finally, the construction of domain knowledge graph encountered some problems, such as the single representation, the extraction sample sparse, the multi-source knowledge conflict, the rule representation difficulty, and the inefficient data management. Therefore, the rationalized representation, accurate and comprehensive extraction, real-time fusion, interpretable reasoning, and high-performance storage are regarded as the next research direction of water conservancy knowledge graph.

Key words: domain knowledge graph; water conservancy; big data; knowledge representation; knowledge extraction; knowledge fusion; knowledge reasoning; knowledge storage

随着人工智能研究的不断发展, 人工智能的主要发展方向经历了从拥有快速计算和记忆存储能力的运算智能, 到拥有视觉、听觉、触觉等感知能力的感知智能, 正在迈向拥有理解和思考能力的认知智能。知识图谱和以知识图谱为代表的知识工程系列技术是认知智能的核心。知识图谱本质是一种揭示实体之间关系的语义网络, 可以对现实世界的事务及其相关关系进行形式化描述^[1], 它强大的语义处理和互联组织能力, 对有效描述数据间的关联关系进而打破信息孤岛的局面具有一定的现实意义。目前, 在一些领域已经出现了面向领域的知识图谱, 例如电影领域的 IMDB^[2]、生物医学领域的 BMKN^[3]、新闻领域的 ECKG^[4]、健康领域的 SHKG^[5]等。从已有的领域知识图谱看, 构建领域知识图谱需要借鉴通用知识图谱的方法, 同时还需要依靠特

基金项目: 国家重点研发计划(2018YFC0407901); 安徽省高等学校自然科学研究重点项目(KJ2019A1277)
作者简介: 冯钧(1969—), 女, 教授, 博士, 主要从事数据管理、智能数据处理与数据挖掘、水利信息化研究。E-mail: fengjun@hhu.edu.cn
通信作者: 杭婷婷, 副教授。E-mail: httsf@hhu.edu.cn
引用本文: 冯钧, 杭婷婷, 陈菊, 等. 领域知识图谱研究进展及其在水利领域的应用[J]. 河海大学学报(自然科学版), 2021, 49(1): 26-34.
FENG Jun, HANG Tinting, CHEN Ju, et al. Research status of domain knowledge graph and its application in water conservancy[J].
Journal of Hohai University(Natural Sciences), 2021, 49(1): 26-34.

定行业数据,具有特定的行业意义,领域知识图谱的构建是当前知识图谱研究的一个重要方向和趋势。

随着水利信息化及其水利信息技术的发展,水利领域长期业务实践积累了实时监测、遥感遥测、水文气象、水利工程、社会经济等多源异构水利大数据,实现了水利监测从点到面的转变,从静态到动态的拓展。随着信息采集和传输技术的飞速发展和领域信息化的进程,领域数据不断更新,数据量日益增加,数据间语义不一致也屡见不鲜。多源异构数据呈现出海量、动态、内容多样、处理复杂的特点。如何让分布存储管理的、语义各异的数据能够互联,充分发掘领域数据价值,促进信息资源的高效利用,是推进智慧水利^[6]的关键,也是水利信息资源查询推荐,语义搜索,智慧防汛^[7-8]和智慧水资源管理^[9]等应用的基础,对于提高水利领域智能化管理水平、辅助管理者进行决策分析具有非常重要的意义^[10]。因此,水利领域知识图谱研究既具有重要的理论意义,也具有显著的实用价值。

本文总结领域知识图谱构建的研究现状,包括构建方式、应用现状等方面的进展;介绍近年来领域知识图谱构建的发展趋势;对水利领域知识图谱构建工作进行展望,提出研究框架和具体的研究内容。

1 领域知识图谱构建研究现状

知识图谱按照覆盖范围可分为通用知识图谱和领域知识图谱。通用知识图谱面向通用领域,以常识性知识为主,其构建过程高度自动化。其关联的大多数是静态的、客观的、明确的三元组事实性知识。领域知识图谱面向某一特定领域,以行业数据为主,其构建过程半自动化。其关联的不仅包含静态知识,也涉及一些动态知识。本文主要探讨领域知识图谱构建。

1.1 领域知识图谱的构建方式

在领域知识图谱的构建方式方面,目前主要有自顶向下和自底向上 2 种构建方式。自顶向下方式是针对特定的行业,由该行业专家定义好顶层本体与数据模式,再将抽取到的实体加入到知识库中。国内外现有的本体建模工具以 Protégé、PlantData 为代表。Protégé 是一套基于 RDF(S),OWL 等语义网规范的开源本体编辑器,拥有图形化界面,适用于原型构建场景。PlantData 是一款商用知识图谱智能平台软件。该软件提供了本体概念类、关系、属性和实例的定义和编辑,屏蔽了具体的本体描述语言,用户只需在概念层次上进行领域本体模型的构建,使得建模更加便捷。自底向上方式主要依赖开放链接数据集和百科网站,从这些结构化的知识中进行自动学习,直接将抽取数据中发现的实体、关系以及属性合并到知识图谱中^[11]。自顶向下的方法有利于抽取新的实例,保证抽取质量。而自底向上的方法则能发现新的模式。因此,目前大部分领域知识图谱的构建方式是自顶向下和自底向上相结合的方式。

