植物遗传资源学报
Journal of Plant Genetic Resources
ISSN 1672-1810,CN 11-4996/S

《植物遗传资源学报》网络首发论文

题目：基于 GraphRAG 的中国马铃薯新品种知识图谱构建
作者：韦一金，任有强，赵慧，樊景超，方沩，闫燊
DOI：10.13430/j.cnki.jpgr.20240919001
收稿日期：2024-09-19
网络首发日期：2024-11-13
引用格式：韦一金，任有强，赵慧，樊景超，方沩，闫燊．基于 GraphRAG 的中国马铃薯新品种知识图谱构建[J/OL]．植物遗传资源学报．
https://doi.org/10.13430/j.cnki.jpgr.20240919001

植物遗传资源学报
Journal of Plant Genetic Resources　DOI：10.13430/j.cnki.jpgr. 20240919001

基于 GraphRAG 的中国马铃薯新品种知识图谱构建

韦一金[1, 2]，任有强[2, 3]，赵　慧[1, 2]，樊景超[1, 2]，方　沩[4]，闫　桑[4]

（[1]中国农业科学院农业信息研究所，北京 100081；[2]国家农业科学数据中心，北京 100081；
[3]三亚中国农业科学院国家南繁研究院，海南 572024；[4]中国农业科学院作物科学研究所，北京 100081）

摘要：马铃薯是世界第四大主粮作物，拥有较高的产量潜力，为应对未来的粮食安全挑战，需要选育具有稳定抗病性的早熟高产马铃薯品种。本研究为助力马铃薯新品种选育，明确目前中国马铃薯选育种现状，以中国知网（CNKI）数据库中 227 种马铃薯选育文献为研究对象，利用 GraphRAG 和 Qwen2-70B-instruct 构建知识图谱并使用 Gephi 实现可视化。基于所构建的知识图谱，分析近几年中国选育的马铃薯新品种的系谱、抗性和生育期，结果表明近几年马铃薯新品种选育使用较多的亲本为冀张薯 8 号、斯凡特、费乌瑞他和早大白等，马铃薯选育种大多对晚疫病有抗性，且生育期大多为中晚熟、晚熟。综上，本研究探索了使用大语言模型快速构建马铃薯新品种选育研究知识图谱的实现路径，并对 227 个马铃薯选育种进行分析，为马铃薯种质资源未来的发掘利用提供参考。

关键词：知识图谱；马铃薯种质资源；大语言模型；农业

Construction of a Knowledge Graph for Selection and Breeding Research of New Potato Varieties in China Based on GraphRAG

WEI Yijin[1, 2]，REN Youqiang[2,3]，ZHAO Hui[1, 2]，FAN Jingchao[1, 2], FANG Wei[4], YAN Shen[4]

（[1]*Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081*；[2]*National Agriculture Science Data Centre, Beijing 100081;* [3]*National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya 572024;* [4]*Institute of Crop Science of Chinese Academy of Agricultural Sciences, Beijing 100081*）

Abstract: Potato is the world's fourth largest staple crop with high yield potential, and selection of early-maturing and high-yielding potato varieties with stable disease resistance is needed to meet future food security challenges. In this study, in order to assist the selection and breeding of new potato varieties and to clarify the current status of potato selection and breeding in China, 227 potato selection and breeding literatures in the China Knowledge Network (CNKI) database were used as the research object, and a knowledge map was constructed by using GraphRAG and Qwen2-70B-instruct and visualized by using Gephi. Based on the constructed knowledge graph, we analyzed the genealogy, resistance and fertility period of new potato varieties selected and bred in China in recent years, and the results showed that the parents used more in the selection and breeding of new potato varieties in recent years were Ji Zhang Yam 8, Svante, Feiuretta, and Early Large White, etc. Most of the potato selections were resistant to late blight, and most of them were of medium-late maturity and late ripeness in their fertility period. In summary, this study explored the realization path of using a large language model to rapidly construct a knowledge graph for potato new variety selection and breeding research, and analyzed 227 potato selections to provide a reference for the future discovery and utilization of potato germplasm resources.

Key words: knowledge graph; potato germplasm resources; large language models; agriculture

马铃薯（*Solanum tuberosum* L.）是继水稻、小麦、玉米之后的世界第四大粮食作物[1]。有限的自然资源、频发的极端天气、人口的持续增长等问题正对世界粮食安全造成威胁，育种工作者面临着巨大的挑战，需要选育兼具高产、抗逆、优质等特性的马铃薯。目前生产上主要使用的栽培马铃薯大多来自于几个具有显

收稿日期：2024-09-19
第一作者研究方向为管理系统工程，E-mail : wyj18376068969@163.com
通信作者：樊景超，研究方向为农业大数据，E-mail : fanjingchao@caas.cn
方　沩，研究方向为作物种质信息管理，E-mail: fangwei@caas.cn
闫　桑，研究方向为作物遗传育种，E-mail: yanshen@xaas.cn
基金项目：中国农业科学院科技创新工程专项（CAAS-ASTIP-2024-AII）
Foundation project：Special Founds for Scientific and Technological Innovation Project of Chinese Academy of Agricultural Sciences(CAAS-ASTIP-2024-AII)

著优良性状的亲本杂交选育而成，如卡它丁、多子白、米拉、疫不加、小叶子和白头翁等[2]，而马铃薯地方种和野生种，往往具有产量低、口感差、外观表现略差等原因，应用比例较低[3]。因此，马铃薯品种间存在遗传基础较狭窄[4]，遗传多样性低[5]等问题。

知识图谱最初提出是为了优化谷歌的搜索引擎，其在存储实体和关系上有一定的优越性[6]。知识图谱通过实体、实体属性及实体间的关系来刻画知识关联，构成了一种揭示实体之间关系的语义网络[7]。这一特性令知识图谱可以完成智能分析、智能查询和智能问答等一系列自然语言处理任务[8]。作为一种知识组织形式，知识图谱以图的形式有效整合信息，可以帮助解决信息杂散和无序所带来的问题[9]，且拥有直观性、扩展性和可塑性强等优势，广泛应用于军事、医学、经济和农业等领域，特别是在智慧农业、领域知识图谱相关研究中受到了广泛的关注与重视[10-14]。但农业领域的知识图谱构建及应用还存在大量数据的高质量批量处理困难，实体分类的精准度不够，数据标注相对单一，模型的稳定性和效率不高等问题。因此，为解决上述问题，明确现有马铃薯选育现状，为未来马铃薯育种工作提供参考，本研究基于大语言模型（LLM，Large Language Model）及相关技术，研究使用 GraphRAG 快速构建农业领域知识图谱方法，形成了 227 个马铃薯品种信息知识图谱，能够为未来的马铃薯选育工作提供参考。

1 材料与方法

本文整体技术路线图如图 1 所示，主要工作可以分为 5 大部分，第一部分针对所收集的马铃薯数据进行数据清洗和预处理，第二部分针对知识图谱构建需求选择基础模型，第三部分则是基于 GraphRAG 进行知识图谱构建工作，第四部分中对所构建的知识图谱进行可视化，包括利用 K-Core 算法构建对应的 K-Core 子图，最后第五部分对马铃薯亲本、抗性、生育期材料进行分析。

