

基于知识图谱的石斛研究可视化分析与构建^①

李石荣¹, 符茂胜¹, 周先存¹, 王成², 余德勇³

1. 皖西学院 电子与信息工程学院, 安徽 六安 237012;

2. 霍山县天下泽雨生物科技发展有限公司, 安徽 六安 237200;

3. 皖西学院 生物与制药工程学院, 安徽 六安 237012

摘要: 石斛作为一种中药材因具有良好的药用价值和市场价值而被广泛地研究与应用. 本研究主要基于知识图谱的石斛研究可视化分析与构建, 预测其未来发展趋势, 为后续研究和临床应用提供重要参考依据. 研究选取中国知网数据库中近 20 年有关石斛研究的文献, 利用 Citespace 软件对作者、机构和关键词进行可视化分析, 结合半监督训练等算法来实现石斛领域的知识图谱构建. 实验结果表明: 1) 作者与机构存在较强地域性, 合作关系较为密切; 2) 研究重点主要围绕提取工艺、临床应用和数据挖掘等方面, 体现了研究的广度和深度; 3) 研究热点变化与市场需求联系紧密; 4) 构建的知识图谱与石斛研究领域结果相符合.

关键词: 石斛; 知识图谱; 可视化分析; 构建

中图分类号: TP391; F302

文献标志码: A

文章编号: 1000-5471(2022)06-0084-13

Visualized Analysis and Construction of Dendrobium Study Based on Knowledge Graph

LI Shirong¹, FU Maosheng¹,
ZHOU Xiancun¹, WANG Cheng², SHE Deyong³

1. College of Electronics and Information Engineering, West Anhui University, Lu'an Anhui 237012, China;

2. Huoshan Tianxia Zeyu Biological Technology Development Co., Ltd., Lu'an Anhui 237200, China;

3. College of Biological and Pharmaceutical Engineering, West Anhui University, Lu'an Anhui 237012, China

Abstract: Dendrobium has been widely researched and used as a kind of Chinese medicinal material due to its good medicinal value and market value. Analysis and construction of visualized dendrobium study based on Knowledge Graph has been completed in this research, which can predict future development trends and provide important reference for follow-up scientific research and clinical application. The references of Dendrobium research in the recent two decades from the China National Knowledge Infrastructure (CNKI) database were selected as the research data. The visualized analysis of authors, institutions and keywords can be done by Citespace software. The construction of knowledge graph in the dendrobium field can be realized by Semi-supervised training algorithm, etc. The experimental results can be shown as follows. 1) Au-

^① 收稿日期: 2021-05-24

基金项目: 国家自然科学基金项目(61572366); 安徽省自然科学基金项目(1908085MF213); 安徽省高校自然科学基金重点项目(KJ2019A0631).

作者简介: 李石荣, 讲师, 硕士, 主要从事智能信息处理、大数据分析等方面的研究.

thors and institutions have strong regionality and close cooperation. 2) The research mainly focuses on extraction technology, clinical application and data mining, which reflects the breadth and depth of research. 3) Changes in research hotspots are closely related to market demand. 4) The constructed knowledge graph is consistent with the results of the dendrobium research field.

Key words: dendrobium; Knowledge Graph; visualized analysis; construction

特色植物资源以其独特的药用价值和商用价值而备受政府的关注,在国民经济中具有重要的作用.党的十九届四中全会《决定》提出:要强化提高人民健康水平的制度保障,坚持以基层为重点、预防为主、防治结合、中西医并重^[1].山区作为中药材产业的重要基地,因其独特的地理位置而拥有丰富的石斛、天麻等多种特色植物资源^[2].特色植物资源的研究与开发可促进山区的经济发展、增加就业岗位和提高国民收入.目前,对特色植物的研究成果较多,任刚等^[3]对铁皮石斛的叶子中所含的主要化学成分进行了研究,并首次提取了若干化合物;袁青松等^[4]针对环境温度、种子品种对天麻抽薹特性的影响进行了相关的研究,得出乌天麻和红天麻的种子在发育过程中抽薹生产的最佳温度;张宗源等^[5]主要研究了组蛋白乙酰化成分对灵芝的生长、多糖和酸生物合成等方面的影响,得出可通过人为调控灵芝生长发育过程进而影响灵芝酸生物合成组蛋白乙酰化.

石斛作为特色植物资源中一种名贵的中药材而受到研究学者和机构的广泛关注.当前对石斛的研究大部分仅限于气候、土壤、地理位置或成分分析等内容^[6-8],在信息化建设发展上相对滞后,如何利用智能化手段挖掘石斛资源潜在的数据信息,构建“一张图”综合服务平台,有利于优化产业结构,便于政府和企业对石斛资源进行优化整合和决策管理,有利于促进新时期地区产业的经济的发展.随着计算机网络和大数据处理技术的不断发展,近些年来,基于大数据的知识图谱可视化分析得到了广泛的研究和应用^[9-15].将知识图谱应用于日常生活和行业发展是未来的发展趋势.目前,知识图谱可视化在智慧教育、智能医疗和智慧农业等多个领域已经实现了成功的应用^[16-22].侯梦薇等^[16]指出医学知识图谱在临床决策知识、语义检索等医疗服务中具有重要的意义,通过集成术语构建大规模的知识图谱可解决医疗中存在的资源不足和需求矛盾等重要问题.Chen等^[18]提出了一种 KnowEdu 系统,利用神经序列标记算法提取教学概念,通过概率关联规则挖掘教育领域的重要信息.陈曦等^[19]指出基于课程知识图谱的预测算法可降低预测误差,获得更好的学生成绩预测效果.侯丽等^[20]对我国农业重金属污染现状进行了知识图谱可视化研究,为农田污染的治理提供了智慧决策.杜师傅等^[21]对国内景观评价方法进行了可视化分析,直观地揭示了国内景观评价方法的发展现状、研究热点和动态方向,为政府和商家提供了智慧性决策方案.王友发等^[22]总结了近20年智能制造研究的热点并对前沿进行挖掘,提出了未来智能制造的可能发展方向.