1.2 领域知识图谱的应用现状

领域知识图谱通常用来辅助各种复杂的分析应用或决策支持。目前,在大多数领域中均存在领域知识图谱的应用。因为应用场景和应用目的不同,不同领域的应用形式也有所不同。下面将从知识应用的角度出发,介绍相关领域知识图谱的应用现状。

a. 电商知识图谱的应用。电商知识图谱的主要应用场景就是导购。导购就是让消费者更容易找到他想要的东西。为此,电商知识图谱学习了大量的行业规范与国家标准,对一些专业词汇进行了更细致的解决。另外,它还可以从公共媒体和专业社区中识别出近期热词。当消费者输入相关热词之后,可以出现跟热词相关的商品。与此同时,电商知识图谱还可以通过场景构建,实现与场景相关的商品推荐。

b. 医疗知识图谱的应用。医疗知识图谱的主要应用包括医疗过程智能辅助、医学科研以及患者服务等方面。其中医疗过程智能辅助是通过医疗知识图谱实现临床辅助决策、合理用药等智能服务。医学科研是基于医疗知识图谱,辅助医务工作者实现疾病风险预测、药物研发等应用服务。患者服务是根据患者过去的就医记录以及相关的医疗知识,为患者提供健康知识推送和健康评估等日常服务。

c. 企业知识图谱的应用。企业知识图谱通过异常关联挖掘、最终控制人等方式为行业客户提供风险管理。其中异常关联挖掘是通过路径分析、关联探索等操作,挖掘企业之间的异常关联,减少企业经营风险和资金风险。最终控制人是寻找持股比例最大的股东,最终追溯至自然人或者国有资产管理部门,向行业用户提供更准确的智能服务。

d. 创投知识图谱的应用。创投知识图谱主要应用包含知识检索和可视化决策支持。其中知识检索是由机器完成用户搜索意图识别,向用户提供准确检索答案。可视化决策支持是通过图谱可视化技术对公司

的全方位信息,投资机构的投资偏好等进行展示,为投融资决策提供支持。

总的来说,知识图谱与各行业的深度融合已经成为一个重要趋势。在这一过程中,涌现出一系列的领域应用,可以解决行业痛点问题。

2 领域知识图谱构建的发展趋势

领域知识图谱构建的主要过程包括知识表示、知识抽取、知识融合、知识推理和知识存储等 5 个方面。尽管目前相关原理和应用都已经取得了较好的成果,但仍在快速发展之中。近年来,领域知识图谱的发展趋势发生了一系列的变化,主要表现在:

a. 在知识表示方面,现阶段一般采用三元组表达事实知识。但是,在决策、推理等相关应用中,需要依赖于大量专家知识、动态知识进行辅助判断,而专家知识的表示已经超出了常规知识表示的范畴。在大数据的赋能下,知识表示的重心将逐步过渡到动态知识是必然趋势。

b. 在知识抽取方面,现阶段的研究主要集中在纯文本信息抽取方面。在训练样本较为丰富的情况下,基于神经网络的抽取模型可以取得较好的抽取效果。但是,领域知识多数处于小样本、零样本以及面向开放域的抽取环境下,知识抽取的重心将逐步过渡到小样本、零样本信息抽取是必然趋势。

c. 在知识融合方面,现阶段的研究主要聚焦于知识融合过程中的某一部分或者只关注知识融合的模式,冲突检测、实体对齐、属性对齐和属性真值发现过程的研究缺乏连续性。另外,随着大量新增知识的更新,知识融合的重心将逐步过渡到新增知识的实时融合是必然趋势。

d. 在知识推理方面,现阶段的研究主要采用基于规则、逻辑的方法挖掘领域图谱中隐含的知识或纠正错误的知识。但是,该方法对规则的依赖度高。图神经网络是连接主义与符号主义的有机结合,不仅使深度学习模型能够应用在图这种非欧几里德结构上,还为深度学习模型赋予了一定的因果推理能力^[12]。知识推理的重心将逐步过渡到面向图结构的深度推理是必然趋势。

e. 在知识存储方面,现阶段一般利用传统的关系型数据库存储领域知识图谱。但是,针对低选择性、复杂查询效率低的问题,知识存储的重心将逐步过渡到分布式 RDF 查询优化是必然趋势。

3 水利领域知识图谱构建

3.1 水利领域知识图谱构建的难点

a. 在水利知识表示方面,领域应用不仅需要静态知识,也需要动态知识。如何对抽取出来的静态知识和动态知识进行合理表示是当前面临的主要技术难点。另外,有很多知识和事实有时间和空间条件,从时空纬度扩展知识表示也是需要解决的技术难点。

b. 在水利知识抽取方面,纯文本信息抽取是当前面临的主要难点。部分文本抽取算法在公共数据集上取得了较好的实验结果,但普遍存在应用到水利领域中扩展性不好等问题。难点在于如何根据领域知识图谱的小样本特性,构建基于小样本的有效模型。

c. 在水利知识融合方面,主要存在以下难点:(a)实体对应不准确,同一实体名在不同数据源中常含有歧义,数据源中存在严重的多源指代问题^[11];(b)不同数据源关于相同实体的相同属性存在表述差异^[13];(c)不同数据源为同一实体的同一属性提供的属性值存在冲突^[11]。

d. 在水利知识推理方面,由于现有水利领域的应用需要高准确性地从图谱中获取信息,因此基于描述逻辑和规则的推理方法能有效用于水利知识推理。难点在于如何设计基于一阶谓词逻辑的推理规则用于知识推理。

e. 在水利知识存储方面,主要存在以下难点:(a)随着水利数据不断丰富,RDF 数据规模日益增加,现有的集中式数据管理系统难以满足对大规模 RDF 数据的存储和查询性能需求,需要高性能的分布式数据管理系统^[14]来实现对大规模 RDF 数据的存储、索引和查询处理;(b)现有的分布式数据管理系统,对特定类型的查询进行了优化^[15],但对水利领域常涉及的低选择性、大直径查询的查询效率低;(c)现有的分布式数据管理系统不能动态适应工作负载^[16]的变化。

