Interdisciplinary Science Letters 交叉科学快报, 2024, 8(1), 37-45
Published Online March 2024 in Hans. https://www.hanspub.org/journal/isl
https://doi.org/10.12677/isl.2024.81006

Hans汉斯

知识图谱内容关系重构与挖掘研究

杜亚勋[1]，张正腾[2]，常云水[1*]

[1]吉林化工学院信息与控制工程学院，吉林 吉林
[2]北华大学电气与信息工程学院，吉林 吉林

收稿日期：2024年2月25日；录用日期：2024年3月19日；发布日期：2024年3月27日

摘 要

知识图谱的挖掘可以帮助发现新的知识，通过挖掘数据之间的隐含关系和规律，揭示事物的本质和内在联系。本文构建了植物与土壤关系的知识图谱，通过知识关系将图谱结构进行动态演化，然后对演化后网络的度、介数、接近度分布特征进行分析，并根据网络中三阶和四阶结构体分布特征识别出了高粱、梨、莴苣等关键的知识实体，可以反映出这些植物实体在植物与土壤关系方面知识应用的重要性和影响力，为基于知识图谱的智能问答设计提供参考。

关键词

挖掘，土壤农作物，知识图谱

Content Relation Reconstruction and Mining for Knowledge Graphs

Yaxun Du[1], Zhengteng Zhang[2], Yunshui Chang[1*]

[1]College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin Jilin
[2]College of Electrical and Information Engineering, Beihua University, Jilin Jilin

Received: Feb. 25th, 2024; accepted: Mar. 19th, 2024; published: Mar. 27th, 2024

Abstract

Knowledge graph mining can help discover new knowledge, by mining the hidden relationships and patterns among data, and revealing the essence and intrinsic connections of things. This paper constructs a knowledge graph of plant-soil relationships, and dynamically evolves the graph structure through knowledge relations. Then, it analyzes the distribution characteristics of degree,

*通讯作者。

文章引用: 杜亚勋, 张正腾, 常云水. 知识图谱内容关系重构与挖掘研究[J]. 交叉科学快报, 2024, 8(1): 37-45.
DOI: 10.12677/isl.2024.81006

betweenness, and closeness of the evolved network, and identifies the key knowledge entities such as sorghum, pear, and lettuce based on the distribution characteristics of three-order and four-order structures in the network. These plant entities can reflect their importance and influence in the knowledge application of plant-soil relationships, and provide reference for the intelligent question answering design based on the knowledge graph.

Keywords

Mining, Soil Crops, Knowledge Graph

1. 引言

信息抽取自动化技术的发展促进了知识图谱规模的增大，面向知识图谱的挖掘是从知识图谱中抽取出隐含的知识，应用于知识图谱补全、去噪、数据解释等问题，具有准确度高、可解释性强的优势[1]。面向知识图谱的规则挖掘的研究方法主要包括基于图结构和统计规则挖掘的推理、基于知识图谱表示学习的推理、基于神经网络的推理、混合推理。基于图结构和统计规则挖掘的推理利用知识图谱的拓扑结构和频繁模式来发现规则，并通过支持度、置信度等指标来评估规则的质量。基于知识图谱表示学习的推理将知识图谱中的实体和关系映射到连续的向量空间，并通过向量运算来完成推理。基于神经网络的推理利用神经网络来学习规则或路径的表示，并通过注意力机制或强化学习等技术来优化推理过程。混合推理，结合了表示学习和规则挖掘的优点，通过迭代的方式来交互地学习规则和表示，并利用规则来指导表示学习或利用表示学习来指导规则挖掘。在应用方面，论文[2]按照动态图谱的框架结构，抽取关键电网数据，完成基于知识图谱的电网数据关联性预测。根据知识图谱结构，完成智能电网多维数据关联挖掘方法的设计与应用。论文[3]从数据底座构建、核心知识图谱挖掘、兼容传统产业链知识 3 个方面，阐述了基于图论的产业网络知识图谱的构建过程。论文[4]针对领域知识图谱的特点，从海洋中药知识图谱中检索补肾类海洋中药及相关效方，利用关联规则、聚类对海洋中药功效、主治、效方配伍、性味归经等之间的关联关系进行分析与聚类，对海洋中药知识图谱进行知识补全。论文[5]提出了一种基于监控视频的知识图谱构建和数据挖掘方法。基于知识图谱利用行人共现算法、轨迹挖掘算法、社团检测算法等对监控视频进行数据挖掘。