目前知识图谱已经在多个行业中实现了成功的研究和应用,然而关于特色植物资源的图谱应用却很少,尤其是关于石斛的知识图谱可视化分析的研究文献更少.石斛作为一种特色植物资源在国民经济和社会发展中具有重要的作用,本文基于 CNKI 数据库中有关石斛的文献,利用大数据技术对获取的数据进行知识抽取和融合,实现面向石斛的大数据知识图谱可视化分析,可直观地描述石斛资源政策、产业和研究方向的相关信息.实验选取了与石斛研究相关的作者、机构和关键词等内容进行了相应的知识图谱可视化分析,根据分析的结果并通过分词和词性标注来进行针对性的信息抽取,通过半监督训练实现经验数据打标签,最后将打标整理好的数据存入数据库并导入 Neo4j 平台实现知识图谱的构建.实验结果表明:近20年来有关石斛的文献研究经历了“缓慢增长、快速增长和稳定波动”3个阶段;作者与机构的聚类网呈现“一主网集中、多小网分散”的关系,有关石斛研究的大部分作者和机构之间存在合作关系,石斛研究的科研单位主要与石斛生长的地理位置有关,石斛研究的作者和机构随着时间的变化也会有所变动;关键词共现和聚类分析结果显示:有关石斛研究的聚类关键词可主要聚集为5大类,即糖尿病、多糖、石斛、遗传稳定性和数据挖掘.聚类糖尿病主要侧重于石斛的中医疗效、免疫功能等内容;聚类多糖主要侧重于石斛主要成分、采收期、提取工艺等内容;聚类石斛主要侧重于石斛的种类、栽培技术、产业推广等内容;聚类遗传稳定性主要侧重于石斛的组织培养、遗传基因、光合作用等内容;聚类数据挖掘主要侧重于石斛的中医传承辅助平台、用药规律、关联规则等内容.关键词聚类图谱可视化结果分别从石斛的药用、成分、属性、遗传

性和智能数据分析方面对石斛研究进行系统性的分析和展示,对热点的分析也反映了近 20 年来研究机构对石斛研究重点的迁移变化.根据石斛研究知识图谱可视化分析的结果,利用 Python 网络爬虫抓取知网摘要关键词,根据分词和词性标注建立信息抽取训练模型,通过经验分析和半监督训练对数据进行打标签,最后将整理好的打标数据存入数据库并导入 Noe4j 平台,实现具有石斛研究针对性的知识图谱结果.本文不仅基于 Citespace 软件实现了石斛研究知识图谱可视化分析,而且根据分析的结果建立半监督训练信息抽取模型,最终构建了针对石斛研究的知识图谱.

本文利用 CNKI 数据库关于石斛研究的数据资源,结合 Citespace 软件对石斛的研究动态、机构合作、摘要关键词与热点等内容做图谱分析,构建了面向石斛资源研究的“一张图”.得到的图谱可以为有关石斛研究人员和机构的研究方向提供参考性建议,为有关地方政府和商户提供重要的决策信息和指导,对促进石斛研究、提高石斛的产业化、带动地方经济的发展和提高国民收入具有重要的意义.

1 知识图谱可视化分析与构建原理

1.1 可视化分析

本文利用 Citespace 软件对知网数据进行可视化处理,流程图如图 1 所示.可视化过程主要分为 4 个步骤:数据准备、参数设置、聚类分析与图谱分析.

1.1.1 数据准备

从 CNKI 数据库下载的数据并不能直接应用于 Citespace 软件进行分析,可通过软件自带数据转换功能,对数据进行格式转换.建立两个文件夹,分别用于存储原始数据和转换数据.

1.1.2 参数设置

Citespace 软件在建立项目过程中需要对参数进行设置,主要参数为时间分割、节点阈值筛选、网络精简、强度关联.

1.1.2.1 时间分割

时间分割的主要功能是对分析的数据在时间段上进行分割,确定时间跨度,分析石斛在各个时间段的主要研究热度和变化趋势.

1.1.2.2 节点阈值筛选

节点阈值筛选主要用于在设定的时间段内提取所需对象的个数.阈值筛选包括在增加规模因子 k 的基础上,通过 g 指数进行排名修正实现对知识单元的抽取.公式为

$$g \leqslant \frac{1}{g} \sum_{i=1}^g (k \cdot c_i)$$

(1)

其中 k 为规模因子,分别取 10, 20, 30, ... 依次进行尝试; c_i 为第 i 文献被引用次数.

1.1.2.3 网络精简

Citespace 软件中主要提供了两种网络精简算法,寻径网络(pathfinder network, PFNET)算法和最小生成树(minimum spanning tree, MST)算法.

1.1.2.3.1 PFNET 算法

PFNET 网络算法的主要结构由两个参数确定,基于闵氏距离的网络节点路径长度 r 和满足三角不等式的连接数量最大值 q .当 $r = 2$ 时,距离测度为欧氏距离.测度空间确定时,三角不等式的关系可定义为

$$w_{ij} \leqslant \left(\sum_k w_{n_k n_{k+1}}^r \right)^{1/k}$$

(2)

式中, w_{ij} 表示第 i 和 j 节点之间的连接权重, $w_{n_k n_{k+1}}$ 表示第 n_k 和 n_{k+1} 节点之间的连接权重, $k = 1, 2, 3, \dots$.当第 $i = n_1$ 和 $j = n_k$ 时,备选路径将经过所有节点,该网络包含所有中间连线.若备选路线权重小于

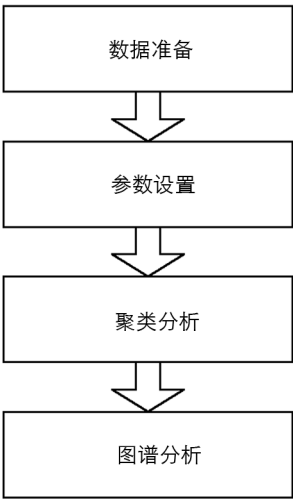


图 1 可视化分析流程图

w_{ij} 的值, 那么等式不成立, 第 i 和 j 节点之间的连接线将被删除. q 参数主要用于备选路线, 取 $[2, N-1]$ 内任意整数, N 为节点的个数. 当 $r \rightarrow \infty$ 时, $q = N-1$, 网络可实现最大裁剪能力.

1.1.2.3.2 MST 算法

最小生成树算法的构造是利用连通网构造实现代价最小的生成树. 给定一个无向图 $G = (V, E)$, V 和 E 分别表示数字顶点集和边集, 若存在边集 T 为 E 的子集并且为无循环图, 同时满足等式(2)且值最小, 则 T 为满足该条件的最小生成树.

$$w(T) = \sum_{(i,j) \in T} w(i,j) \quad (3)$$

式中 (i, j) 表示顶点 i 和 j 连接的边, $w(i, j)$ 表示 (i, j) 的权重.

无论是采用 MST 算法还是 PFNET 算法, 目的都是希望对生成的网络进行裁剪, 降低网络的复杂性和提高图谱的可读性. 与 PFNET 算法相比, MST 算法得到的图谱会更加简洁清晰, 但丢失的节点信息更多. 对于图谱过于庞大和混乱时, 需要选择 PFNET 算法得到的修剪效果会更好.

1.1.2.4 强度关联

Citespace 软件提供了网络分布中用于计算连接强度的 cosine 算法, 如公式(4):

$$\text{cosine}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\sqrt{s_i s_j}} \quad (4)$$

标准化后的数值在 $0 \sim 1$ 之间, 其中 c_{ij} 表示节点 i 和 j 共现次数, s_i 表示节点 i 共现的频次, s_j 表示节点 j 共现的频次.