3.2 水利领域知识图谱的总体框架

为解决上述水利领域知识图谱构建研究的 5 个难点,并实现建立水利领域知识图谱的目标,本文提出了

如图 1 所示的研究框架。在该研究框架下,首先对水利知识表示进行研究,建立 2 种不同的表示形式;其次,针对不同类型的水利数据,研究相对应的水利知识抽取方法;然后,研究了水利知识融合和推理的具体方法;最后,在充分利用水利大数据和相关存储技术的基础上,对水利领域知识进行存储,支撑相关应用。

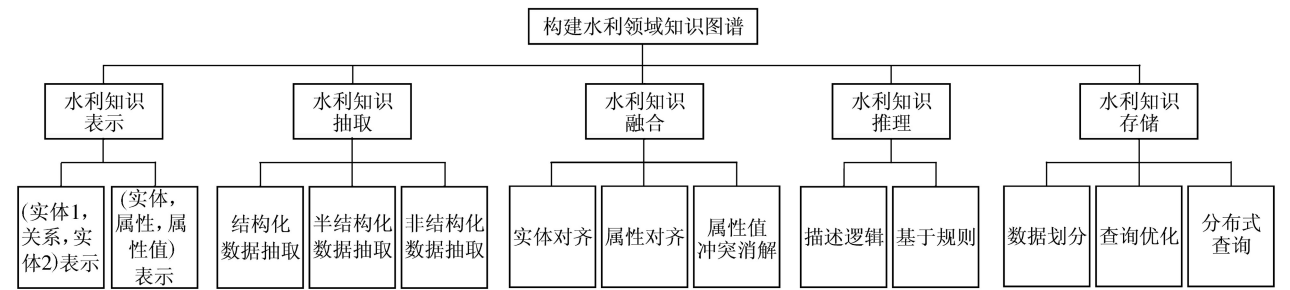


图 1 水利领域知识图谱构建研究框架

Fig.1 Modeling framework of domain knowledge graph in water conservancy

3.3 水利领域知识图谱构建的研究内容

水利领域知识图谱的构建流程可以被归纳为 5 个模块,即水利知识表示、水利知识抽取、水利知识融合、水利知识推理以及水利知识存储。水利知识表示是将水利知识表达成计算机可存储、可计算的结构化知识。水利知识抽取可以从大量结构化、半结构化和非结构化的水利数据中提取知识要素。水利知识融合可以消除实体、关系、属性与对象之间的歧义,并为水利知识图谱更新旧知识或补充新知识。水利知识推理是在已有水利知识的基础上进一步挖掘隐含知识或者缺失事实,从而丰富、扩展水利知识库。水利知识存储是设计有效的存储模式来支持对水利数据的有效管理。

3.3.1 水利知识表示

三元组是知识图谱的一种通用表示形式^[17],由 2 个具有语义连接关系的水利实体和实体间关系组成,是水利知识的直观表示。三元组的基本形式主要包括(实体 1,关系,实体 2)和(实体,属性,属性值)等。概念主要指水利对象类,例如水资源分区、流域分区、湖泊、测站、河流、水库及水电站等;实体是知识图谱中的最基本元素,例如湖西区、长江流域、汾湖、吴江水厂、太浦河、青山水库、龙头水电站等;关系存在于不同实体之间,例如属于、位于、流入、包含等;属性主要指对象可能具有的特征及参数,例如湖泊代码、湖泊名称、跨界类型等;属性值指对象特定属性的值,例如 FH407、FHBA1B00000M、跨省等。

通过一个全局唯一的 ID 号来标识实体,实体间内在特征通过属性-属性值来进行刻画,实体之间的关联通过关系来描述。三元组的存在表示一个已有的事实。例如关于太湖的描述为:太湖流域包括江苏省苏南地区、浙江省杭嘉湖地区、上海市大陆部分(不含崇明、长兴、横沙三道)和安徽省宣城的小部分地区,总面积 3.69 万 km²。流域水面积 5 551 km²;河道总长约 12 万 km,河道密度达 3.3 km/km²。流域地形呈周边高、中间低的碟状地形,地势平坦,河道比降小,水流流速缓慢。太湖的描述可以通过表 1 的三元组进行表示。

所有三元组合可以并构成一个图(图 2),其中节点表示实体,有向边表示实体之间的关系,不同的关系边的标签不同。

3.3.2 水利知识抽取

在水利信息技术飞速发展的今天,水利知识大量存在于水利信息系统的结构化数据、半结构化的表格、网页以及非结构化的文本数据中。针对不同类型的水利数据,采用不同的知识抽取方法。对于结构化数据,研究基于 D2R 技术的知识图谱构建方法,利用信息系统中的结构化对象数据,抽取出静态对象及其相关关