随着全球人口的增长和粮食需求的不断增加，农业生产面临着巨大的挑战。土壤作为农业生产的基础，其质量和特性对农作物的生长和产量具有重要影响。了解土壤与农作物之间复杂的关联关系对于优化种植方案、提高农产品质量和数量至关重要[6]。然而，传统的农业生产方式往往未能充分考虑土壤特性对农作物的影响，导致资源浪费和环境压力加剧。在探索土壤特性与农作物关系的研究领域中，相关工作已经取得了一定进展。过去的研究主要集中在土壤类型与农作物适应性之间的关联、施肥方案对作物产量的影响以及土壤改良对农作物生长的作用等方面。首先，早期的研究侧重于探讨不同土壤类型对农作物生长的影响。这些研究通过对比不同土壤特性(如质地、PH 值、养分含量等) [7] [8] [9]下作物的生长状况，探究了土壤类型对作物适应性的影响。然而，这些研究往往局限于特定的土壤类型和少数作物，未能形成全面的土壤－作物关系模型。另外，一些研究致力于探讨土壤改良对农作物生长的作用[10] [11]

[12]。这些研究通过添加改良剂、优化耕作措施等方法，试图改善土壤结构和养分供应，从而提升作物产量和质量。然而，这些方法往往忽略了土壤特性与作物生长之间更为细致的关联，无法提供全面的土壤－作物关系信息。虽然过去的研究工作为理解土壤与植物之间的关系提供了宝贵的参考，但仍存在着一些不足之处。缺乏综合考虑土壤多种特性与作物生长之间的关联、未能建立全面的土壤－作物关系[13] [14]模型等问题限制了这些研究的深入和应用。

本文利用已发表论文中的数据，包括土壤类型、作物种类等信息[15] [16]，通过数据整合和知识提取，构建一个植物与土壤图谱模型。图谱以节点和关系的方式呈现土壤特性、不同作物以及它们之间的关联，从而更好地揭示土壤与植物之间的复杂联系。然后，依据植物实体土壤特征相似度关系对图谱进行演化，对演化后网络的度、介数、接近度分布特征进行分析，并根据网络中三阶和四阶结构体分布特征识别出了高粱、梨、莴苣等关键的知识实体。可以降低语义复杂度，提高自然语言的查询效率和准确度。

2. 构建知识图谱

行业知识图谱与各行各业相结合，对不同用户提供精准推送[17]的服务，并在构建完成后并入通用知识图谱，为将来的深入研究提供数据支撑。2012 年，谷歌公司为了优化用户的搜索体验，首次提出知识图谱的概念[18]，将现实世界中各种类型的数据转化为“资源－属性－属性值”三元组结构。

2.1. 数据获取

构建高质量的知识图谱需要高质量的数据源。因此本文结合土壤作物知识图谱的特点，对其领域的数据进行筛选，最终选择表 1 所示的土壤文献作为本文的数据来源。

Table 1. Sources of data
表 1. 数据来源

序号	题目	作者	作者单位
1	微塑料对土壤植物生长发育影响研究进展	邓悦	河南大学
2	植物－土壤反馈对植物生长发育及土壤氮素利用和损失的影响	孙修婷	东北师范大学
3	土壤水分和养分对沙质草地优势植物叶片氮回收效率的影响	张晶	中国科学院西北生态环境资源研究院
4	植物－土壤反馈对西南亚高山森林 4 种草本植物种子萌发和幼苗生长的影响	刘露	中国科学院成都生物研究所
5	不同土壤水分对植物光合作用的影响研究进展	杨佳鹤	宁夏大学
6	土壤类型对超积累植物东南景天叶际微生物群落结构和功能的影响	姜悦	浙江大学

2.2. 知识抽取

使用基于关键词匹配的方法文献中查找特定的关键词，以从中提取所需信息。

将已经下载的文献编入文本文件。并使用 python 进行关键词匹配，从“属于”，“吸收”，“影响”这三个关系中提取“实体－属性－值”。遍历获取的属性结果，存储为 excel 表格式，部分数据如表 2 所示。