1.1.3 聚类分析

Citespace 主要利用最大期望(EM)聚类算法对网络节点进行聚类处理. 首先, 将所有节点随机地分配到 K 个聚类簇中; 其次, 统计每个节点在 K 个聚类簇中的分布比例; 再次, 求出模型所需的参数. 输入 m 个样本观察数据 $x = (x^{(1)}, x^{(2)}, \dots, x^{(m)})$, 找到样本的模型参数 θ , 极大化模型的对数似然函数为

$$\theta = \arg \max_{\theta} \sum_{i=1}^m \log P(x^{(i)}; \theta) \quad (5)$$

由于观察数据中存在未观察到的隐含数据 $z = (z^{(1)}, z^{(2)}, \dots, z^{(m)})$, 此时极大化模型的似然函数可变为

$$\theta = \arg \max_{\theta} \sum_{i=1}^m \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta) \quad (6)$$

为求出模型参数, 对式子进行缩放可得

$$\begin{aligned} \sum_{i=1}^m \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta) &= \sum_{i=1}^m \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \geq \\ &= \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \end{aligned} \quad (7)$$

$Q_i(z^{(i)})$ 为未知的分布, 根据詹森不等式, 由上式可以得到

$$Q_i(z^{(i)}) = \frac{P(x^{(i)}, z^{(i)}; \theta)}{\sum_z P(x^{(i)}, z^{(i)}; \theta)} = P(z^{(i)} | x^{(i)}; \theta) \quad (8)$$

公式(6)中包含隐含数据的对数似然下界, 通过去掉常数部分, 可将极大化对数似然下界表示为

$$\arg \max_{\theta} \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) \log P(x^{(i)}, z^{(i)}; \theta) \quad (9)$$

$\sum_{z^{(i)}} Q_i(z^{(i)}) \log P(x^{(i)}, z^{(i)}; \theta)$ 可以理解为 $\log P(x^{(i)}, z^{(i)}; \theta)$ 基于条件概率分布 $Q_i(z^{(i)})$ 的期望.

最后, 将得到的每个词分别分配给 K 聚类簇的概率, 选取 K 个数值中最大值作为该词所属的聚类簇.

Citespace 提供了从标题、关键词和摘要中提取聚类命名, 提供的算法主要有潜语义索引算法、假设检验算法和互信息算法, 可提取研究术语并强调研究特征.

1.1.4 图谱分析

选择合适的研究术语、时间分割跨度、阈值、精简算法进行可视化显示, 借助软件提供的可视化工具

选项对图谱进行美化,也可以利用提供的网络计算功能对图谱网络进行深度分析,最后对分析的结果进行验证.

1.2 知识图谱构建

本文利用信息抽取技术、半监督学习算法和 Neo4j 平台等相结合方法对知网数据库近 20 年的文献作者、机构、年份和关键词等进行处理,构建针对石斛研究领域的知识图谱.

1.2.1 信息抽取

知识图谱的构建离不开对实体进行识别,借助词性标签,从有关石斛的摘要内容中提取重要词语.当需要的关键词语跨越多个词语时,仅仅利用词性标签是不够的,需要解析句子中的词语之间的依赖关系即依赖解析.如提取关键词和修饰词、提取复合词、提取词语之间的标点符号等.图谱实体之间的关系抽取也需要使用依赖解析,提取实体关系需要构建实体与关系之间的架构.

1.2.2 半监督学习算法

半监督学习充分将监督学习和无监督学习的优势相结合,通过少量的数据样本对总体数据进行训练,降低人工成本和提高工作效率.本文利用 Bootstrapping 算法对石斛研究关键词语进行关系抽取,对数据进行训练后实现标签设置.主要步骤有:

- 1) 人工标记少量石斛研究相关数据作为种子集并训练一个初始实体识别模型;
- 2) 设定条件概率阈值,将实体识别模型对未标记数据进行预测处理获取数据关系标签和概率大小,当输出概率大于阈值时,将数据标记为可靠数据并组成一个可靠集;
- 3) 当该可靠集数据大于 500 条时,将可靠集与种子集合并成一个新的标注数据集,重新训练获取实体识别模型;
- 4) 重复上述步骤,直至结束.

1.2.3 Neo4j 平台显示

选择 Neo4j 作为知识图谱构建平台,将提取的实体和关系进行打标签并整理好,导入数据库并在 Neo4j 平台中显示.

2 实验结果与分析

由于 Citespace 软件对 CNKI 的数据分析功能有限,本实验就文献的概况、作者与机构分析、关键词与热点分析等几个部分内容作重点分析,最后利用半监督模型训练后在 Neo4j 平台构建知识图谱.

2.1 数据收集和处理

本文在中国知网下载了有关石斛的文献,选取主体为“石斛”进行检索,时间、作者、单位等条件不限,初步统计文献有 9 024 篇,删除检索结果中的新闻、会议通知和重复等内容,最终获得文献共计 5 729 篇作为本次实验的基础数据.

本文是基于石斛的知识图谱可视化分析,主要就研究领域的作者、机构、关键词等核心内容进行分析与筛选,展示石斛在不同阶段研究关注的热点、发展动态等,揭示未来的石斛发展趋势.

2.2 文献基础分析

不同年份关于石斛研究的文献数量可以反映出该领域研究的热点和趋势,从图 2 中可以看出有关石斛近 20 年研究文献的变化趋势,有关石斛的研究经历了 3 个阶段:缓慢增长、快速增长和稳定波动.

缓慢增长阶段:2001—2011 年,有关石斛的研究文献在 10 年内从约 100 篇稳步增长到约 400 篇,石斛濒危问题得到了较好的解决,关于石斛的商业化生产仍处于起步阶段,关于石斛的研究、种植和销售等方面引起了广大研究学者和政府部门人员的关注^[23].

快速增长阶段:2012—2015 年,石斛的研究文献在这个时间段内实现了快速增长,2015 年达到了约 800 篇,关于石斛的研究和商业化处于快速发展时期,期间如铁皮石斛、霍山石斛等品种在中药研究、品种培育和商业推广等方面取得了众多成果^[24-25].

稳定波动阶段:2016—2019 年,石斛的研究文献在这个期间处于缓慢增长和波动阶段,关于石斛的研究和商业化基本上处于成熟阶段,石斛未来的研究朝向多元化和智能化发展^[26].