表 1 太湖描述的三元组表示			
Table 1 Triple representation of Taihu Lake			
基本形式	实体 1	关系	实体 2
(实体 1, 关系, 实体 2)	太湖流域	包括	苏南地区
	太湖流域	包括	杭嘉湖地区
	太湖流域	包括	上海市大陆部分
	太湖流域	包括	宣城的小部分地区
基本形式	实体	属性	属性值
(实体, 属性, 属性值)	太湖流域	总面积	3.69 万 km ²
	太湖流域	水面积	5 551 km ²
	太湖流域	河道总长	12 万 km
	太湖流域	河道密度	3.3 km/km ²
	太湖流域	地形	碟状
	太湖流域	地势	平坦
	太湖流域	河道比降	小
	太湖流域	水流流速	缓慢

系。结构化数据抽取如图 3(a) 所示,基本步骤包括:
(a)通过分析关系型数据库判断可以建立联系的 2 张表是否有外键关联。如果没有外键关联,需要人工设置外键或者在映射文件中写入外键。
(b)建立了外键关系之后,将 2 张表映射成 RDF 之后就可以实现语义互联。通过上述一系列操作,可以将 2 个实体之间存在的关系进行合理表示。

对于半结构化数据,利用包装器将分布在互联网上半结构化的 HTML 页面中的属性和属性值抽取出来。半结构化数据抽取如图 3(b) 所示,基本步骤包括:
(a)HTML 页面清洗及解析。将页面转换为 DOM 树形结构。
(b)页面去噪。去除页面中与主题信息无关的其他信息。
(c)包装器自动生成。自动获取需求信息节点的 XPath 路径,定义规则模板,结合 XPath 路径表达式实现抽取规则的自动构造。通过上述一系列操作,可以抽取与实体有关的属性和属性值信息。

对于非结构化数据,利用基于远程监督和神经网络的方法抽取水利文本中的知识。非结构化数据抽取如图 4(c) 所示,基本步骤包括:
(a)采用远程监督的方法利用知识库自动生成标注数据,再通过离群点检测的方法去除其中的错误标注。
(b)采用基于监督学习的神经网络方法,先在标注好的数据上进行训练,再对未标注的数据进行测试,抽取出来未标注文本中包含的实体和它们之间的关系。通过上述一系列操作,可以补充知识图谱中所需要的一些静态知识和动态知识。

3.3.3 水利知识融合

鉴于百科类网站具有一个页面围绕一个实体进行描述、页面组织结构相对统一、信息质量相对较高的特点,百科类网站成为领域知识库进行知识融合的主要数据来源^[18],其信息框中的关于实体的属性-属性值对是对该页面实体信息的高度提炼。对不同百科中描述相同实体的知识卡片进行融合,可以获得关于水利对象的更全面、质量更高的知识。

针对前述关于多知识库融合的难点,研究基于中文维基百科、百度百科、互动百科的知识卡片的水利知识融合方法。图 4 所示为水利知识融合流程。通过基于多特征的命名实体消歧、基于词典的属性对齐和基于贝叶斯分析的属性真值发现模块,消除实体、关系、属性及其对象之间的歧义,最终获得跟水利对象有关的属性及相应的属性值。

图 5 是三大百科以及本地知识库对于水利对象“太湖”融合后的查询结果。蓝色的方块代表初步形成的水利领域知识图谱,红色的方块代表中文维基百科,黄色的方块代表百度百科,绿色的方块代表互动百科。从“太湖”的融合结果可以看出,本地水利领域知识图谱提供的信息资源具备良好的行业覆盖面和行业深度,为水利知识图谱的构建提供了核心支撑。中文维基百科则更多地从专业领域对其进行描述,提供的更多是较严谨的知识。百度百科和互动百科的知识卡片存在很多重复,且覆盖的属性更符合普通大众的娱乐需要,如关于太湖的适宜游玩季节、建议游玩时长、门票价格等。

3.3.4 水利知识推理

知识推理旨在从图谱已有的知识推理得到新的事实^[19]。由于水利知识来源多样化,水利知识和数据的收集局限于终端采集方式而缺乏整体性,需要结合水利知识推理方法,来对相关知识进行补充。例如,水利领域知识图谱中存在由不同数据源得到的 2 个三元组:(太湖,出口,太浦闸)和(太浦闸,属于,太浦河),可以利用知识推理来获取新的事实知识(太湖,流入,太浦河)。目前主要的领域知识推理的方法有:基于规则推理的方法^[20]、基于本体推理的方法^[21-22]、基于表示模型的方法^[23-25]、基于神经网络的方法^[26]。

通过对水利领域的业务需求进行分析,可以发现水利领域知识图谱需要为即时查询、决策提供支撑,因此决定了水利领域知识图谱构建的高准确性要求。另外,水利领域知识图谱的层次性较强,根据管理单位、地理空间、河网管网的分层关系可以在实际应用场景中将图谱切分,以降低搜索空间。结合水利知识图谱存在的高准确性要求和可切分特点,最适合的知识推理方法是基于规则推理的方法。该方法通过结合现有的一些水利领域知识,手工定义一些推理规则,去服务水利知识推理。其具体过程如下:(a)在概念层,通过一阶谓

