数字乡村建设领域知识图谱构建及应用研究

孔瑞琪，赵蕾*

（西安财经大学信息学院，陕西西安　710000）

摘　要：【目的】数字乡村建设领域包含大量复杂而分散的信息，使得科研、政策制定和实际应用面临着知识获取和整合的挑战。构建数字乡村领域的知识图谱成为解决这一问题的有效途径，能够整合各类信息，深化对数字乡村发展的理解，提升智能决策的水平。【方法】研究采用基于Bert的CasRel实体关系联合抽取模型，通过分析和挖掘相关文本数据构建数字乡村建设领域语料库，提取数字乡村领域内的重要实体和对应的关系，并且将抽取得到的数据采用三元组的形式存储于Neo4j图数据库。【结果】利用图数据库的可视化功能，可以直观地呈现数字乡村知识图谱的结构和内在关联。【结论】这不仅有助于深入理解数字乡村各要素之间的相互关系，而且为政策制定、决策支持等方面提供了可视化的数据支持。

关键词：数字乡村；语料库；知识抽取；知识图谱；Neo4j

1　引言

数字乡村是建设数字中国的重要组成部分，也是乡村振兴的战略方向。《数字乡村建设指南》涵盖信息基础设施、数字应用场景和建设运营管理等多个方面，为各地推进数字农村建设提供了参考。知识图谱可以整合数字乡村建设指南中的各种信息，包括政策条文、技术要点、实施方案等，形成一个全面的、结构化的知识网络。这有助于决策者和从业者更全面地理解指南内容，准确掌握各个要素之间的关联。本研究提出借用知识图谱这种基于数据和科学方法的定量分析工具，通过构建知识图谱全面揭示数字乡村建设发展中存在的问题，有助于优化和提升数字乡村建设战略的实施效果，为理解该领域的广泛应用提供了新的视角和方法论支持。

在当今大数据时代的背景下，各领域的专业知识和数据规模不断扩大、质量不断提高。知识图谱已经在金融[1]、医疗[2]和电力[3]等多个领域得到了广泛应用，而在农业农村领域的研究处于初步探索阶段，庞天娇[4]利用知识图谱挖掘农村污水治理领域现有知识和经验，提高了农村污水治理水平以及路径制定的科学性和可解释性。吴雅娟[5]基于构建美丽乡村数据集和知识图谱，实现了美丽乡村知识图谱智能问答。陈晓玲[6]绘制了我国数字经济赋能乡村振兴研究知识图谱，研究分析发现数字经济赋能乡村振兴领域的相关研究具有明显的政策导向性，有了国家政策的扶持，数字乡村建设快速推进。张晨阳等人[7]构建村镇公共服务设施知识图谱，从整体关联的角度评价村镇公共服务设施体系，挖掘设施间存在的空间和功能关系，将零散的设施资源关联成更有效的服务网络。研究表明，知识图谱在分析数字乡村研究趋势和现状方面已被应用，但仍缺乏能全面描述数字乡村建设的领域数据集和知识图谱。虽然利用标注方法构建语料库是可行的，但不同领域的语料库构建方法存在差异，这些方法很难移植到数字乡村建设领域的知识图谱构建中。因此，可以通过梳理数字乡村建设的结构，提取样本数据中的实体和关系，进而构建知识图谱。利用知识图谱实现数字乡村建设治理的可视化分析，不仅能够直观展现建设情况，而且能为治理部门提供便利支持，从而加快数字乡村建设的推进进程。综上所述，构建数字乡村建设领域的语料库以支持知识图谱的构建和研究具有重要的实际意义。

针对当前知识图谱的研究和构建需求，本文提出了数字乡村建设知识图谱构建方法。首先，设计数字乡村建设领域的知识本体；其次，收集领域内的新闻和学术论文等非结构化文本数据补充现有文本材料；最后，在构建面向数字乡村建设领域的知识图谱后，深入挖掘相关实体及其关系，为数字乡村的发展提供全面而系统的知识支持。通过知识图谱分析，可以明晰数字乡村要素之间的关系，识别潜在的依赖关