Table 2. Knowledge extraction relationship table
表 2. 知识抽取关系表

植物	关系	种类	存在	根系	属性	土壤
小麦	属于	禾本科	有	小麦根系	吸收	土壤氮磷钾比重是 1 比 0.4 比 0.6
水稻	属于	禾本科	有	水稻根系	吸收	土壤氮磷钾比重是 1 比 0.3 比 0.7
高粱	属于	禾本科	有	高粱根系	吸收	土壤氮磷钾比重是 1 比 0.6 比 0.9
燕麦	属于	禾本科	有	燕麦根系	吸收	土壤氮磷钾比重是 1 比 0.5 比 0.8
大麦	属于	禾本科	有	大麦根系	吸收	土壤氮磷钾比重是 1 比 0.4 比 0.7
竹子	属于	禾本科	有	竹子根系	吸收	土壤氮磷钾比重是 1 比 0.7 比 1.2
甘蔗	属于	禾本科	有	甘蔗根系	吸收	土壤氮磷钾比重是 1 比 0.6 比 1.1
牧草	属于	禾本科	有	牧草根系	吸收	土壤氮磷钾比重是 1 比 0.5 比 0.9
狗尾草	属于	禾本科	有	狗尾草根系	吸收	土壤氮磷钾比重是 1 比 0.4 比 0.6
稗草	属于	禾本科	有	稗草根系	吸收	土壤氮磷钾比重是 1 比 0.3 比 0.5
玫瑰	属于	蔷薇科	有	浅根系	吸收	土壤氮磷钾比重是 1 比 0.5 比 0.9

依据知识抽取关系表，共确定知识图谱的 4 种实体类型和 3 种关系类型。如表 3 所示。

Table 3. Entity relationship table
表 3. 实体关系表

实体	关系	实体
植物	属于	种类
根系	吸收	土壤
植物	存在	根系

2.3. 图数据库存储

本文采用 Neo4j 图数据库存储数据。首先逐行读取知识抽取关系表的内容，该表的第一列为植物名称，后面列为该植物的属性，逐一抽取创建实体和关系，直到文件结束，植物与土壤知识图谱创建完成。图 1 是部分使用 py2neo 模块在 pycharm 上的展示，植物－属于－种类的相关信息。

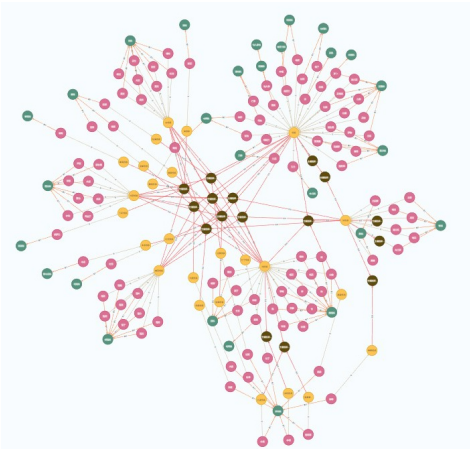

Figure 1. Knowledge map of plants and soils
图 1. 植物与土壤知识图谱

3. 知识图谱挖掘

3.1. 植物关系复杂网络模型构建

依据知识抽取关系表中土壤的氮磷钾比例结算植物的关系紧密度 D_{ij}。

$$D_{ij}=\frac{\sum_{p}^{k}\left(\left|u_i+u_j\right|-\left|u_i-u_j\right|\right)}{\sum_{p}^{k}\left|u_i+u_j\right|}$$

其中，u_i 和 u_j 分别是植物 i 和植物 j 的磷(p)和钾(k)比例。获得所有植物间的关系紧密度以后，构建植物关系复杂网络模型，网络中的节点是知识图谱中的植物实体，网络中的边是植物间的关系度，当植物关系度的值大于等于 0.95 时，则认为两个植物间存在一条边，否则认为两个植物间不存在边。其复杂网络模型如图 2 所示，节点数量为 104，边的数量为 1390，网络边比率为 0.2595，聚类系数为 0.7929。