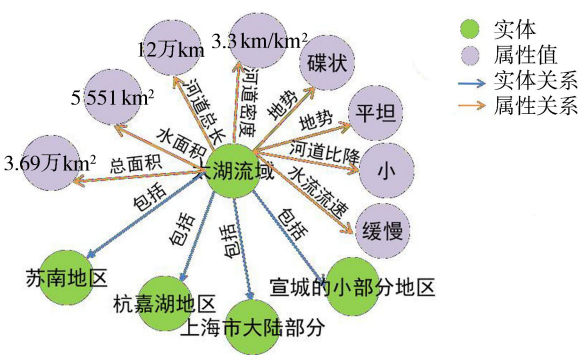
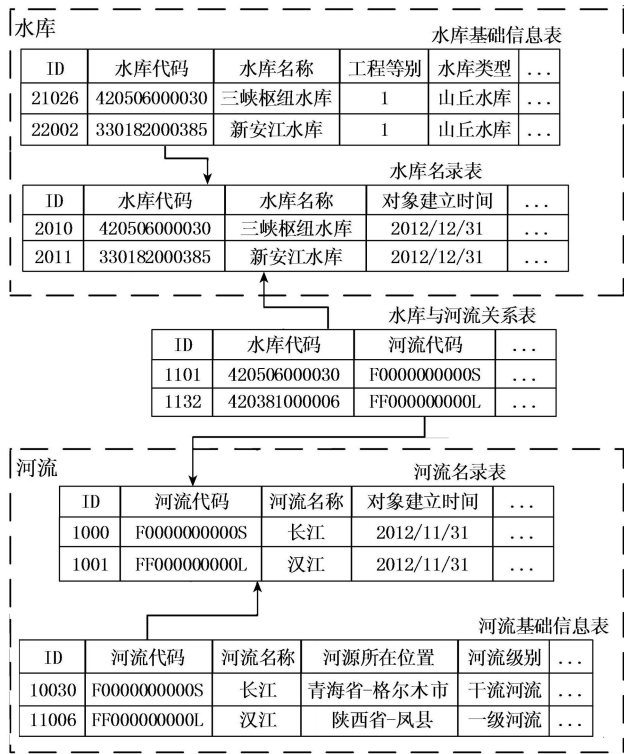
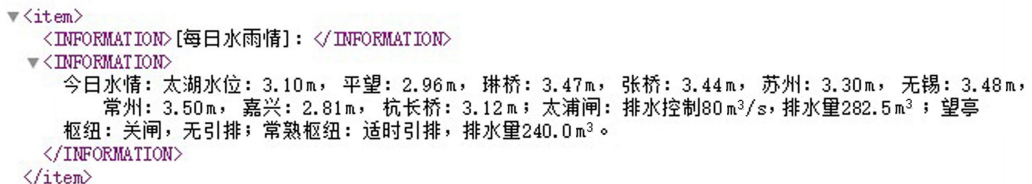
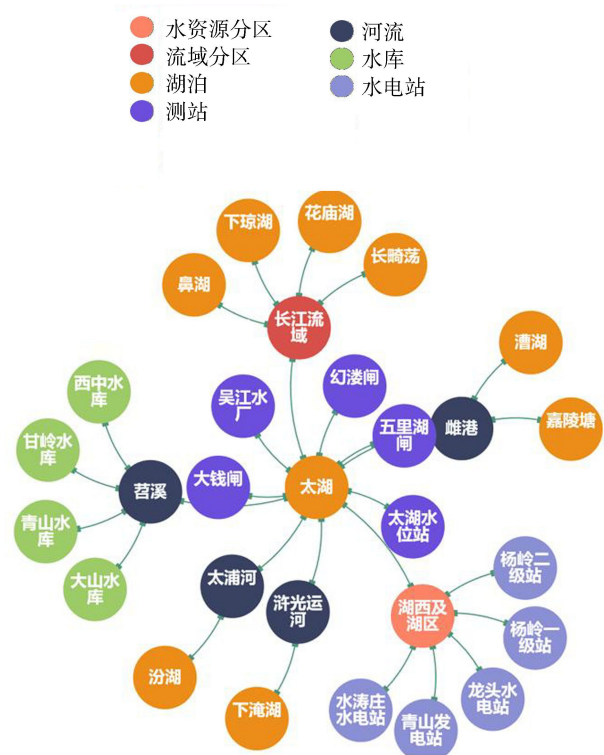


图 2 水利知识表示示意图

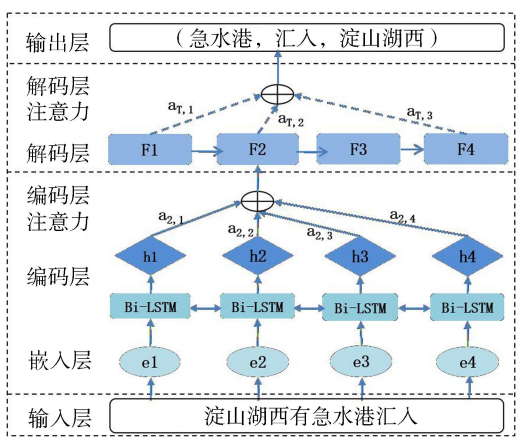
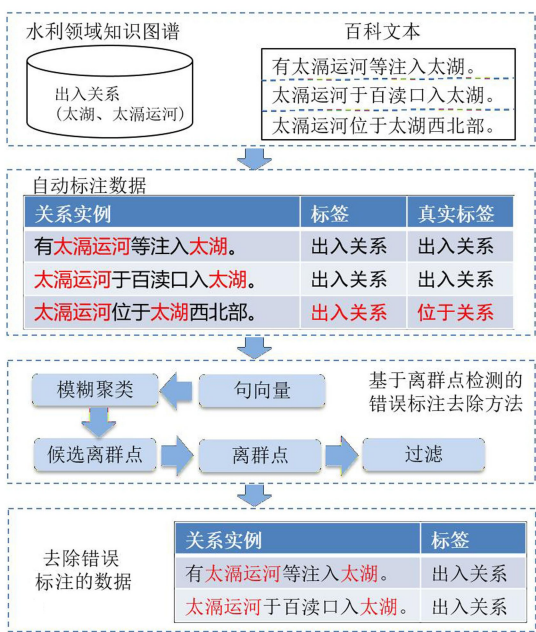
Fig. 2 Schematic diagram of knowledge representation in water conservancy