基金项目：2023年西安财经大学研究生创新基金项目“数字经济赋能农业经济发展的测度研究——基于知识图谱和经济计量模型的分析”（23YCZC17）。

作者简介：孔瑞琪，硕士研究生，研究方向：统计信息智能处理。Email：kongruiqi624@163.com

通讯作者：赵蕾，副教授，研究方向：统计信息智能处理。Email：leizhao@xaufe.edu.cn

系，从而更好理解指南中的指导逻辑，提升对数字化农村发展的认知水平，并为决策者提供智能化的决策依据。

2　语料库和知识图谱构建框架

2.1　语料库构建框架

语料库建设[8]包括以下步骤：（1）数据收集：收集数字乡村建设领域的相关文本，包括学术论文、政府报告、行业白皮书、新闻报道等，这些将作为语料库的主要来源；（2）数据处理：包括文本清洗、文本分析、预处理和实体关系标注任务。语料库构建方法总体框架如图1所示。

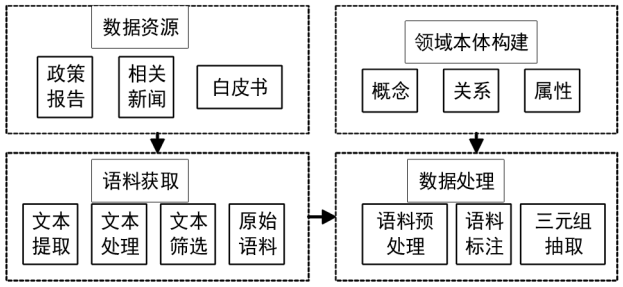

图1 语料库构建方法框架

2.2　知识图谱构建框架

通过综合自顶向下的全局框架设计和自底向上的数据挖掘优势，可以构建出既具完整性又具有动态扩展性和高质量的知识图谱[9]。采用流水线式结构的知识图谱构建技术框架，包括5个阶段：领域本体设计、数据收集和预处理、知识抽取、知识融合以及知识存储，具体技术流程见图2。

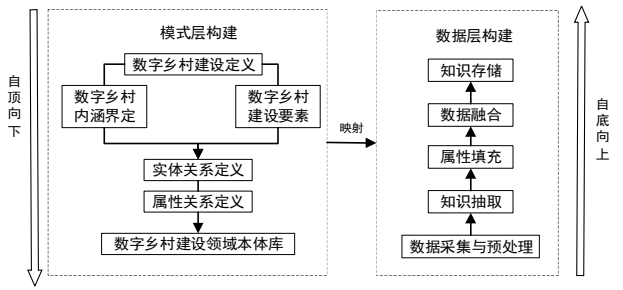

图2 知识图谱构建框架

3　数字乡村建设领域知识图谱构建方法

3.1　数据获取及处理

本研究以公开发布的《数字乡村建设指南》为主，提取相关的实体和关系，为了使数据类型更加多样化，还包括其余涉及此类信息的文本，例如新闻报道。在获取原始数据后对数据进行预处理操作，以确保文本数据质量、提高模型训练效率。在研究数字乡村建设文本内容的特点后，确定实体和关系类型作为构建数据集实体关系分类体系及标注规范，具体如表1所示。最终得到的数据集的文本数有216组，每条数据集包含多个三元组。

表1 数据集实体及实体类型

实体1	关系	实体2	实体1	关系	实体2
数字乡村建设服务	途径	建设途径	建设途径	数字化	应用场景
	采取	措施		包括	建设内容
应用场景	主体	组织名	组织名	采取	措施
	采取	措施		提供	服务
	提供	服务		对接	组织名
	数字化	建设内容			
建设项目	主体	组织名	建设内容	项目	建设项目
	采取	措施		主体	组织名
	提供	服务		采取	措施
	数字化	应用场景		提供	服务