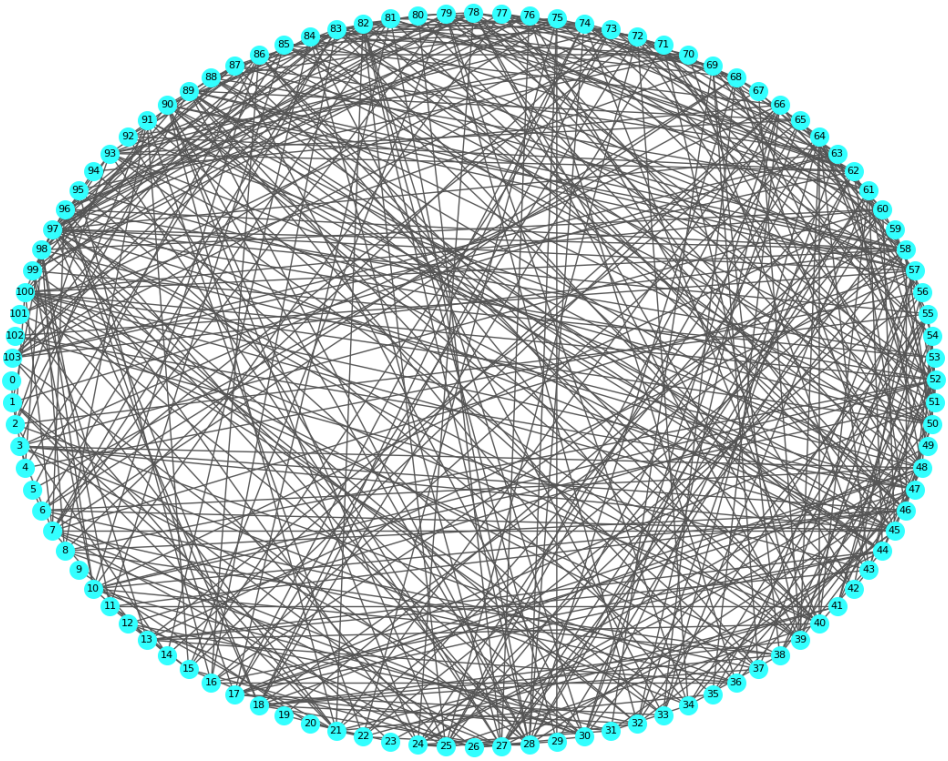

Figure 2. Plant relationship network model
图 2. 植物关系网络模型

3.2. 基本属性分布特征分析

3.2.1. 度分布

度分布是指网络中节点的度的概率分布或频率分布。度分布反映了网络的结构和性质，也可以用来刻画网络的异质性或同质性，即节点之间的连接是否均匀或不均匀。植物关系网络模型中植物度分布如图 3 所示，可以看出，其并不符合幂律分布，该网络不是无标度网络。度为 12 的节点比例最高，度值较低的节点所占比例很低，说明大多数植物特性较为相近。

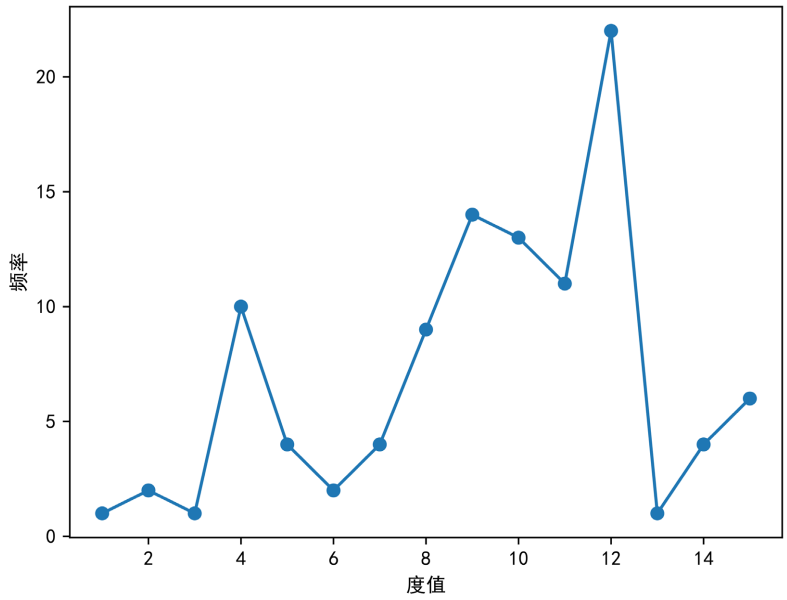

Figure 3. Degree distribution
图 3. 度分布

3.2.2. 介数分布

介分布是指网络中节点的介数的概率分布或频率分布。介数是一种反映节点在网络中的重要性和影响力的指标，它表示网络中所有最短路径中经过该节点或边的路径的数目占总的最短路径数的比例。植物关系网络模型中植物度分布如图 4 所示，介数值为 4 的节点所占比例最高，分布特征类似于正态分布。