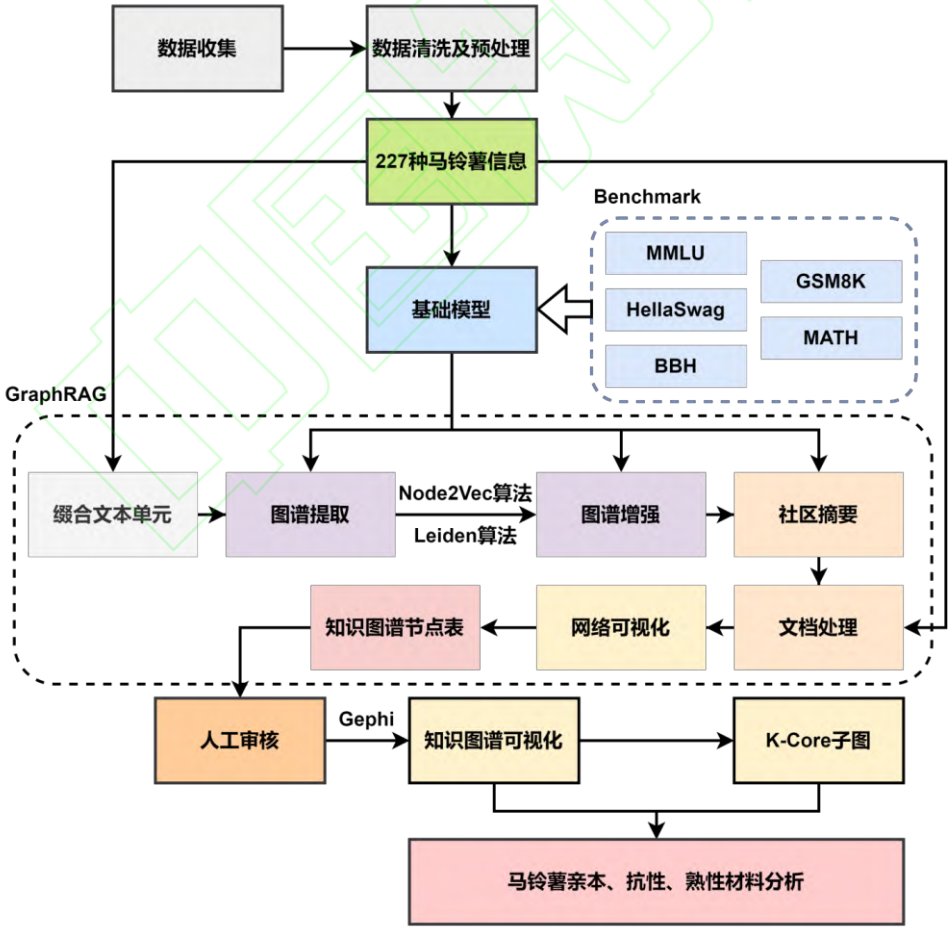

图 1 技术路线图

Fig.1 Technology Roadmap

1.1 数据来源

为确保检索文献的准确性、科学性及可信度，本文以中国最全面、规模较大的学术期刊中国知网（CNKI）作为样本数据来源，分类号选择中图分类 S532 马铃薯（土豆），以“马铃薯新品种选育”为核心检索词，检索 CNKI 收录的高质量学术论文期刊。对检索结果进行反复检查，剔除与马铃薯选育研究不相关的条目以及信息较少的马铃薯，共得到 227 个马铃薯信息。

1.2 基础模型选型

本研究在 GraphRAG 内使用的大语言模型为 Qwen2-70B-instruct[15]。本研究在构建知识图谱时侧重于使用具有一定上下文长度的、具备一定语言理解能力和逻辑推理能力的 LLM，因此在选择所使用的模型时主要关注模型在 MMLU[16]、HellaSwag[17]、BBH[18]、GSM8K[19]和 MATH[20]基准下的分数。本研究比较了目前的一部分主流 LLM 在这些 benchmark 上的分数，包括 Command-r-plus-104b、Qwen1.5-110B、Yi-34B、LLaMA3-70B 、 Deepseek-v2-236B 和 Qwen2-70B-instruct ， 分 数 比 较 情 况 如 表 1 所 示 。 其 中 Command-r-plus-104b 模型是在双盲实验中表现出了优秀的性能而成为主流 LLM 之一，目前没有可靠的在 BBH 和 MATH 基准下的公开分数。通过比较与 MATH 基准类似的用于评估 LLM 解决数学问题能力的 GSM8K 基准下的分数可知，Command-r-plus-104b 模型的数学能力与 Qwen2-70B-instruct 相比差距较大，在知识的推理上可能能力略逊色于 Qwen2-70B-instruct，因此本研究最终选择在综合分数较好 Qwen2-70B-instruct 模型用于构建知识图谱。

表 1 主流大语言模型在指定基准下的分数

Table 1 Scores of Mainstream Large Language Models for Specified Benchmarks

模型 Model	基准 Benchmark(Metric)					上下文
	MMLU	HellaSwag	BBH	GSM8K	MATH	Context
Command-r-plus-104b	75.7	**88.6**	-	70.7	-	128k
Qwen1.5-110B[21]	80.4	87.5	74.8	85.4	49.6	32k
Yi-34B[22]	76.3	87.19	54.3	67.2	14.4	200k
LLaMA3-70B[23]	79.5	88.0	76.6	79.2	41.0	8k
Deepseek-v2-236B[24]	78.5	84.2	78.9	79.2	43.6	128k
Qwen2-70B-instruct	**82.3**	87.6	**82.4**	**91.1**	**59.7**	128k

1.3 知识图谱构建

本研究使用 GraphRAG[25]进行知识图谱构建。GraphRAG 是一种基于知识图谱的检索增强技术，其中 RAG（Retrieval-Augmented Generation, RAG）是一种结合检索技术和生成技术的方法，通过增强生成过程来提高检索结果的准确性、相关性和多样性。通过构建图模型的知识表达，将实体和关系之间的联系用图的形式展示出来，然后利用大语言模型（Large Language Model, LLM）进行检索增强。这种方法将知识图谱视为一个超大规模的词汇表，实体和关系则对应于单词，从而在检索时能将实体和关系作为单元进行联合建模。

1.4 知识图谱可视化

本研究使用 Gephi 进行知识图谱可视化。Gephi 软件主要用于网络科学的方法处理与分析关系数据，通过不同方式的布局对图进行可视化处理，或通过对生长网络动态模拟进行解读与分析。Gephi 不仅可以处理大规模网络数据集，也可在节点层面对网络属性进行统计分析，使用不同布局算法对网络进行可视化处理与模拟分析。本研究在具体可视化时，设定节点大小与节点连接数存在等比例关系，即节点连接的点越多则节点的尺寸越大。

1.5 K-Core 算法

本研究在具体分析部分，使用 K-Core 算法在所构建的知识图谱基础上筛选符合指定核心度的紧密关联的子图结构用于分析。K-Core 算法是一种子图挖掘算法，用于寻找图中符合指定核心度的顶点的集合，即要求每个顶点至少与该子图中的其他 k 个顶点相关联。该算法通常用于对一个图进行子图划分，通过去除

不重要的顶点，将符合逾期的子图暴露出来进行进一步分析，有助于找到图网络中核心度较高的节点集合，从而更好地理解网络的本质结构。

1.6 实验设置

实验配置如表 2 所示。本文实验使用的操作系统为 Windows11 工作站专业版，GPU 为 NVIDIA RTX 6000，编程语言为 Python3.11。知识图谱构建使用 GraphRAG，并且在 GraphRAG 中使用的大语言模型为 Qwen2-70B-instruc，Embedding 模型为 nomic-embed-text。此外，本文实验使用 Ollama 框架实现 Qwen2-70B-instruc 的本地化部署，使用 Gephi 实现知识图谱可视化。

2 结果与分析

本研究对 GraphRAG 构建的知识图谱进行人工数据检验校正，最终获得一个包含 1135 个节点和 1755 条边的知识图谱。观察所构建的知识图谱易见现有马铃薯新品种选育研究存在明显的特征特性倾向，知识图谱整体如图 2 所示，所构建知识图谱的中部呈现多个大尺寸节点。其中节点大小及颜色与节点的连接数有关，连接数越多节点越大并且颜色越浅。