研究采用中文多元组联合标注平台百度EasyDL。它是一款不需要进行配置的开源标注工具，导入处理好的txt文件后便可进行实体关系标注，并可以根据数据集的要求修改实体和关系类别。在实体关系标注后将标注的数据转换成三元组形式，以JSON文件进行保存用于后续的实验开发实体关系抽取。

3.2　CasRel 实体关系联合抽取模型

数字乡村建设领域的政策文本内容通常是非结构化数据的形式。现有两种基于深度学习的实体关系抽取方法：流水线法和联合法[10]。借助深度学习模型可以提取出结构化的三元组信息以构建相关的知识图谱。流水线方法面临着实体识别引起的误差积累问题，不能解决三元组重叠的问题。而研究面向的政策文本中包含了很多重叠的三元组，因此若想实现对实体和关系的同步提取，采用实体—关系的联合提取方法可以有效解决三元组重叠的问题[11]。CasRel 模型（如图3）由两个部分组成：编码层和级联解码器。编码层主要是将输入的文本内容转换为特征向量形式来得到文本的信息特征；解码端的主要内容是使用主体标注器识别所有可能的实体，再将客体标注器用于标识与关系对应的客体并提取三元组信息[12]。

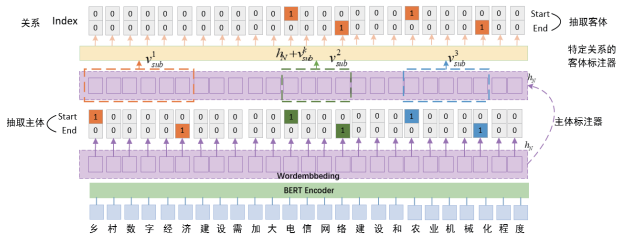

图3 CasRel 模型结构图

若输入的文本中有多个主体，则使用就近匹配的方式完成主体匹配任务。如图3所示，距离“乡”最近的标记为“1”的字是“济”，二者作为主体的起始和终止位置，能够识别出完整的实体“乡村数字经济”，同理也识别出其他主体“电信网络”和“农业机械化”，当主体“电信网络”的起始位置“电”确定后，那么在“电”之前的单词将被屏蔽，避免无法形成词汇的主体。特定关系的客体标注器与主体标注器的结构类似，也是通过标记实体的起始和终止位置来确定完整的实体。对于主体“乡村数字经济”进行关系遍历，对于“建设”这个关系能够找到特定的客体“电信网络”和“农业机械化”。通过上述步骤可以完成三元组的识别任务，然后将抽取得到数字乡村建设领域的三元组信息存入Neo4j中，构建出数字乡村建设领域知识图谱[13]。

3.3 基于Neo4j的知识图谱存储

Neo4j采用图数据库模型能够直观、高效地表示和查询复杂的关系网络，并且其高可扩展性和灵活性使其适用于处理大规模数据。知识图谱本质上是一个由实体和关系构成的网络，这种模型非常适合知识图谱的存储和管理。在数字乡村建设中，可以实时监控和调整资源分配和项目进度。因此，采用 Neo4j 图数据库作为可视化的工具。知识图谱可视化的步骤包括：将三元组信息转换为CSV文件格式；将实体节点和关系导入数据库以进行可视化。在完成知识图的可视化后，Neo4j提供的Cypher查询语言非常强大且易于使用，允许用户高效地查询、更新和操作图数据。Cypher的语法设计与SQL类似，但更适合图数据的操作，使得复杂关系的查询变得简洁明了[14]。

4 数字乡村建设知识图谱应用

4.1 数字乡村建设领域知识结构

以“数字乡村建设”为关键字查询，得出涵盖的具体内容，如图4所示。在网络基础设施方面，应加强电信网络和广播电视网络建设项目，以提供网络连接的基本能力并提升网络接入能力；公共支撑平台的建设包括公共数据平台和各种应用平台；数字应用场景描述了例如乡村网络文化和数字乡村经济等将信息化深度融合在农村生活各领域的数字化场景。各地区可根据自身发展实际，选择适合的应用场景进行参考。