(a) 结构化数据抽取



(b) 半结构化数据抽取



(c) 非结构化数据抽取

图 3 水利知识抽取示意图

Fig. 3 Schematic diagrams of knowledge extraction in water conservancy

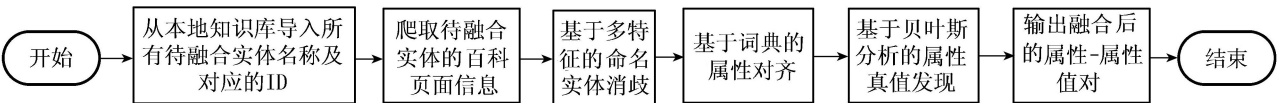


图 4 水利知识融合流程

Fig. 4 Flow chart of knowledge fusion in water conservancy

太湖F2D9B541D36B4EA9A 实体查询	
太湖	Water_Graph Ch_Wiki Baidu Interactive
属性名	属性值
湖泊名称	太湖
左下角经度(度)	119.89150000
左下角纬度(度)	30.92793000
右上角经度(度)	120.61510000
右上角纬度(度)	31.54958000
跨界类型	跨省
湖泊所在位置	江苏吴江市、苏州吴中区、苏州虎丘区、苏州相城区、无锡滨湖区、宜兴市、常州武进区、浙江长兴县、湖州吴兴区
水面面积(km ²)	2341.00
湖泊容积(万m ³)	61830.00
平均水深(m)	2.60
最大水深(m)	7.30
中文名称:	太湖
英文名称	Taihu Lake
别名	太湖、震泽、五湖
所属水系	长江水系
所属地区	中国华东
地理位置	江苏省南部、长江中下游
湖泊类型:	淡水湖
表面积	2338 km ²
位置	中华人民共和国江苏省南部和浙江省北部交界处
属于	中华人民共和国
适宜游玩季节	3-5月和9-10月
建议游玩时长	2 d
门票价格	105元
开放时间	7:00-17:30

图 5 水利知识融合示意图

Fig. 5 Schematic diagram of knowledge fusion in water conservancy

词逻辑表示定义相关推理规则。(b)在实例层,再通过实例去实例化推理规则,找到符合推理规则的关系事实。表 2 为部分推理规则及其相关含义。

3.3.5 水利知识存储

水利知识存储的优化目标是减少冗余数据的存储,提高查询的效率。为了达到上述目标,采用以下处理手段:

(a) 针对集中式系统难以满足对大规模水利 RDF 数据的存储和查询处理的问题,采用了一个无共享的集群,以分布式的方式处理大规模 RDF 数据。(b) 针对水利领域涉及的低选择性、大直径查询效率低,对查询工作负载伸缩性差的问题,研究了基于垂直划分和哈希划分的混合关系存储模式。通过监控查询工作负载中的频繁模式,使用频繁模式指导水利 RDF 数据进行增量重划分,以提高对查询工作负载的伸缩性。(c) 通过设置代价评估模型,进行代数优化和连接顺序优化,从而优化分布式查询的效率。

水利知识存储流程如图 6 所示。该流程首先对经过质量评估后的水利知识进行基于主语的哈希划分形成三元组表(TT);然后,对哈希划分后的三元组表进行垂直划分,形成只包含主语-宾语列的垂直划分表(VP);最后,通过查询监控器监控查询工作负载,挖掘频繁模式,对频繁模式所对应的垂直划分表进行半连接计算,形成频繁谓词扩展垂直划分表(FP-ExtVP)。上述不同类型的表都以 Parquet 格式存储到集群的各

表 2 水利知识推理规则

Table 2 Rules of knowledge reasoning in water conservancy

编号	推理规则	含义
1	(河流,流入,水库),(水电站,属于,水库)→(水电站,位于,河流)	水电站在水库所在的河流上
2	(泵站,拥有,取水口),(泵站口,位于,湖泊),(湖泊,属于,流域分区)→(取水口,属于,流域分区)	取水口属于泵站所在湖泊的流域分区
3	(桥梁,位于,河段),(河段,属于,河流)→(桥梁,横跨,河流)	桥梁横跨河段所属的河流