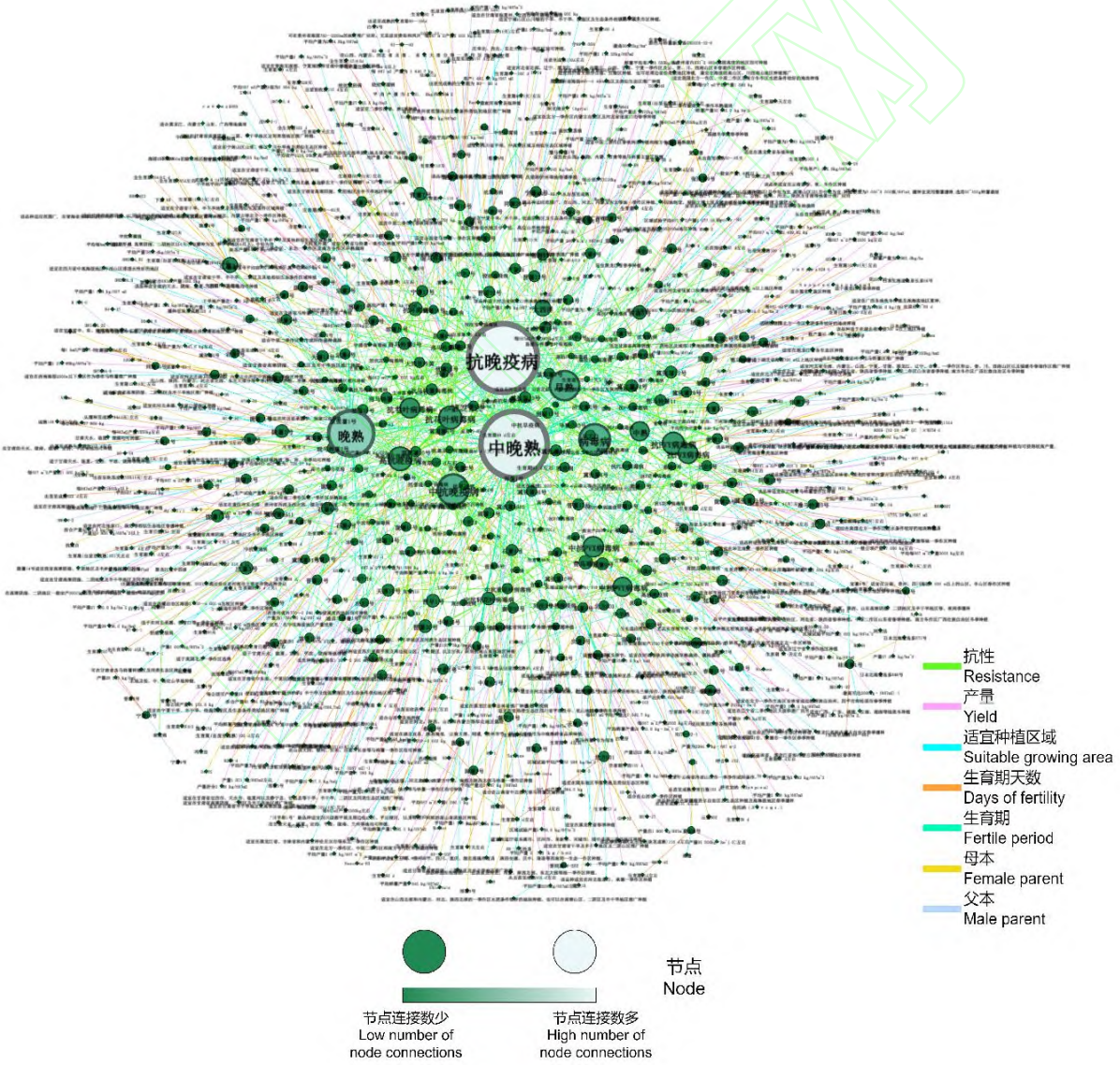

图 2 马铃薯育成种知识图谱整体可视化

Fig.2 Overall visualization of the potato breeding seed knowledge graph

使用 K-Core 算法获得 4-Core 子图筛选紧密关联的节点，如图 3 所示，易见这些节点主要与马铃薯抗性、

生育期相关，如现有马铃薯育成种大多具备晚疫病抗性，且生育期大多为中晚熟、晚熟。因此，本研究在对马铃薯育成种亲本材料进行分析的基础上，以马铃薯育成种普遍具备的抗晚疫病、中晚熟和晚熟特性为出发点，对其抗性和生育期信息单独提取分析。

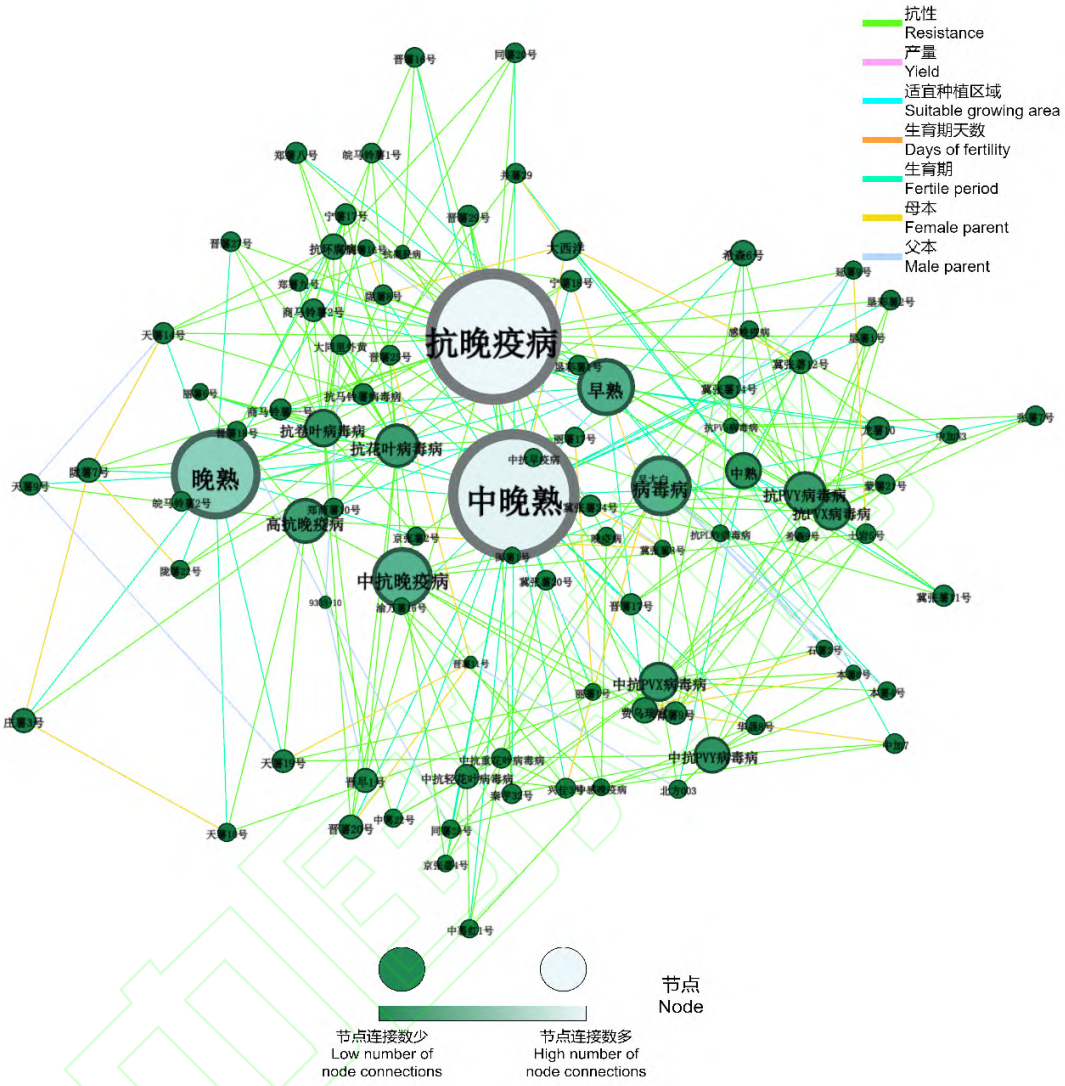

图 3 马铃薯育成种知识图谱 4-Core 子图

Fig.3 Potato Breeding Seed Knowledge Map 4-Core Submap

2.1 马铃薯品种选育中的重要亲本分析

系谱分析广泛应用于多种作物的遗传多样性研究[26]，可以评估品种和育种无性系之间的亲缘关系，从而便于品种间的杂交设计，提高关联遗传学分析的能力。世界范围内的系谱分析表明，在杂交育种中使用优质亲本种质可以培育出优秀的马铃薯品种[27]。本研究针对马铃薯育成种的亲本材料，从所构建的知识图谱中提取相关信息以分析马铃薯品种选育研究中的亲本材料使用情况，对所提取信息进行可视化结果如图 4 所示。