文献的来源期刊可直接反映石斛研究的热点, 如图 3 所示, 有关石斛研究的文献期刊主要为与中药材和农业发展相关的期刊。

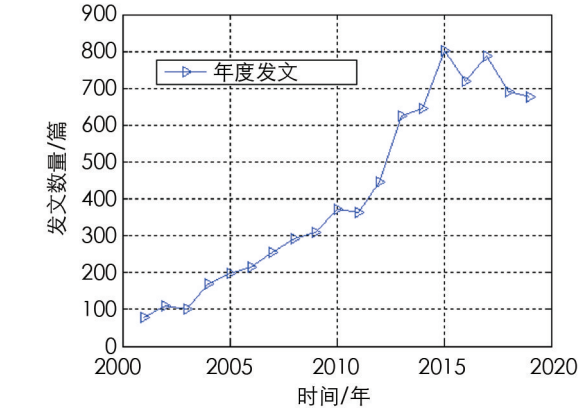


图 2 石斛相关文献发表量年度趋势

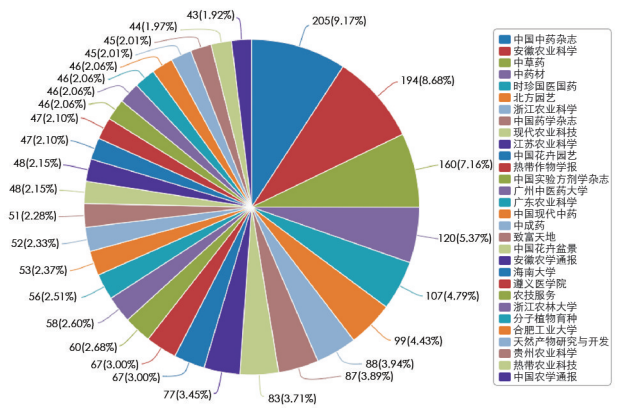


图 3 石斛研究文献期刊分布

2.3 作者与研究机构分析

作者是研究石斛的主体, 作者之间的联系可以观察出在石斛研究过程中的合作交流情况, 运用 Citespace 软件可以获得关于石斛研究的作者图谱原图、MST 图谱和 PFNET 图谱. 如图 4—图 6 所示, 原图谱、MST 图谱和 PFNET 图谱的节点均为 799 个, 连线分别为 1 419, 710 和 1 105 个, 网络密度分别为 0.004 5, 0.002 2 和 0.003 5. 与原图谱相比, MST 图谱和 PFNET 图谱更加清晰简捷, 但部分节点信息会被省略, 与 PFNET 图谱相比, MST 图谱丢失了更多的结构信息。

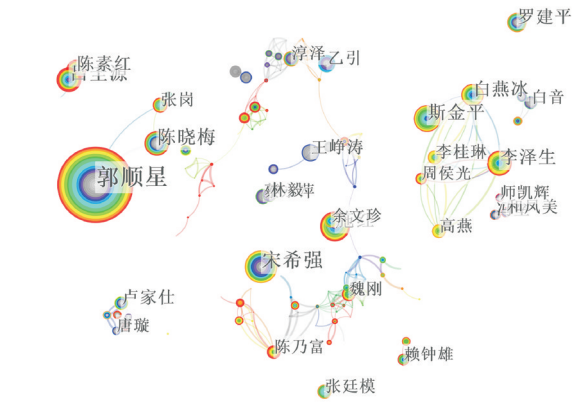


图 4 作者图谱原图

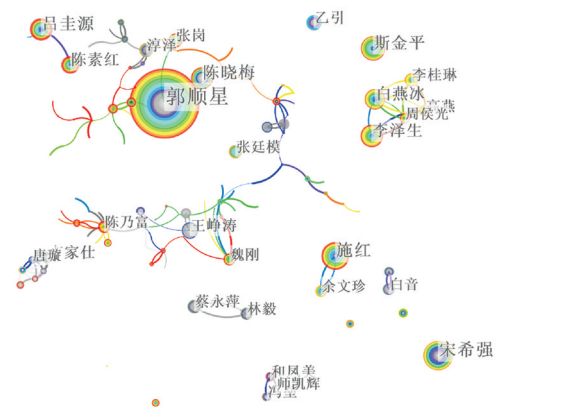


图 5 MST 作者图谱

统计核心作者的发文数量, 有关石斛研究的发文量大于 20 篇的作者共有 42 人, 从图 5—图 6 中可以看出作者之间的关系主要分为几大聚类关系, 大部分作者之间存在一种合作研究关系. 表 1 展示了石斛研究的主要作者, 从图 5—图 6 图谱中可以发现, 这些作者在整个石斛的研究过程中起到重要的作用. 其中陈乃富、魏刚、郭顺星等作者组成了一张较大的研究关系网, 而宋希强、李泽生等作者各自组成了较小的研究关系网, 石斛研究领域作者关系网呈现的是“大网集中, 多网分散”的趋势. 表 2 给出的是作者在不同时间段对石斛研究的主要内容变化, 体现石斛研究演变趋势。

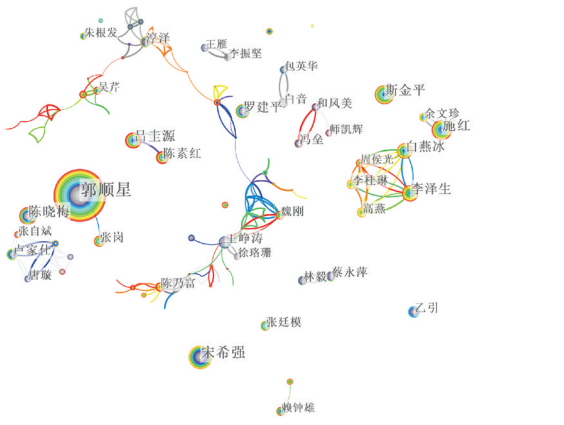


图 6 PFNET 作者图谱

表 1 石斛研究部分核心作者发文量统计表

作者	郭顺星	徐璐珊	张廷模	魏刚	陈乃富	淳泽	乙引	宋希强	李泽生
发文量	43	49	28	37	31	32	37	35	56

表 2 作者石斛研究主要内容变化

时间段	研究重点	研究主要内容
2001—2011	氨基酸遗传基因培养生长	研究石斛所含主要成分，分析石斛的基因遗传特性，对石斛的生长环境进行研究和分析
2012—2015	药用石斛提取工艺产业化	研究石斛主要成分的药用价值，研究石斛成分的提取工艺技术，研究石斛的产业化发展
2015—2020	石斛多糖品质免疫调节	主要分析和提取铁皮石斛、金钗石斛等品种中的多糖成分，对石斛工艺的品质进行研究分析，分析石斛主要成分对免疫系统的作用

机构是石斛研究的主要平台，石斛研究的主要机构包括科研院校、研究所和企业等. 图 7 为 MST 算法处理的机构知识图谱，可以发现有关石斛的研究基本上为科研院校和研究所，其中郭顺星研究员所在的中国医学科学院北京协和医学院药用植物研究所、魏刚研究员所在的广州中医药大学和陈乃富教授所在的皖西学院生物与制药工程学院等单位均为石斛主要的研究和发文机构，图谱结果可以看出研究机构之间的合作关系强度. 从图谱结果的侧面分析可以得出，广州中医药大学发文数量最高达到了 189 篇，其余包括皖西学院在内的 9 家研究机构发文量也突破了 100 篇，体现了这些高校和科研院所对石斛研究的深度.