个节点上。图 6 中红色模块代表三元组表,橙色模块代表垂直划分表,蓝色模块代表频繁模式所对应的垂直划分表,绿色模块代表频繁谓词扩展垂直划分表。

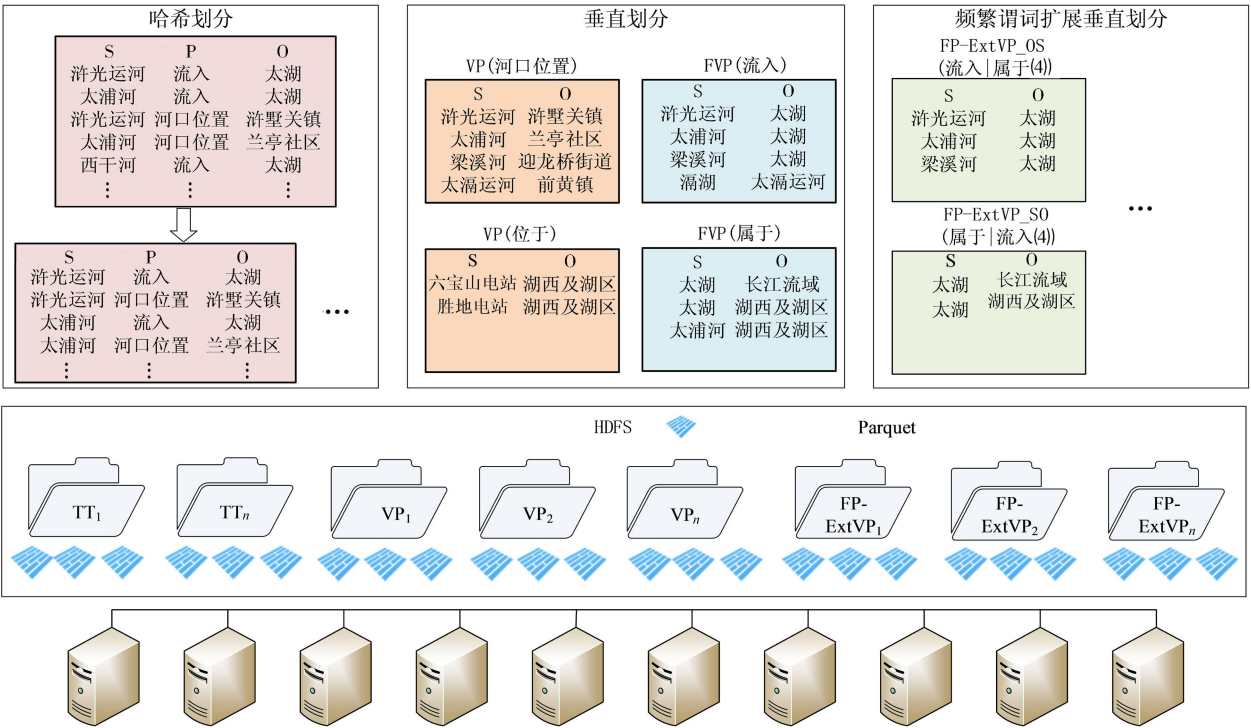


图 6 水利知识存储流程

Fig. 6 Flow chart of knowledge storage in water conservancy

4 结 语

随着水利信息化水平的逐步深入,水利领域积累了大量水利数据。通过构建水利领域知识图谱,可以为全面认知水利数据提供一种技术手段,其具有重要的理论和实际价值。本文对领域知识图谱构建的研究现状和发展趋势进行了全面调研和深入分析,并结合水利领域知识图谱构建过程中存在的难点,提出水利领域知识图谱的整体框架和主要研究内容。

知识图谱在水利领域的意义不仅仅在于它是一个水利大数据的知识库,也是支撑水利业务管理的一个主要技术手段,它能够推进水利数据自动化和智能化处理,实现水利对象的全面互联,智能应用,从而促进水利行业水利现代化。

参考文献:

[1] 金菊良,陈鹏飞,陈梦璐,等. 基于知识图谱的水资源承载力研究的文献计量分析[J]. 水资源保护,2019,35(6):14-24. (JIN Juliang,CHEN Pengfei,CHEN Menglu,et al. Bibliometric analysis of research progress on water resources carrying capacity based on knowledge map[J]. Water Resources Protection,2019,35(6):14-24. (in Chinese))

[2] IMDB Official. IMDB[EB/OL]. [2020-11-23]. <http://www.imda.com>.

[3] BAI T,GONG L,WANG Y,et al. A method for exploring implicit concept relatedness in biomedical knowledge network[J]. BMC Bioinformatics,2016,17(9):53-66.

[4] ROSPOCHER M,VAN ERP M,VOSSEN P,et al. Building event-centric knowledge graphs from news[J]. Journal of Web Semantics,2016,37:132-151.

[5] SHI L,LI S,YANG X,et al. Semantic health knowledge graph:Semantic integration of heterogeneous medical knowledge and services[J]. BioMed Research International,2017,1(4):1-12.