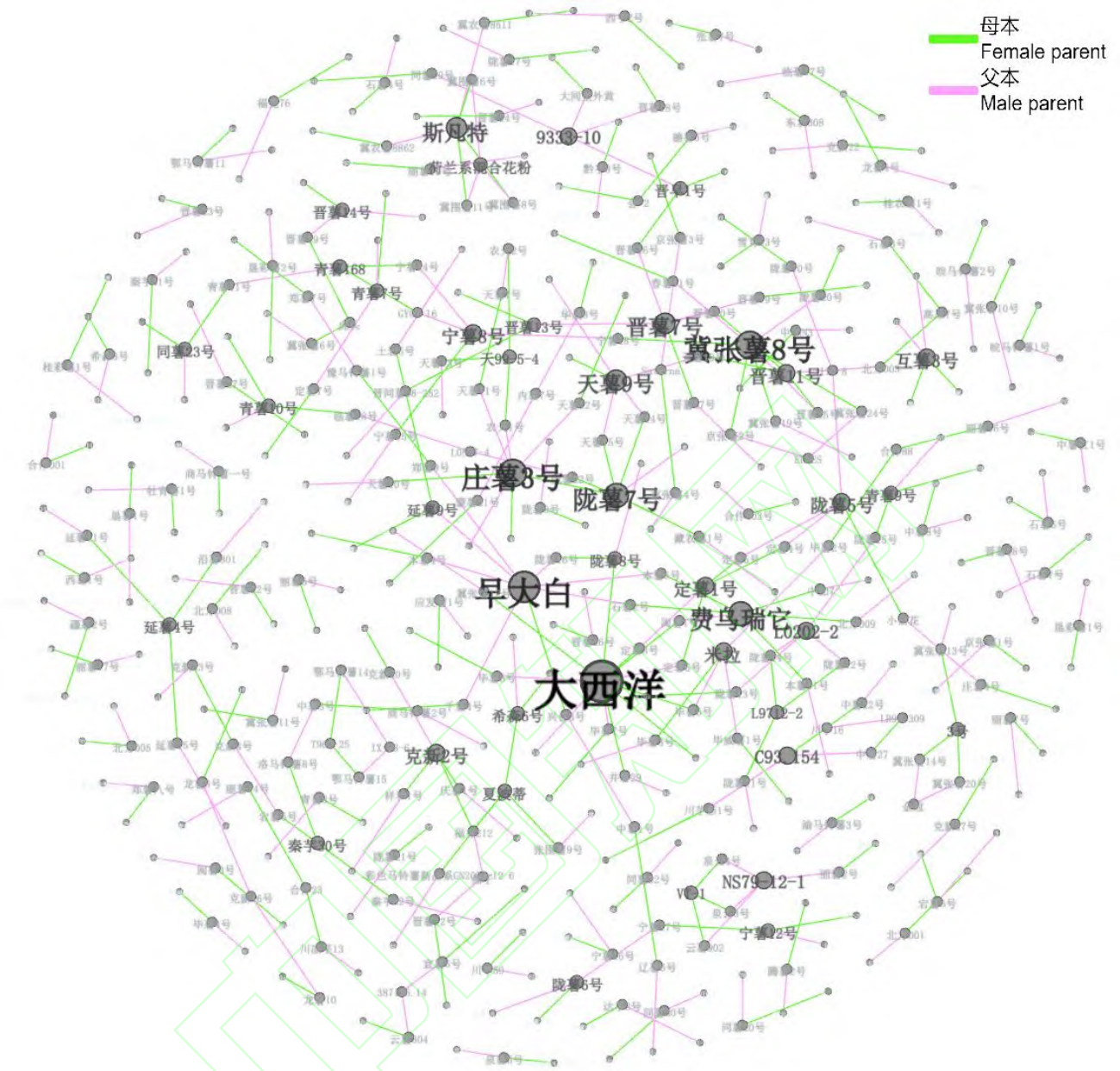

图 4 马铃薯育成种亲本材料信息可视化

Fig. 4 Visualization of information on parental material of potato breeding stock

图 4 所示知识图谱为本文所研究马铃薯的亲本材料信息，其中包括部分国外引进品种。结合所构建知识图谱和相关马铃薯的品种选育文献可知，大西洋、早大白、庄薯 3 号、冀张薯 8 号、费乌瑞它、晋薯 7 号、斯凡特，这 7 个马铃薯品种与 5 个以上马铃薯之间存在遗传关系，在近年马铃薯新品种选育研究中属于常用育种材料。其中大西洋培育了毕薯 7 号、定薯 6 号、并薯 29、毕薯 6 号、毕薯 5 号、冀张薯 12 号、陇薯 8 号、闽薯 1 号、定薯 3 号，早大白培育了千薯 4 号、兴佳 3 号、本薯 6 号、陇薯 16 号、石薯 2 号、延薯 9 号、郑薯 9 号、本薯 4 号，庄薯 3 号培育了天薯 12 号、农天 1 号、天薯 11 号、天薯 10 号、陇薯 7 号，冀张薯 8 号培育了冀张薯 19 号、晋薯 25 号、冀张薯 24 号、京张薯 2 号、宁薯 18 号、春薯 11 号、陇薯 20 号，费乌瑞它培育了本薯 6 号、石薯 2 号、北方 009、中加 7、本薯 11 号、闽薯 1 号，晋薯 7 号培育了晋薯 26 号、晋薯 20 号、晋薯 19 号、晋薯 17 号、晋薯 13 号，斯凡特培育了冀围薯 8 号、冀农薯 8511、冀围薯 11 号、冀围薯 6 号、冀农薯 8862。而早期的核心亲本，如卡它丁、多子白、米拉、疫不加、小叶子和白头翁等[3]，在近年马铃薯新品种选育研究中使用较为少。

使用 K-Core 算法并设定 k 值为 2，得到马铃薯亲本信息的 2-Core 子图如图 5 所示。可见近年马铃薯新品种中，存在 8 个拥有紧密遗传关系的集合，共包含 62 个节点，在马铃薯育成种亲本材料知识图谱中占比为 13.08%。该 2-Core 子图所展示了具体遗传关系传递情况，上文所示的 7 个马铃薯常用亲本材料在该子图中均有呈现。

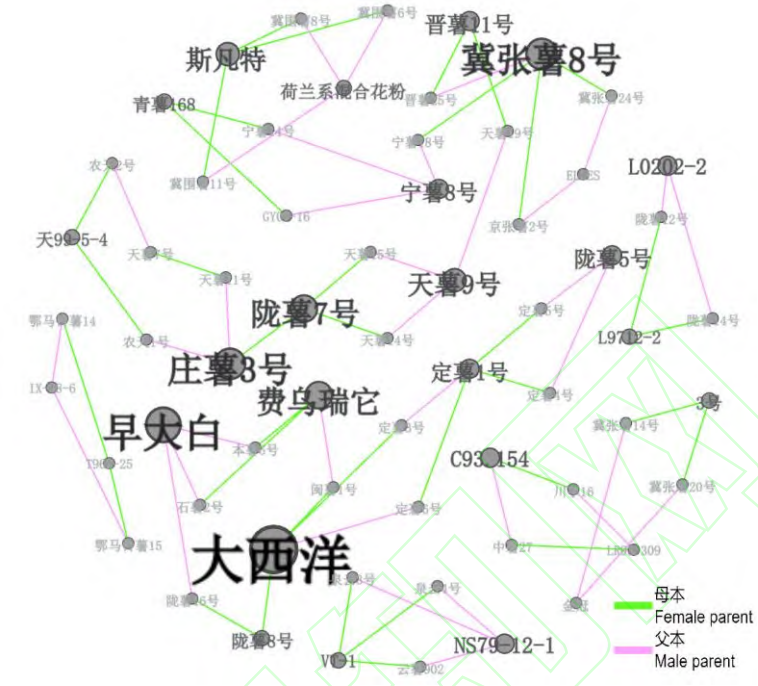

图 5 马铃薯育成种亲本材料 2-Core 子图

Fig. 5 Potato Breeding Seed Parental Material 2-Core Subplot

2.2 马铃薯不同抗性材料分析

马铃薯晚疫病[28]等病害会制约马铃薯发展，并造成重大经济损失，尤其是在多雨、冷凉、适于晚疫病流行的地区和年份。而且病原菌的快速变异会导致马铃薯抗病品种的抗性易丧失、病害防治难度加大。引进、鉴定、筛选马铃薯抗病资源，挖掘持久性抗病基因是病害防治的关键。本研究针对马铃薯育成种的抗性，从所构建的知识图谱中提取相关信息以进一步分析其抗性情况，对所提取信息进行可视化结果如图 6 所示，对应的 4-Core 子图如图 7 所示。