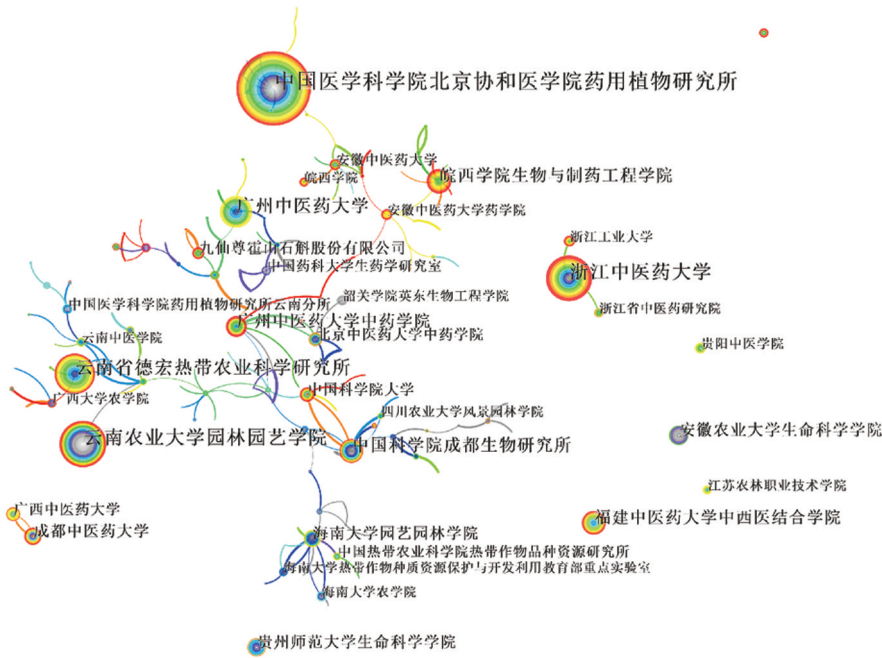


图 7 研究机构知识图谱

由于作者是研究机构的主体，基于时序的作者知识图谱可视化分析可直接反映出研究机构在石斛领域的变化情况. 图 8 给出的是通过聚类得到的近 20 年来机构基于时序的知识图谱可视化图. 聚类结果主要是通过关键词进行聚类分析获取，每个聚类结果中包含多个关键词，不影响作者的时序图谱. 时序图谱可以看出研究机构对石斛研究的参考文献从什么时间开始出现、聚类结果对应的研究机构所出成果从哪些年份开始增多以及聚类结果在哪些年份关注度开始降低等内容. 从图 8 中可以看出，郭顺星、陈晓梅等所在的中国医学科学院北京协和医学院药用植物研究所从 2000 年开始就已对石斛的栽培生长、主要成分和基因分析等内容进行了研究并持续至今，在石斛研究领域具有一定的广度和深度；罗建平、刘咏等所在的合肥工业大学生物与食品工程学院主要从 2005 年开始对霍山石斛的原球茎、多糖等进行了研究；陈乃富、韩邦兴等所在的皖西学院主要从 2009 年开始对霍山石斛的生长栽培、遗传性和药用性等进行了研究；李泽

生、李桂琳等所在的云南省德宏热带农业科学研究所主要从 2011 年开始对石斛的生长、产业化等内容进行了研究。

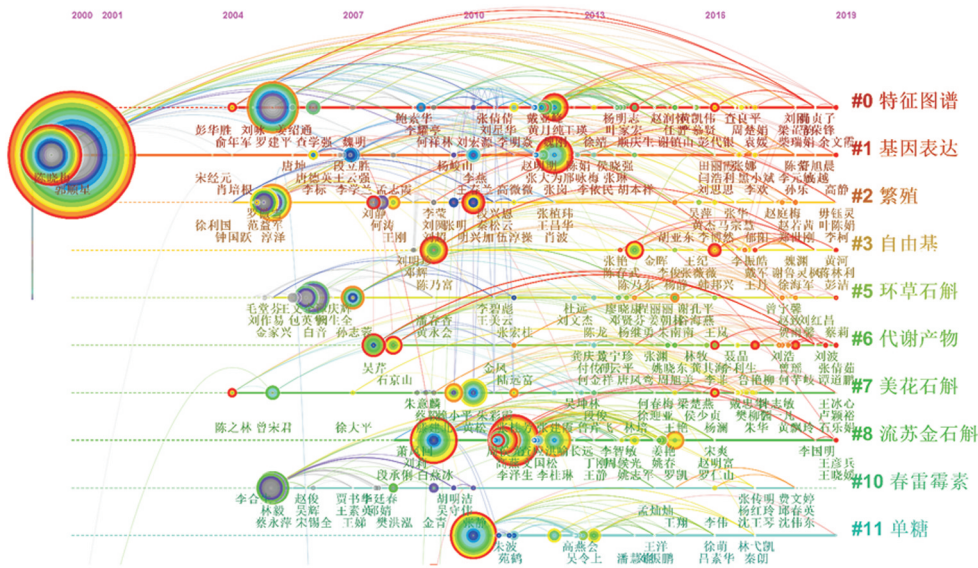


图 8 研究机构主体知识图谱时序图

图 9 展示了近 20 年来根据时间段对石斛研究的主要研究院所关联强度显示图. 可以发现广西壮族自治区农科院花卉研究所和云南农业大学园林园艺学院是最早开始研究石斛的研究所和高校, 这与广西、云南地理环境和盛产石斛有关, 但持续时间仅到 2007 年. 联系最强的前三个单位是安徽农业大学生命科学学院、贵州师范大学生命科学学院和皖西学院生物与制药工程学院, 这与贵州金钗石斛和安徽霍山石斛的药用价值高和近些年得到政府和科研院所的重视有关. 时间最近的两个单位为皖西学院生物与制药工程学院和浙江工业大学, 这也体现了近几年来这两个单位对石斛的研究成果较多。

Top 18 Institutions with the Strongest Citation Bursts

Institutions	Year	Strength	Begin	End	2000 - 2019
广西壮族自治区农业科学院花卉研究所	2000	5.3194	2000	2007	
云南农业大学园林园艺学院	2000	7.7385	2000	2006	
中国药科大学生药学研究室	2000	6.6617	2004	2008	
贵州师范大学生物技术与工程学院	2000	4.7468	2004	2007	
合肥工业大学生物与食品工程学院	2000	7.8808	2005	2010	
安徽农业大学生命科学学院	2000	11.6573	2005	2010	
玉林师范学院化学与生物系	2000	5.531	2005	2007	
韶关学院英东生物工程学院	2000	8.653	2006	2009	
中国林业科学研究院林业研究所	2000	4.608	2006	2012	
贵州师范大学生命科学学院	2000	8.9159	2009	2012	
海南大学农学院	2000	5.4437	2009	2013	
中国科学院研究生院	2000	5.3907	2009	2012	
海南大学园艺园林学院	2000	6.5978	2009	2012	
广州中医药大学	2000	6.1769	2010	2014	
贵阳中医学院	2000	5.8581	2014	2017	
浙江农林大学亚热带森林培育国家重点实验室培育基地	2000	5.1843	2014	2016	
皖西学院生物与制药工程学院	2000	8.7288	2015	2019	
浙江工业大学	2000	5.8744	2016	2019	