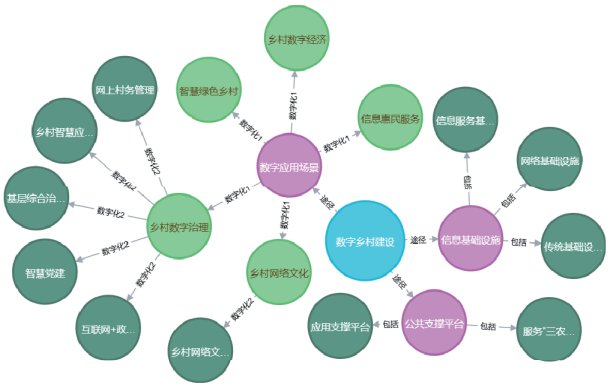

图4 数字乡村建设领域知识结构

4.2 乡村网络文化建设关键实体

若参考网络文化建设相关，以“乡村网络文化”为关键字查询，如图5所示可获得的信息有：“乡村网络文化”是数字化应用场景建设中的一方面，其又包含了“农村网络文化阵地建设”等数字化应用。以“农村网络文化阵地建设”为例，可获得的信息有：农村网络文化阵地建设在主流思想网上传播方面省级层面和县级层面负责对应的管理内容，省级层面负责宣传贯彻党的文化宣传方针政策，指导县级宣传部门开设运营各类自媒体官方宣传账号，对党和国家政策、社会主义主流思想进行常态化宣传、解读。

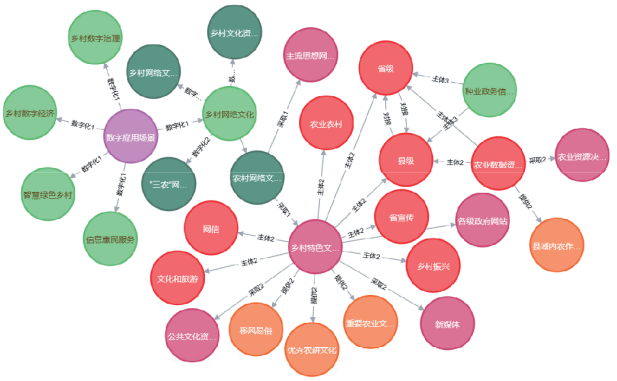

图5 乡村网络文化建设关键实体

4.3 信息惠民服务建设发展趋势

在查询“互联网+医疗健康”关键字时，可获得其相关的实体，同时可以获得与其他实体相关联的内

容如图6，例如“智慧养老”和“互联网+医疗健康”都同时提供远程医疗服务。远程医疗不仅能优化医疗资源的配置，而且有利于医院间的信息交流，包括远程传输和共享病历及医疗图像。该建设便利了医务人员的科学研究和信息检索，提高了工作效率和水平。最后，在突发公共事件中，远程医疗可以快速集中各种优质医疗资源，最大限度地保护人民的生命安全。在疫情时代，从各省级卫健委官网得知，中国各省的卫健委多次利用远程平台邀请相关医学专家开展专题会议和培训，针对疫情防控做出重要指示，为疫情得到有效防控提供了有效保障。由此可见，挖掘信息时发现的重复内容可以作为首要建设方向。通过有效整合社会资源、政府资源和信息资源等各类养老服务资源，能够实现养老服务信息的共建共享。