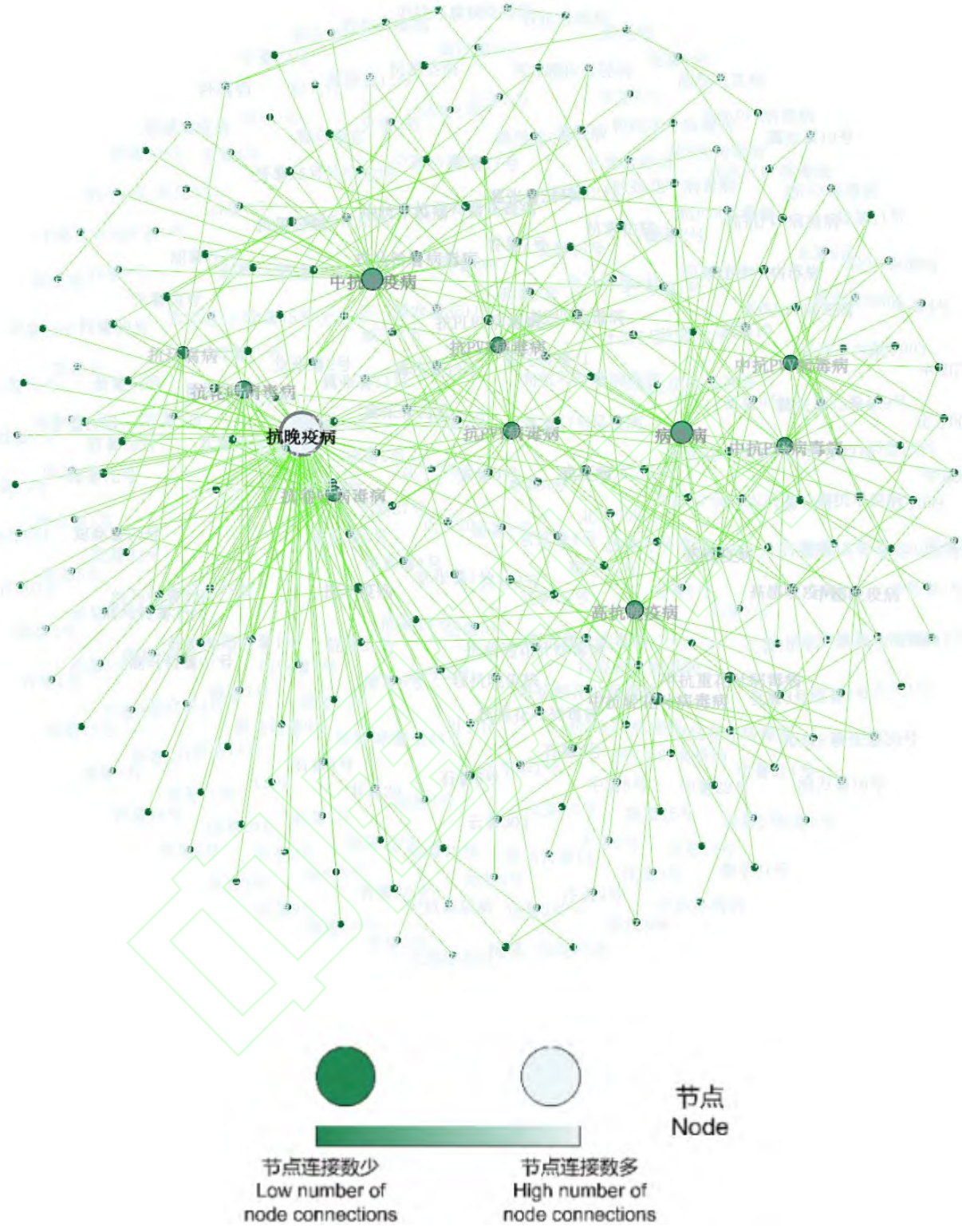

图 6 马铃薯育成种抗性材料可视化

Fig. 6 Visualization of resistance materials in potato breeding stock

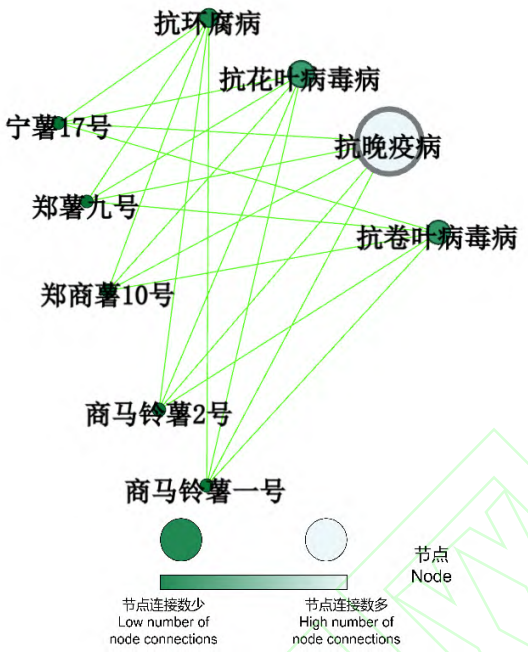

图 7 马铃薯育成种抗性材料可视化 4-Core 子图

Fig. 7 Potato Breeding Seed Resistance Material Visualization 4-Core Subplot

本研究对每种马铃薯拥有的抗性数和每种抗性对应的马铃薯品种数进行统计分析。可知，大部分马铃薯育成种拥有多种抗性，9 种马铃薯拥有 5 个以上的抗性，114 种马铃薯拥有 2 种以上抗性。大部分马铃薯育成种具有对晚疫病抗性，抗病毒病的品种数仅次于抗晚疫病的品种数。具体而言，抗晚疫病的马铃薯有 131 种，抗病毒病（包括 PVX 病毒、PVY 病毒、PVA 病毒、PVM 病毒、PVS 病毒、PLRV 病毒）的马铃薯有 109 种，抗早疫病的马铃薯有 13 种，抗环腐病的马铃薯有 13 种，抗抗黑胫病的马铃薯有 6 种，抗青枯病的马铃薯有 5 种，抗黑茎病的马铃薯有 3 种，抗旱的马铃薯有 3 种，抗疮痂病的马铃薯有 2 种。

此外，本研究还对拥有晚疫病抗性和病毒病抗性的马铃薯的亲本进行分析，统计培育出了 2 种及以上拥有对应抗性的马铃薯品种的育种材料，抗晚疫病和抗病毒病品种常用育种材料见表 2。

表 2 抗晚疫病和抗病毒病品种常用育种材料

Table 2 Common breeding materials for late blight and virus disease resistant varieties