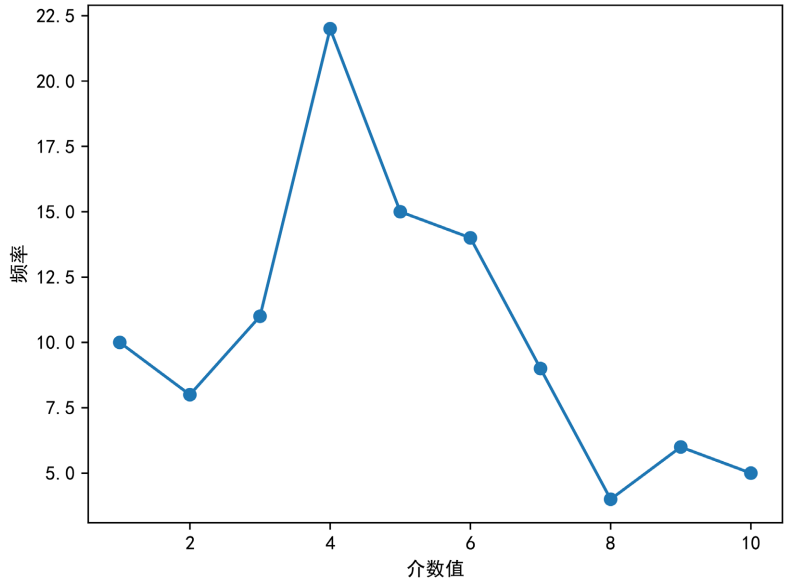

Figure 4. Intermediary distribution
图 4. 介数分布

3.3.3. 接近度分布

接近度分布是指网络中节点或边的接近度的概率分布或频率分布。接近度是一种反映节点或边在网

络中的位置和重要性的指标，它表示节点到其他节点或边的平均距离的倒数。接近度越大，说明节点越靠近网络的中心，越能快速地与其他节点交流信息。植物关系网络模型中植物度分布如图 5 所示，接近度为 16 的节点所占比率最高。

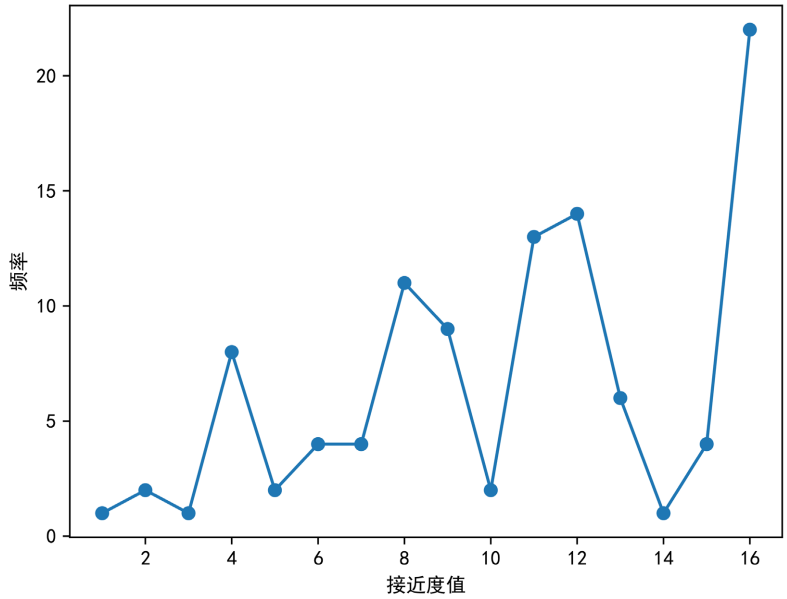

Figure 5. Proximity distribution
图 5. 接近度分布

3.3. 基于三阶和四阶结构体分布的重要植物识别

三阶和四阶结构体分布是指网络中三角形和四面体的单纯形的概率分布或频率分布。多阶结构体分布可以反映网络的高阶拓扑特征，例如网络的示性数、贝蒂数、圈数、洞数等。多阶结构体分布也可以用来分析网络的动力学性质，例如网络的同步能力、控制能力、传播能力等。三阶和四阶结构体如图 6 所示，其中，三阶结构体有两个，四阶结构体有 6 个，其分布的计算方法取决于网络的类型和特征。本文计算了植物关系网络模型中所有节点的关联三阶结构体和四阶结构体数量，将该数量作为节点的重要度，识别出的最重要的 10 个节点的名称如表 4 所示，其中高粱、梨和莴苣是植物与土壤关系知识谱图中最重要的三个植物节点。