图 9 石斛研究机构关联强度图

多数石斛的生长与地理位置具有较强的关系. 表 3 为我国主要的几种石斛与地理位置、生长环境关系统计表. 表 3 中可以看出球花石斛、霍山石斛和铁皮石斛等不同品种所在的生长地区有所差异, 主要与不同品种石斛对生长环境的要求各不相同有关, 这与图 7、图 9 所示的研究机构知识图谱和关联强度图相契合. 石斛研究机构主要以云南、安徽、广西、四川和贵州等少数省份为主, 部分经济发达地区如北京、广州等城市的少数研究机构在石斛研究领域也有所涉及.

表 3 石斛种类与地理位置统计表

石斛种类	盛产地区	环境要求
球花石斛	云南	附生于海拔 750 米到 1800 米的山地林中树干上
霍山石斛	安徽	生长在云雾缭绕的悬崖峭壁崖石缝隙间和参天古树上
铁皮石斛	云南、广西、浙江、贵州、湖北	半阴湿的岩石和湿度较大的原始森林高大树干上, 温暖湿润气候和半阴半阳的环境
流苏石斛	广西、贵州、云南	生于密林中树干上或山谷阴湿岩石上
金钗石斛	四川、广西、云南、贵州	对大气、土壤、水质要求高, 怕严寒、喜高温高湿, 对温暖、湿润气候要求十分严格

2.4 关键词与热点分析

2.4.1 关键词共现分析

关键词是表达文献核心内容的重要信息, 也是研究该领域重点的提炼, 在一定程度上可以反映该领域的研究发展方向和研究的方法技术. 运用 Citespace 软件可以生成有关石斛研究的关键词共现知识图谱, 如图 10—图 12 所示, 分别为图谱原图、MST 图谱和 PFNET 图谱. 由于图谱中的关键词较多, 得到的原图图谱比较庞大和混乱, MST 算法和 PFNET 算法可解决这个问题. 如图 11 所示, MST 图谱可以展示有关石斛研究的核心关键词, 其中铁皮石斛、多糖、组织培养等是石斛研究的重要关键词, 关键词部分主要涉及到石斛的品种、培育生长、成分提取与分析、药用价值和产业化等内容. 石斛的品种主要分为铁皮石斛、霍山石斛和金钗石斛等, 其中铁皮石斛需要生长在海拔较高的山谷和树木上, 霍山石斛需要生长在海拔较高的悬崖峭壁石缝中或古树上; 铁皮石斛的花、茎、叶均可入药, 而金钗石斛只有茎可以入药; 石斛中的多糖可用于治疗糖尿病, 其中霍山石斛的药用滋阴功能较其他两种石斛更强; 霍山石斛由于需要野生栽培, 故其产业化推广应用最差.

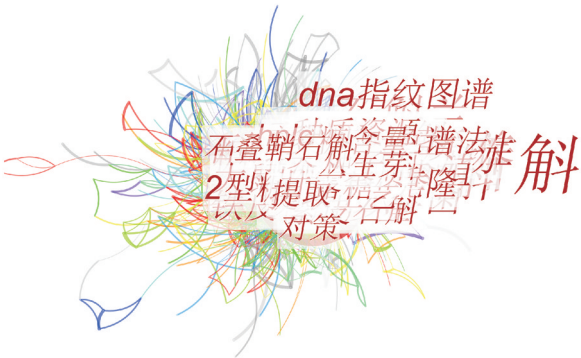


图 10 石斛研究关键词共现图谱原图



图 11 石斛研究关键词共现 MST 图谱



图 12 石斛研究关键词共现 PFNET 图谱

2.4.2 关键词聚类分析

Citespace 软件中关于关键词聚类时间线的可视化视图可用于描绘聚类时间的关系以及聚类中文献的历史跨度。如图 13 所示, 有关石斛研究的关键词聚类大体分为 11 类, 本文重点分析糖尿病、多糖、石斛、遗传稳定性和数据挖掘 5 大聚类。

从聚类结果来看, 聚类结构较显著(聚类模块值为 0.58), 聚类较合理(聚类平均轮廓值为 0.52)。下面着重分析实验得到的 5 个聚类。



图 13 关键词知识图谱时序图

- 1) 糖尿病. 主要内容有作用机制、降糖降脂、大鼠模型、免疫、脂多糖、抗炎等. 可以看出从 2000 年开始研究石斛中所含的重要成分对糖尿病的作用机制, 到 2004 年左右开始通过对糖尿病模型大鼠进行实验, 2007 年开始提出了中医药疗法, 2012 年开始研究石斛对降低人体血糖、抗炎和提高免疫力均有一定的疗效, 2017 年开始提出西洋参、灵芝等中药对改善糖尿病具有一定的效果。
- 2) 多糖. 主要内容有含量测定、采收期、提取工艺、原球茎等. 2000 年开始研究石斛多糖含量测定技术和成分提取方法, 2005 年开始研究石斛多糖抗氧化性能对抗衰老的作用, 2015 年开始研究石斛活性多糖对提高人体白细胞数量和提高免疫力等功效, 2018 年开始研究并提出铁皮石斛花中含有的多糖具有抗脂质过氧化作用和提高人体免疫力。
- 3) 石斛. 主要内容有铁皮石斛、产品质量、栽培技术、采收加工、产业开发等. 该部分为类别涵盖最多的一类聚类, 该聚类合并了铁皮石斛和霍山石斛等聚类关键词, 并首先对石斛的品种进行分类, 对比不同品种石斛的品质质量; 其次对不同石斛的栽培技术进行研究, 实现人工产业化栽培并进行专业的采收加工; 最后实现石斛产品的产业化推广, 推动地方经济建设的发展。
- 4) 遗传稳定性. 主要内容有组织培养、可控环境、遗传多样性、光合作用、筛选、适应性等内容. 主要研究石斛在不同条件下遗传稳定性及萌果之间的差异及有关遗传稳定性的问题。
- 5) 数据挖掘. 主要内容有中医传承辅助平台、用药规律、关联规则、因子分析等. 关于石斛数据挖掘的主要时间节点在 2016 年国务院发布《中国的中医药》白皮书后, 强调要着力推动中医药发展, 实现中医药健康养生文化的创新性. 主要利用临床数据信息的基础数据库对石斛在药物进行描述性分析和系统聚类分析时, 分析石斛在临床治疗上的疗效。