抗性 Resistance	育种材料 Breeding materials	育成抗性品种数 Number of resistant varieties bred
抗晚疫病	冀张薯 8 号，大西洋	6
Late blight resistance	晋薯 11 号，L0202-2	5
	L9712-2，庄薯 3 号，斯凡特，宁薯 8 号，早大白，	4
	陇薯 7 号,C93.154,定薯 1 号,晋薯 7 号,克新 2 号,天薯 9 号,荷兰系混合花粉,秦芋 30 号, 9333-10	3
	虎头,青薯 168,天薯 7 号,小白花,合作 88,ELLES,陇薯 5 号，119-8,延薯 4 号,3 号, 合作 23,米拉,L0527-4,豫马铃薯 1 号,NS79-12-1,中薯 3 号	2
抗病毒病	费乌瑞它	4
Resistance to potato viruses	早大白	3
	高原 7 号,冀张薯 8 号,XS9304,克新 16 号	2

具体分析可知，冀张薯 24 号同时对早疫病、晚疫病和 5 种病毒病有抗性，因此对冀张薯 24 号的抗性系谱进行分析。如图 8 所示，由于京张薯 2 号与冀张薯 24 号拥有同一对亲本，将京张薯 2 号与冀张薯 24 号进行对比，可知京张薯 2 号和冀张薯 24 号均拥有早疫病、晚疫病、PLRV 病毒病和 PVY 病毒病的抗性，

并且冀张薯 24 号比京张薯 2 号多拥有 3 种病毒病抗性。但是，作为京张薯 2 号和冀张薯 24 号母本的冀张薯 8 号只拥有 PVA 病毒病和 PVY 病毒病的抗性并且轻中感晚疫病，京张薯 2 号和冀张薯 24 号均继承了 PVY 病毒病抗性，并对晚疫病具备抗性。由于 ELLES 缺乏抗性信息，因此可推测京张薯 2 号和冀张薯 24 号的晚疫病抗性大概率来自于父本 ELLES，且 ELLES 很可能还拥有 PVX 病毒病、PVM 病毒病和 PLRV 病毒病抗性。

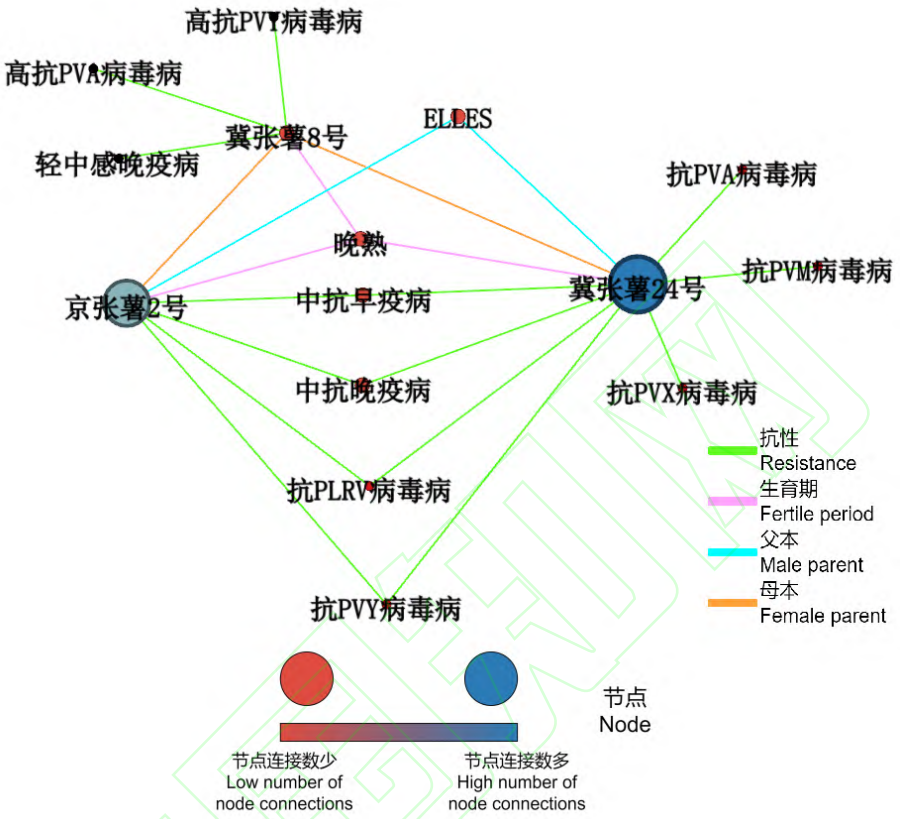

图 8 冀张薯 24 号抗性系谱分析

Fig. 8 Analysis of resistance genealogy of Jizhangshu 24

2.3 马铃薯不同生育期材料分析

根据生育期长短，生产上通常将作物品种划分为早熟品种、中熟品种和晚熟品种。其中，早熟作物生育期短，在降低植株感病率，提高土地复种指数，增加作物周年产量，促进农业高质量发展等方面具有重要意义。但是早熟基因与不良基因的连锁限制了早熟品种的培育和利用[29]。对马铃薯育成种的生育期进行分析，有助于获得早熟且具有稳定抗性的马铃薯品种。本研究针对马铃薯育成种的生育期，从所构建的知识图谱中提取相关信息以进一步分析不同生育期材料，对所提取信息进行可视化结果如图 9 所示。由于一种马铃薯只有一种生育期，因此无法构建 K-Core 子图。

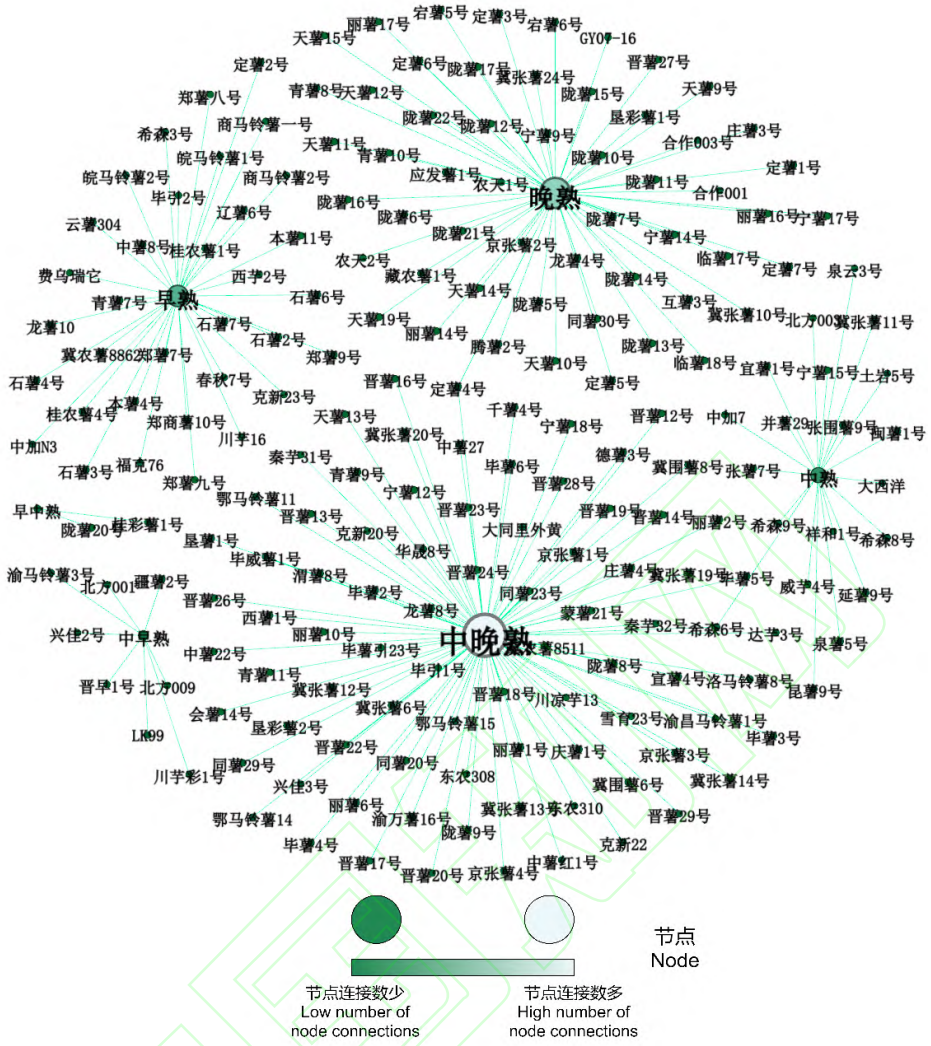

图 9 马铃薯育成种不同生育期材料可视化

Fig. 9 Visualization of potato materials at different stages of fertility

对所提取的生育期信息进行统计得到具体生育期对应马铃薯品种数，其中早熟 35 种，早中熟 2 种，中早熟 8 种，中熟 20 种，中晚熟 84 种，晚熟 59 种。。近年选育得到的马铃薯育成种的生育期大多为中晚熟、晚熟，早熟品种相对较少。针对早熟基因与不良基因的连锁对早熟品种的培育和利用造成限制的问题，本研究对拥有不同抗性的早熟品种进行统计，得到拥有不同抗性的早熟品种如表 3 所示。