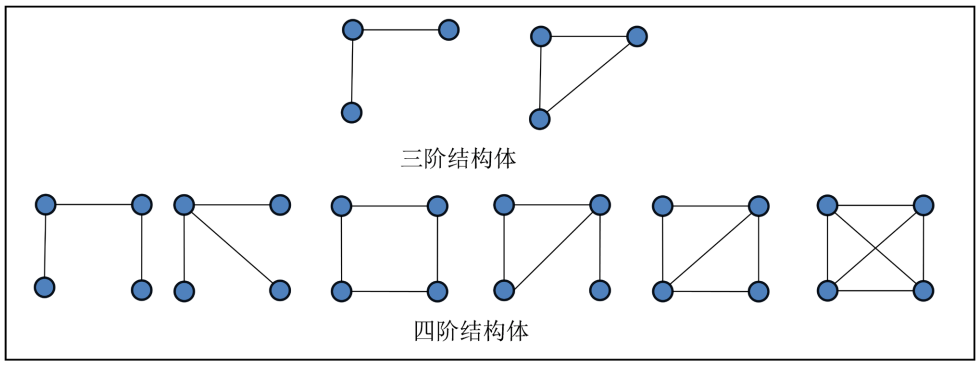

Figure 6. Third- and fourth-order structures
图 6. 三阶和四阶结构体

Table 4. Plant entity importance based on third-order and fourth-order structure distribution
表 4. 基于三阶和四阶结构体分布的植物实体重要度

植物实体重要度排序	基于三阶结构体分布的植物名称	基于四阶结构体分布的植物名称
1	高粱	高粱
2	梨	梨
3	莴苣	莴苣
4	黄豆	水稻
5	郁金香	黄豆
6	兰花	苹果
7	水稻	蒲公英
8	苹果	郁金香
9	蒲公英	马鞭草
10	马鞭草	兰花

4. 结论

本研究旨在提出一种知识图谱内容的深入挖掘思路，以植物生长与土壤特性的联系为研究对象，从多源数据中汇集了土壤类型、作物种类、种植季节、施肥方案等关键信息，利用这些数据构建了植物与土壤关系知识图谱，并基于知识图谱的内容关系重构了植物关系复杂网络模型，通过网络模型的结构特征挖掘出植物知识特性并识别了知识图谱中重要的植物知识节点。该知识图谱挖掘思路不仅可以用于土壤与植物领域，其可以用于所有的知识图谱，挖掘出重要的知识节点或知识关系。

由于自然语言处理问题可以转化为知识图谱中的实体和关系识别问题，因此知识图谱中关键知识实体的识别，可以根据实体的重要性、相关性和新颖性等指标进行排序，有助于提高知识图谱的压缩率和存储效率，通过保留关键节点和它们之间的关系，可以去除冗余或无关的信息，减少知识图谱的规模和复杂度；有助于提高知识图谱的检索效率和准确度，通过将用户的查询与关键节点进行匹配，可以快速定位到相关的知识子图，避免遍历整个知识图谱，提高查询的响应速度和质量；有助于提高知识图谱的可解释性和可视化效果，通过将关键节点作为知识图谱的核心元素，可以更好地展示知识图谱的结构和语义，帮助用户理解和探索知识图谱的内容和特征。有助于提高知识图谱的应用价值和创新潜力，通过将关键节点作为知识图谱的关注点，可以更好地发现知识图谱中的规律和趋势，支持知识的推理和发现，促进知识的创新和应用。

本文对知识图谱演化网络模型的拓扑结构进行了分析，发现其度、介数和接近度分布均不符合传统的无标度性和小世界性，结构特征规律不够明显。因此，只通过网络的节点(一阶)或边(二阶)的分布特征无法识别关键知识实体。进一步分析三阶和四阶结构体的分布特征，将每个知识实体所在三阶和四阶结构体的数量作为量化指标为知识图谱中所有知识实体排序，通过比较两种排序结果的前十个知识实体，发现通过三阶和四阶结构体分布特征识别出的最关键的十个知识实体是相同的，只是排序的顺序略有不同。说明通过有效的知识图谱演化重构后，通过高阶结构体的分布特征可以有效的识别知识图谱中的关键知识实体。

基金项目

本论文由国家级大学生创新创业项目“基于土壤检测的专家指导系统”(202310201011)资助。

参考文献

[1] 刘洪波, 等. 面向知识图谱的规则挖掘研究综述[J]. 计算机工程与应用, 2023, 59(14): 30-38.