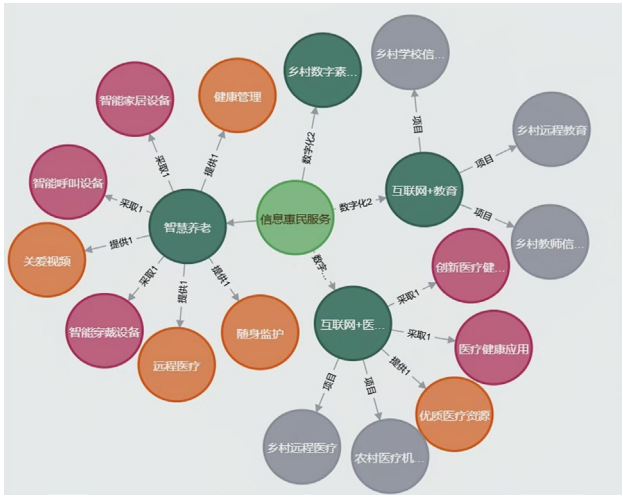

图6 信息惠民服务建设

5　结论

《数字乡村建设指南》明确了各类应用场景的建设内容、主要任务和注意事项，并提供了省、县两级的指导性建议。同时，总结提炼了地方政府的有益经验和做法，整合了整体参考架构和应用场景，以典型案例呼应建设模式，为数字乡村建设提供了系统的解释和详细的指导。由于数字农村领域对知识表示的研究相对有限，且强调知识图谱结构之间的相关性和关系特征，利用知识图谱可以有效呈现整体知识及其相互关系，为数字乡村建设提供新的路径。研究围绕数字乡村建设知识图谱的构建，使用基于BERT的CasRel实体关系联合提取模型提取数字农村建设领域知识图谱构建中的三元组数据集。该方法利用BERT模型对上下文信息的建模能力，在一定程度上解决了实体识别问题。

参考文献

[1] 祝由，贾冉，王纲金，等.供应链金融风险评估研究综述——基于知识图谱技术[J].系统工程理论与实践，2023，43（3）:795-812.

[2] 王润周，张新生.基于混合动态掩码与多策略融合的医疗知识图谱问答[J].计算机科学与探索，2024，18（10）:2770-2786.

[3] 陈宏山，周鹏，高红亮，等.基于时间知识图谱嵌入的电力恐怖主义事件预测[J/OL].哈尔滨理工大学学报，1-10.

[4] 庞天娇.基于知识挖掘的农村污水治理路径优化研究[D].重庆:重庆大学，2022.

[5] 吴雅娟.基于美丽乡村知识图谱的智能问答研究[D].南京:南京邮电大学，2023.

[6] 陈晓玲.数字经济赋能乡村振兴的研究热点与趋势——基于CiteSpace知识图谱分析[J].山西农经，2023（8）:12-15.

[7] 张晨阳，史北祥.基于知识图谱技术的村镇公共服务设施网络研究[J].西部人居环境学刊，2022，37（4）:26-32.

[8] 李秋荣，刘晓晓，王波等.滑坡地质灾害语料库构建与命名实体识别[J/OL].南京信息工程大学学报，1-17.

[9] 侯琛，牛培宇.农业知识图谱技术研究现状与展望[J].农业机械学报，2024，55（6）:1-17.

[10] 张西硕，柳林，王海龙，等.知识图谱中实体关系抽取方法研究[J].计算机科学与探索，2024，18（3）:574-596.

[11] 张仰森，刘帅康，刘洋，等.基于深度学习的实体关系联合抽取研究综述[J].电子学报，2023，51（4）:1093-1116.

[12] 蒋萌，杨春成，尚海滨，等.地理实体与重叠空间关系联合抽取的改进CasRel模型法[J].测绘学报，2023，52（8）:1387-1397.

[13] 成全，蒋世辉，李卓卓.基于改进CasRel实体关系抽取模型的在线健康信息语义发现研究[J].数据分析与知识发现，2024，8（10）:112-124.

[14] 刘彦超，刘键，席上琳，等.基于Neo4j的中轴线艺术价值数字化知识图谱研究[J].包装工程，2024，45（8）:211-223.