表 3 不同抗性对应早熟品种

Table 3 Different resistance corresponds to early maturing varieties

抗性 Resistance	早熟品种 Early maturing variety
抗晚疫病 Late blight resistance	桂农薯 4 号，郑薯 7 号，西芋 2 号，青薯 7 号，商马铃薯 2 号，费乌瑞它，石薯 4 号，郑薯八号，郑薯九号，云薯 304，郑商薯 10 号，石薯 7 号，商马铃薯一号，桂农薯 1 号，川芋 16，福克 76，皖马铃薯 2 号，皖马铃薯 1 号，郑薯 9 号
抗病毒病 Resistance to potato viruses	克新 23 号，本薯 4 号，青薯 7 号，商马铃薯 2 号，费乌瑞它，中薯 8 号，郑薯八号，郑薯九号，云薯 304，中加 N3，石薯 3 号，辽薯 6 号，郑商薯 10 号，毕引 2 号，石薯 2 号，冀农薯 8862，商马铃薯一号，川芋 16，皖马铃薯 2 号，皖马铃薯 1 号，石薯 6 号，龙薯 10
抗早疫病 Early blight resistance	商马铃薯 2 号，郑商薯 10 号，费乌瑞它，石薯 7 号，石薯 4 号，商马铃薯一号
抗环腐病 Resistance to common	商马铃薯 2 号，郑商薯 10 号，郑薯九号，郑薯八号，商马铃薯一号，青薯 7 号

scab		
抗黑胫病		青薯 7 号
Resistance to blackleg disease		

3 讨论

充分开发利用马铃薯野生资源和原始栽培种质资源对马铃薯育种发展具有重要意义[30]。知识图谱作为一种将知识以图的形式进行有效组织的方法，能够应对信息杂散和无序所带来的问题[6]。通过对现有马铃薯品种信息的系统梳理与整合，可以明确马铃薯选育现状，进而为未来马铃薯育种工作提供参考。

农业领域现有知识图谱大多使用深度学习方法进行构建，但仍面临若干挑战，包括如何高效地批量处理海量数据、如何实现高精度的实体分类、如何实施多样化的数据标注策略，以及如何提高模型的稳定性和效率问题。使用深度学习方法时，训练模型通常需要大量的标注数据。这些数据用于训练模型识别实体、关系以及属性等元素，并将它们转化为向量表示，以便进行进一步的特征交互和学习。深度学习方法相较于传统人工方法所需人工和时间成本已大幅减少，但仍需花费一定的时间人工标注数据用于模型训练，并且参与知识图谱构建的研究人员所掌握的知识深度也会影响所构建的知识图谱呈现的知识。在大量的通用数据上进行预训练所得到的大语言模型拥有强大的性能，能够以更少的时间和人工成本完成知识图谱构建工作。大语言模型掌握广泛的知识，能避免数据标注人员知识的缺乏导致知识图谱质量不高问题，更有可能挖掘出隐藏的信息，从而推动马铃薯种质资源应用与创新。因此，本研究使用自然语言处理领域新技术 GraphRAG，探索由非结构化数据快速构建马铃薯品种选育研究知识图谱方案。

本研究提出的知识图谱构建方案相较于需要大量标注数据进行训练并且泛化性不高的深度学习的方法而言，自动化程度更高且泛化能力更强。在自动化程度上，利用 GraphRAG 进行知识图谱构建时，无需二次训练即可实现自动化实体识别、关系抽取和知识补全。尽管本研究在 GraphRAG 中使用的 Qwen2-70B-instruct 未经微调，缺少相关的领域知识，存在出现错误的可能性，但可以在获得节点表后进行人工审核以解决该问题。在泛化能力上，该方法不需要准备大量标注数据进行训练，只需要更换输入数据即可实现任何领域的知识图谱构建。

此外，由于同样的实体在不同领域内所代表的含义有可能存在一定差异，在 GraphRAG 中使用通用 LLM 会导致实体识别错误，使得所构建的知识图谱质量不理想。因此，利用通用 LLM 和 GraphRAG 实现自动化知识图谱构建时，对数据质量的要求要高于深度学习方法，数据内的专业名词、多义词、同义词等应是易于理解区分的，否则会对输出质量造成影响。

本研究构建知识图谱所使用的数据仅包含了马铃薯的抗性、生育期和亲本信息，缺少马铃薯亲本材料的抗性、生育期等信息，暂未对马铃薯的抗性、生育期结合系谱的继承关系进行深入挖掘。下一步的研究方向是探索更高精度的自动化育种知识图谱构建方法，并增加马铃薯育成种的数据资料和新登记的马铃薯信息，深入探索马铃薯选育中特性结合系谱的继承关系，以加速马铃薯育种。

4 结论

本研究主要完成了两部分工作，分别是基于 GraphRAG 的领域知识图谱构建和基于所构建的知识图谱对 227 个马铃薯育成种进行亲本、生育期、抗性材料分析。在基于 GraphRAG 的领域知识图谱构建部分工作中，本研究以中国知网（CNKI）数据库中的 227 种马铃薯育成种数据为基础，利用 Qwen2-70B-instruct 模型和 GraphRAG 构建知识图谱，对形成的知识图谱进行了进一步的人工审核和校验，获得对应的知识图谱节点表，使用 Gephi 软件对其进行可视化。在基于所构建的知识图谱分析亲本、生育期和抗性材料分析部分工作中，本研究单独提取所讨论信息并进行可视化，得到对应可视化结果和统计分析结果。综上，本研究探索了大语言模型相关技术在构建马铃薯种质资源领域知识图谱任务上的实现路径，并基于得到的知识图谱分析近年中国马铃薯选育材料，实现了对所构建知识图谱的应用，对未来大语言模型在马铃薯种质资源上的进一步应用具有一定指导意义。

参考文献

[1] 王鹏,李芳弟,颉炜清,窦俊焕,罗照霞,郭天顺,赵中梁,齐小东,杨晨,宋怡,吕汰.甘肃早熟马铃薯种质资源引进鉴定试验.种子, 2020,39(9):58-65

Wang P, Li F D, Jie W Q, Do J H, Lu Z X, Guo T S, Zhao Z L, Qi X D, Yang C, Song Y, Lu T Introduction and identification of early maturing potato germplasm resources in Gansu. Seed, 2020,39(9):58-65

[2] 金黎平.我国马铃薯育种和品种应用.农业技术与装备,2007(9):14-15

Jin L P. Potato breeding and variety application in China. Agricultural Technology and Equipment, 2007(9):14-15

[3] 段绍光.马铃薯种质资源遗传多样性评价和重要性状的遗传分析.北京: 中国农业科学院,2017

Duan S G. Genetic diversity evaluation and genetic analysis of important traits in potato germplasm resources.Beijing: Chinese Academy of Agricultural Sciences, 2017

[4] Gopal J. Heterosis and combining ability analysis for resistance to early blight (*Alternaria solani*) in potato. Potato Research, 1998, 41: 311-317

[5] 王磊.198 份 CIP 马铃薯种质资源的表型性状和晚疫病抗性的遗传多样性研究.兰州: 甘肃农业大学,2021

Wang L. Genetic diversity study on phenotypic traits and late blight resistance of 198 CIP potato germplasm resources.Lanzhou: Gansu Agricultural University, 2021

[6] 王巧珍.马铃薯产业链知识图谱构建研究.兰州: 甘肃农业大学,2023

Wang Q Z. Research on the construction of potato industry chain knowledge graph.Lanzhou: Gansu Agricultural University, 2023

[7] Xiao G, Corman J. Ontology-mediated SPARQL query answering over knowledge graphs. Big Data Research, 2021, 23: 100177.