图 13 所示的关键词共现时间线聚类图谱中包含 1 500 多个节点和 3 500 多条连接线, 图谱可视性较

弱, 为提高图谱可视性, 可通过提高(c , cc , ccv)参数, 得到如图 14 所示的图谱。

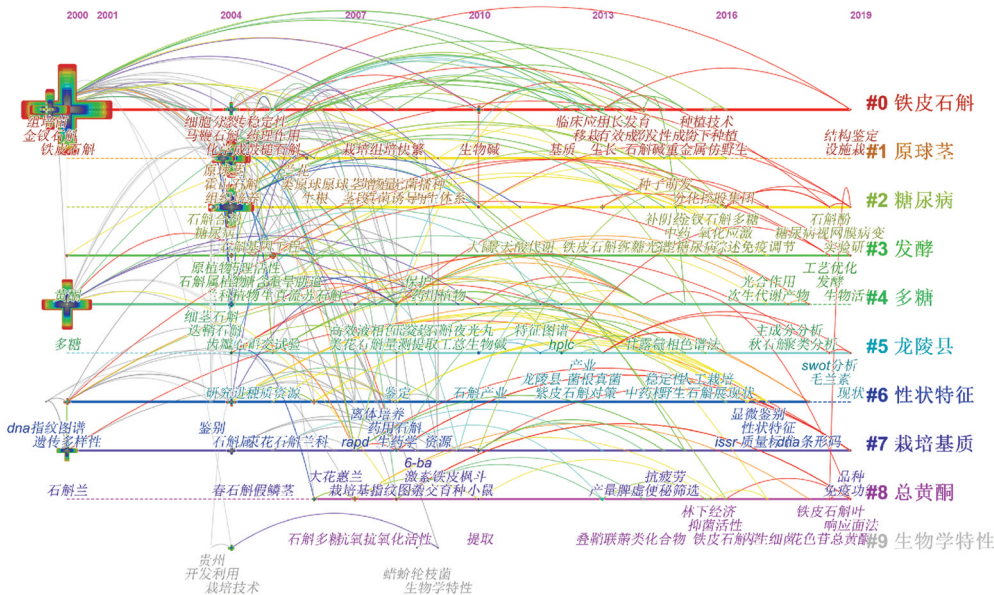


图 14 关键词知识图谱简化时序图

简化图谱的节点和连接线降低为 237 个和 462 条, 简化后的图谱聚类结果与原图谱相似, 可以看出聚类的结果重点围绕石斛的品种、成分、生长、症状治疗、遗传特性和数据挖掘等重点内容. 与图 13 相比, 图 14 中可以更清晰地看出每个聚类开始出现的时间、哪些年份聚类成果开始增多以及哪些年份关注度开始降低等内容, 图谱隐藏的信息更多。

2.4.3 热点分析

热点可以反映出一段时间内科研院所对该领域研究的动态变化过程. 如图 15 所示, 运用 Citespace 软件可以生成有关石斛研究的热点可视化图, 共有 26 个突变型关键词, 强度在 5.124 5 到 18.254 4 之间, 年度在 2000—2019 年。

从图 15 中可以看出, 石斛研究的起点较早、时间的跨度较长, 有关石斛的研究热点数量较多、强度较突出, 虽然部分核心研究热点未能展示出来, 但从侧面可以看出石斛领域研究和产业化发展历程. 首先, 有关石斛研究出现最早的热点为 DNA 指纹图谱和多糖, 主要研究石斛的特征和所含主要成分; 其次是石斛的品种和培育, 主要是对石斛进行更深层次的研究, 将石斛纳入中医药领域; 再次是石斛产业的需求分析, 将石斛进一步推广应用, 带动就业和经济发展; 最后着重就市场上研究和应用价值较高的铁皮石斛进行深度研究, 提取的多糖可以用于提高人体免疫力和降低血糖。

2.5 知识图谱构建结果

基于半监督训练的知识图谱构建可以弥补 Citespace 软件的缺点, 它可以根据石斛研究的特点构建针对性的实体关系抽取模型. 图 16 为基于半监督学习训练模型并在 Neo4j 平台中构建的知识图谱显示结果. 图中选取了作者、发表期刊、关键词、研究单位、2018 年 5 种实体进行训练, 可以清晰地看出 5 种实体之间的关系. 铁皮石斛、多糖、免疫调节等关键词

Top 26 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2000 - 2019
dna指纹图谱	2000	6.4639	2000	2007	
黄酮	2000	5.4198	2000	2006	
多糖	2000	6.4206	2000	2003	
原球茎	2000	10.001	2004	2009	
环萼石斛	2000	7.0027	2004	2011	
选种石斛	2000	5.4633	2004	2010	
诱导	2000	5.8332	2004	2011	
试育苗	2000	7.8105	2004	2009	
悬浮培养	2000	5.1487	2004	2007	
快速繁殖	2000	8.1223	2004	2008	
青石斛	2000	18.2544	2005	2011	
假鳞茎	2000	7.962	2005	2009	
石斛兰	2000	16.8285	2005	2009	
生根	2000	5.9909	2006	2014	
rapid	2000	6.0595	2006	2009	
流苏石斛	2000	5.1245	2006	2012	
类原球茎	2000	9.1509	2006	2012	
生物学特性	2000	5.4046	2006	2011	
培养基	2000	5.8861	2006	2008	
组织培养	2000	10.0748	2007	2009	
美花石斛	2000	5.4788	2007	2011	
药用石斛	2000	5.8671	2008	2010	
石斛产业	2000	6.6539	2010	2014	
齿瓣石斛	2000	7.7936	2010	2013	
叠鞘石斛	2000	7.7038	2013	2016	
铁皮石斛多糖	2000	5.3144	2016	2019	

图 15 石斛研究热点分析图

内容是 2018 年的主要研究对象, 每个研究对象对应的作者、期刊和研究单位可以在图中清晰地显示. 通过可视化结果分析、半监督训练模型来进行实体识别和数据打标, 构建的图谱可以较好地展示石斛领域的研究关系网.