[2] 许中平, 等. 基于知识图谱的智能电网多维数据关联挖掘方法[J]. 电子设计工程, 2023, 31(11): 84-87+92.

[3] 李振军, 等. 基于图论的产业网络知识图谱挖掘与构建[J]. 大数据, 2023, 9(6): 174-183.

[4] 洪海蓝, 等. 基于海洋中药知识图谱的数据挖掘与知识补全研究[J]. 中医药信息, 2023, 40(10): 27-34.

[5] 金磊, 等. 基于监控视频的知识图谱数据挖掘[J]. 工业控制计算机, 2022, 35(5): 76-78+81.

[6] 黄旭东, 杨永红, 曹秀文, 等. 白龙江高山林线木本植物组成与地被物和土壤持水特性[J]. 贵州林业科技, 2023, 51(2): 60-64.

[7] 李丛笑, 张彦, 覃茜瑾, 等. 黄河三角洲不同植物群落土壤有机碳特征及其影响因子[J/OL]. 环境科学, 1-13. https://doi.org/10.13227/j.hjkx.202308141, 2024-03-21.

[8] 刘炜璇, 李依蒙, 江红星, 等. 吉林莫莫格国家级自然保护区四种典型植物群落下土壤微生物组成的对比分析[J/OL]. 生态学杂志, 1-12. http://101.42.170.182:8085/kcms/detail/21.1148.Q.20230914.1106.006.html, 2024-03-21.

[9] 彭雪梅, 阙涛, 胡鑫, 等. 喀斯特地区公路边坡植物多样性与土壤养分关系分析[J]. 天津农林科技, 2023(3): 1-5.

[10] 张志明, 孙小妹, 包段红, 等. 祁连山北麓荒漠草原 5 种优势植物生物量与土壤养分特征[J/OL]. 干旱区地理, 1-15. http://101.42.170.182:8085/kcms/detail/65.1103.x.20231130.1457.001.html, 2024-03-21.

[11] 冯爱平, 康鹏宇, 刘传朋, 等. 山东沂南土壤-植物系统中硒生物有效性评价[J]. 吉林大学学报(地球科学版), 2023, 53(4): 1216-1227.

[12] 魏瑶瑶, 郑恩, 高铭雨, 等. 陕北山地苹果园土壤饱和导水率和植物导水率特征[J]. 西北农业学报, 2023, 32(11): 1813-1820.

[13] 孙修婷. 植物-土壤反馈对植物生长发育及土壤氮素利用和损失的影响[D]: [硕士学位论文]. 长春: 东北师范大学, 2023. https://doi.org/10.27011/d.cnki.gdbsu.2023.000691

[14] 刘露, 赵文强, 梁婷, 等. 植物-土壤反馈对西南亚高山森林 4 种草本植物种子萌发和幼苗生长的影响[J/OL]. 生态学杂志, 1-12. http://101.42.170.182:8085/kcms/detail/21.1148.Q.20231103.1335.006.html, 2024-03-21.

[15] 杨佳鹤, 何进宇, 刘飞杨, 等. 不同土壤水分对植物光合作用的影响研究进展[J]. 节水灌溉, 2023(11): 39-46.

[16] 刘明蕊, 刘世婷, 马春燕, 等. 草地植物和土壤对温度和降水变化的响应研究进展[J/OL]. 生态学杂志, 1-12. http://101.42.170.182:8085/kcms/detail/21.1148.Q.20231026.1119.002.html, 2024-03-21.

[17] 史入文. 融合知识图谱和情感分析的推荐算法研究[D]: [硕士学位论文]. 上海: 上海社会科学院, 2020. https://doi.org/10.27310/d.cnki.gshsy.2020.000107

[18] Fensel, D., Şimşek, U., Angele, K., *et al.* (2020) Introduction: What Is A Knowledge Graph? In: *Knowledge Graphs*, Springer, Cham, 1-10. https://doi.org/10.1007/978-3-030-37439-6_1