[8] 刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述.计算机研究与发展,2016,53(3):582-600

Liu Q, Li Y, Duan H, Liu Y, Qin Z G. A review of knowledge graph construction techniques. Journal of Computer Research and Development, 2016, 53(3):582-600

[9] 孙亚茹,杨莹,王永剑.基于知信图卷积神经网络的开放域知识图谱自动构建模型.计算机工程,2022,48(10):116-122

Sun Y R, Yang Y, Wang Y J. An automatic construction model of open-domain knowledge graph based on knowledge trust graph convolutional neural network. Computer Engineering, 2022, 48(10):116-122

[10] 赵赛,杨婉霞,王巧珍,王梦瑶,熊磊.基于马铃薯病虫害知识图谱的问答系统.农业工程,2023,13(8):29-37

Zhao S, Yang W X, Wang Q Z, Wang M Y, Xiong L. A question-answering system based on a knowledge graph of potato diseases and pests. Agricultural Engineering, 2023, 13(8):29-37

[11] 徐帅博.基于枸杞病虫害知识图谱的问答系统研究与实现.银川: 宁夏大学,2020

Xu S B. Research and implementation of a question-answering system based on a knowledge graph of goji berry diseases and pests. Yinchuan: Ningxia University, 2020

[12] 宋璐璐.基于知识图谱的水稻病虫害问答系统的设计与实现.雅安: 四川农业大学,2023

Song L L. Design and implementation of a rice pest and disease Q&A system based on knowledge graphs. Ya'an: Sichuan Agricultural University, 2023

[13] 陈亚东,鲜国建,寇远涛,郭淑敏,刘现武.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

Chen Y D, Xian G J, Kou Y T, Guo S M, Liu X W. Construction of knowledge graph for apple industry in China. Chinese Journal of Agricultural Resources and Regional Planning, 2017, 38(11):40-45

[14] 申存,黄廷磊,梁霄.基于多粒度特征表示的知识图谱问答.计算机与现代化,2018(9):5-10

Shen C, Huang T L, Liang X. Knowledge graph question answering based on multi-granularity feature representation. Computer and Modernization, 2018(9):5-10

[15] Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, Li C, Li C, Liu D, Huang F, Dong G, Wei H, Lin H, Tang J, Wang J, Yang J, Tu J, Zhang J, Ma J, Yang J, Xu J, Zhou J, Bai J, He J, Lin J, Dang K, Lu K, Chen K, Yang K, Li M, Xue M, Ni N, Zhang P, Wang P, Peng R, Men R, Gao R, Lin R, Wang S, Bai S, Tan S, Zhu T, Li T, Liu T, Ge W, Deng X, Zhou X, Ren X, Zhang X, Wei X, Ren X, Liu X, Yang F, Yao Y, Zhang Y, Wan Y, Chu Y, Liu Y, Cui Z, Zhang Z, Guo Z, Fan Z. Qwen2 technical report. (2024-07-15) [2024-09-10] https://doi.org/10.48550/arXiv.2407.10671

[16] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring massive multitask language understanding. (2009-09-17) [2024-09-10]. https://doi.org/10.48550/arXiv.2009.03300

[17] Zellers R, Holtzman A, Bisk Y, Farhadi A, Choi Y. Hellaswag: Can a machine really finish your sentence? (2019-05-19) [2024-09-10]. https://doi.org/10.48550/arXiv.1905.07830

[18] Suzgun M, Scales N, Schärli N, Gehrmann S, Tay Y, Chung H W, Chowdhery A, Le Q V, Chi E H, Zhou D, Wei J. Challenging big-bench tasks and whether chain-of-thought can solve them. (2022-10-17) [2024-09-10]. https://doi.org/10.48550/arXiv.2210.09261

[19] Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, Plappert M, Tworek J, Hilton J, Nakano R, Hesse C, Schulman J. Training verifiers to solve math word problems. (2021-11-18) [2024-09-10]. https://doi.org/10.48550/arXiv.2110.14168

[20] Hendrycks D, Burns C, Kadavath S, Arora A, Basart S, Tang E, Song D, Steinhardt J. Measuring mathematical problem solving with the math dataset. (2021-11-08) [2024-09-10]. https://doi.org/10.48550/arXiv.2103.03874

[21] Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, Yang F, Ge W B, Han Y, Huang F, Hui B Y, Ji L, Li M, Lin J Y, Lin R J, Liu D Y, Liu G, Lu C Q, Lu K M, Ma J X, Men R, Ren X Z, Ren X C, Tan C Q, Tan S N, Tu J H, Wang P, Wang S J, Wang W, Wu S G, Xu B F, Xu J, Yang A, Yang H, Yang J, Yang S S, Yao Y, Yu B W, Yuan H Y, Yuan Z, Zhang J W, Zhang X X, Zhang Y C, Zhang Z R, Zhou C, Zhou J R, Zhou X H, Zhu T H. Qwen technical report. (2023-09-28) [2024-09-10]. https://doi.org/10.48550/arXiv.2309.16609

[22] Young A, Chen B, Li C, Huang C, Zhang G, Zhang G, Li H, Zhu J, Chen J, Chang J, Yu K, Liu P, Liu Q, Yue S, Yang S, Yang S, Yu T, Xie W, Huang W, Hu X, Ren X, Niu X, Nie P, Xu Y, Liu Y, Wang Y, Cai Y, Gu Z, Liu Z, Dai Z. Yi: Open foundation models by 01. ai. (2024-03-07) [2024-09-10]. https://doi.org/10.48550/arXiv.2403.04652

[23] Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, et al. The Llama 3 Herd of Models.(2024-08-15) [2024-09-10]. https://doi.org/10.48550/arXiv.2407.21783

[24] DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. (2024-06-19) [2024-09-10]. https://doi.org/10.48550/arXiv.2405.04434

[25] Edge D, Trinh H, Cheng N, et al. From local to global: A graph rag approach to query-focused summarization. (2024-04-24) [2024-09-10]. https://doi.org/10.48550/arXiv.2404.16130

[26] Li X C, Xu J F, Duan S G, Bian C S, Hu J, Shen H L, Li G C, Jin L P. Pedigree-based deciphering of genome-wide conserved patterns in an elite potato parental line. Frontiers in Plant Science, 2018, 9: 690

[27] 徐建飞,金黎平.马铃薯遗传育种研究:现状与展望.中国农业科学,2017,50(6):990-1015

Xu J F, Jin L P. Research on potato genetic breeding: current status and prospects. Chinese Journal of Agricultural Science, 2017, 50(6):990-1015

[28] 段绍光.马铃薯种质资源遗传多样性评价和重要性状的遗传分析.北京: 中国农业科学院,2017

Duan S G. Genetic diversity assessment and genetic analysis of important traits in potato germplasm resources. Beijing: Chinese Academy of Agricultural Sciences, 2017

[29] 李玲,单建伟,王丽,刘计涛,宋波涛,李小波.作物熟性遗传及相关基因功能研究进展.分子植物育种,2024,http://kns.cnki.net/kcms/detail/46.1068.S.20240226.1210.004.html

Li L, Shan J W, Wang L, Liu J T, Song B T, Li X B. Advances in the genetic study of crop maturity and the functions of related genes. Molecular Plant Breeding, 2024,http://kns.cnki.net/kcms/detail/46.1068.S.20240226.1210.004.html

[30] Danan S, Veyrieras J B, Lefebvre V. Construction of a potato consensus map and QTL meta-analysis offer new insights into the genetic architecture of late blight resistance and plant maturity traits. BMC Plant Biology, 2011, 11: 1-17