[6] 蔡阳. 智慧水利建设现状分析与发展思考[J]. 水利信息化,2018(4):1-6. (CAI Yang. Status analysis and development thinking of smart water conservancy construction[J]. Water Resources Information,2018(4):1-6. (in Chinese))

[7] 陈华,徐坚,盛晟,等. 面向流域防洪安全的态势图谱构建及可视化方法[J]. 河海大学学报(自然科学版),2020,48(6):408-505. (CHEN Hua,XU Jian,SHENG Sheng,et al. Construction and visualization methods of situation map for flood control in

- river basin[J]. Journal of Hohai University(Natural Sciences),2020,48(6):408-505. (in Chinese))
- [8] 巫义锐,汪浩航,魏大保,等. 时空多特征流域场景模式库构建方法[J]. 河海大学学报(自然科学版),2020,48(6):514-520. (WU Yirui,WANG Haohang,WEI Dabao,et al. Construction method of watershed scene pattern library via spatio-temporal multiple features[J]. Journal of Hohai University(Natural Sciences),2020,48(6):514-520. (in Chinese))
- [9] 冯钧,王云峰,邬炜,等. 城市内涝事理图谱构建方法及应用[J]. 河海大学学报(自然科学版),2020,48(6):479-487. (FENG Jun,WANG Yunfeng,WU Wei,et al. Construction method and application of event logic graph for urban waterlogging[J]. Journal of Hohai University(Natural Sciences),2020,48(6):479-487. (in Chinese))
- [10] 叶枫,张鹏,夏润亮,等. 基于新一代大数据处理引擎 Flink 的“智慧滁河”系统[J]. 水资源保护,2019,35(2):90-94. (YE Feng,ZHANG Peng,XIA Runliang,et al. “Smart Chuhe River” system based on Flink,a new generation of big data processing engine[J]. Water Resources Protection,2019,35(2):90-94. (in Chinese))
- [11] 侯梦薇,卫荣,陆亮,等. 知识图谱研究综述及其在医疗领域的应用[J]. 计算机研究与发展,2018,55(12):2587-2599. (HOU Mengwei,WEI Rong,LU Liang,et al. Research review of knowledge graph and its application in medical domain[J]. Journal of Computer Research and Development,2018,55(12):2587-2599. (in Chinese))
- [12] LAMB L,GARCEZ A,GORI M,et al. Graph neural networks meet neural-symbolic computing:a survey and perspective[C]//Proc of the 29th International Joint Conference on Artificial Intelligence Survey track. San Francisco:Morgan Kaufmann,2020:4877-4884.
- [13] WANG H,FANG Z,ZHANG L,et al. Effective online knowledge graph fusion[C]//Proceeding of the International Semantic Web Conference. Heidelberg:Springer-Verlag,2015:286-302.
- [14] 王鑫,邹磊,王朝坤,等. 知识图谱数据管理研究综述[J]. 软件学报,2019,30(7):2139-2174. (WANG Xin,ZOU Lei,WANG Chaokun,et al. Research on knowledge graph data management;a survey[J]. Journal of Software,2019,30(7):2139-2174. (in Chinese))
- [15] HASSAN M,BANSAL S K. Data partitioning scheme for efficient distributed RDF querying using Apache Spark[C]//Proceeding of the 13th IEEE International Conference on Semantic Computing. New Jersey:IEEE,2019:24-31.
- [16] HARB R,ABDELAZIZ I,KALNIS P,et al. Accelerating SPARQL queries by exploiting hash-based locality and adaptive partitioning[J]. The VLDB Journal,2016,25(3):355-380.
- [17] WANG Q,MAO Z,WANG B,et al. Knowledge graph embedding:a survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering,2017,29(12):2724-2743.
- [18] 徐波. 百科知识图谱构建[EB/OL]. [2020-11-23]. <http://zhuanlan.zhihu.com/p/30871301>.
- [19] 官赛萍,靳小龙,贾岩涛,等. 面向知识图谱的知识推理研究进展[J]. 软件学报,2018,29(10):74-102. (GUAN Saiping,JIN Xiaolong,JIA Yantao,et al. Knowledge reasoning over knowledge graph:a survey[J]. Journal of Software,2018,29(10):2966-2994. (in Chinese))
- [20] SUDA M,WEIDENBACH C,WISCHNEWSKI P. On the saturation of YAGO[C]//Proceeding of the International Joint Conference on Automated Reasoning. Berlin:Springer-Verlag,2010:441-456.
- [21] BUHMANN L,LEHMANN J. Pattern based knowledge base enrichment[C]//Proceeding of the International Semantic Web Conference. Berlin:Springer-Verlag,2013:33-48.
- [22] JIANG Shangpu,LOWD D,DOU Dejing. Learning to refine an automatically extracted knowledge base using markov logic[C]//Proceeding of the 12th International Conference on Data Mining. New Jersey:IEEE,2012:912-917.
- [23] BORDES A,USUNIER N,GARCIA-DURAN A,et al. Translating embeddings for modeling multi-relational data[C]//Proceeding of the 26th International Conference on Neural Information Processing Systems. New York:ACM Press,2013:2787-2795.
- [24] CHEN Xuelu,CHEN Muhao,SHI Weijia,et al. Embedding uncertain knowledge graphs[C]//Proceeding of the 33th AAAI Conference on Artificial Intelligence. Menlo Park:AAAI Press,2019,33(1):3363-3370.
- [25] ZHANG Wen,PAUDEL B,WANG Liang,et al. Iteratively learning embeddings and rules for knowledge graph reasoning[C]//Proceeding of the 28th World Wide Web Conference on World Wide Web. New York:ACM Press,2019:2366-2377.
- [26] SOCHER R,CHEN Danqi,MANNING C D,et al. Reasoning with neural tensor networks for knowledge base completion[C]//Proceeding of the 26th International Conference on Neural Information Processing Systems. New York:ACM Press,2013:926-934.

(收稿日期:2020-11-24 编辑:刘晓艳)