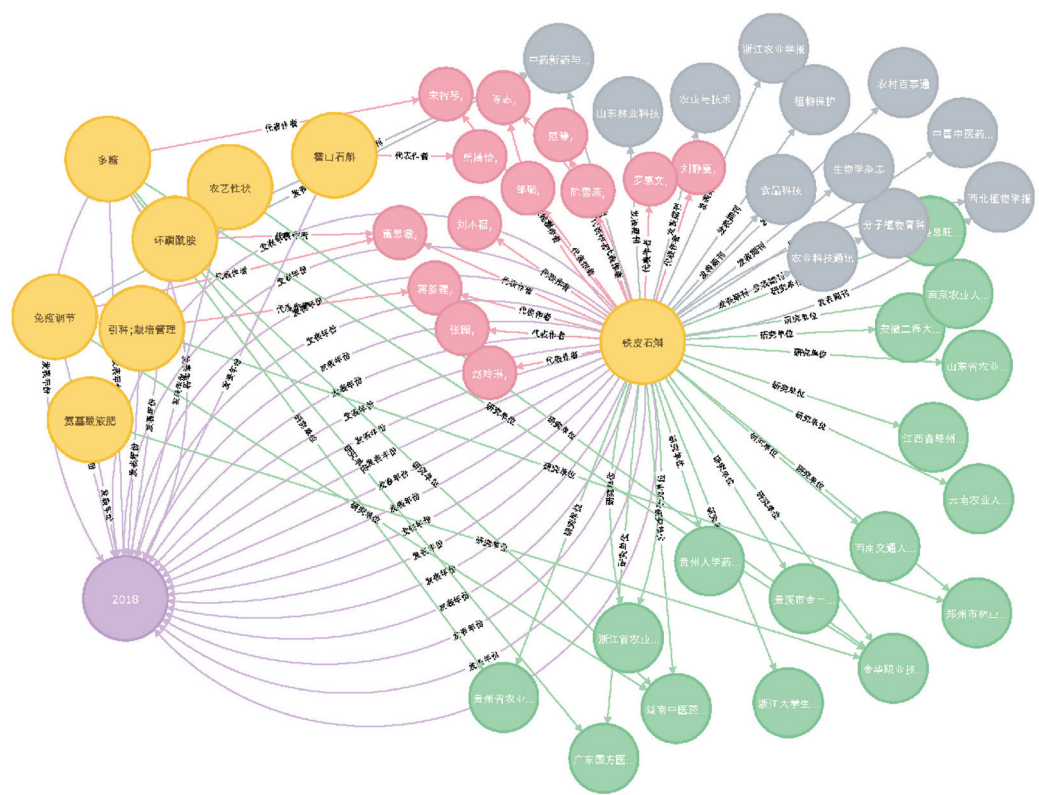


图 16 基于半监督训练的知识图谱构建图

3 总结

本文主要通过 Citespace 软件对知网有关石斛文献的摘要和关键词数据进行知识图谱可视化分析, 并基于半监督训练实现石斛领域的知识图谱构建. 结果表明, 有关石斛的研究呈现了“缓慢增长、快速增长和缓慢波动”的过程, 有关石斛的研究趋于成熟化并倾向于专业化; 研究文献来源分布广泛, 具有较强的专业性和地方性特色; 通过作者与机构的聚类图谱可以发现, 有关石斛研究的大部分作者和机构具有相互合作关系, 呈现较强关联性, 研究的科研单位更多与地理位置有关, 少数发达地区城市有关研究机构在石斛领域也有所涉及; 通过关键词和热点的聚类分析发现整个石斛产业的发展变化趋势, 分别从石斛的医疗效果、成分分析与提取、产业化推广、遗传分析和数据挖掘等聚类部分较好地看出了石斛研究的核心方向. 构建的知识图谱可以通过半监督学习对实体进行知识抽取和打标, 实现针对性的石斛研究知识图谱可视化. 由于训练和打标过程中需要人工选取实体和确定关系, 工作量较大, 下一步研究将尝试结合自然语言处理、神经网络和深度学习等算法实现实体抽取和图谱构建.

参考文献:

[1] 习近平. 关于《中共中央关于坚持和完善中国特色社会主义制度 推进国家治理体系和治理能力现代化若干重大问题的决定》的说明 [N]. 人民日报, 2019-11-06(4).

[2] 陈雁, 肖杨. 大别山地区百合科药用植物资源调查 [J]. 中国民族民间医药, 2008, 17(10): 1-4.

[3] 任刚, 陈优婷, 叶金宝, 等. 铁皮石斛叶的化学成分研究 [J]. 中草药, 2020, 51(14): 3637-3644.

[4] 袁青松, 王绘, 江维克, 等. 温度和品种对天麻抽蔓特性的影响 [J]. 中国中药杂志, 2020, 45(3): 485-490.

[5] 张宗源, 蒋咏梅, 章文贤. 组蛋白乙酰化对灵芝生长、灵芝多糖和灵芝酸生物合成的影响 [J]. 中国农业科学, 2020,

- 53(3): 632-641.
- [6] 江贤敏, 王正明, 潘利华, 等. 提取温度对霍山石斛多糖理化性质及肠道黏膜免疫活性的影响 [J]. 食品科学, 2017, 38(7): 176-181.
- [7] 张四杰, 钱正, 刘京晶, 等. 铁皮石斛花中花色成分抗氧化性和稳定性研究 [J]. 中国中药杂志, 2018, 43(10): 2025-2031.
- [8] 刘德锋, 彭琳, 郜鲁涛, 等. 基于地理信息系统(GIS)的云南省铁皮石斛种植适宜性区划 [J]. 江苏农业科学, 2017, 45(4): 227-230.
- [9] 刘烨宸, 李华昱. 领域知识图谱研究综述 [J]. 计算机系统应用, 2020, 29(6): 1-12.
- [10] 饶子昀, 张毅, 刘俊涛, 等. 应用知识图谱的推荐方法与系统 [J]. 自动化学报, 2021, 47(9): 2061-2077.
- [11] 王勇超, 罗胜文, 杨英宝, 等. 知识图谱可视化综述 [J]. 计算机辅助设计与图形学学报, 2019, 31(10): 1666-1676.
- [12] 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述 [J]. 计算机系统应用, 2019, 28(6): 1-12.
- [13] 李涛, 王次臣, 李华康. 知识图谱的发展与构建 [J]. 南京理工大学学报, 2017, 41(1): 22-34.
- [14] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述 [J]. 电子科技大学学报, 2016, 45(4): 589-606.
- [15] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53(3): 582-600.
- [16] 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.
- [17] YU T, LI J H, YU Q, et al. Knowledge Graph for TCM Health Preservation: Design, Construction, and Applications [J]. Artificial Intelligence in Medicine, 2017, 77: 48-52.
- [18] CHEN P H, LU Y, ZHENG V W, et al. KnowEdu: A System to Construct Knowledge Graph for Education [J]. IEEE Access, 2018, 6: 31553-31563.
- [19] 陈曦, 梅广, 张金金, 等. 融合知识图谱和协同过滤的学生成绩预测方法 [J]. 计算机应用, 2020, 40(2): 595-601.
- [20] 侯丽, 高阳, 刘路路. 我国农田重金属污染生态补偿研究现状与展望——基于 CiteSpace 知识图谱分析 [J]. 中国农业大学学报, 2020, 25(8): 132-143.
- [21] 杜师傅, 李娇, 杨芳绒, 等. 国内景观评价方法研究现状及趋势——基于 Cite Space 的文献计量分析 [J]. 西南大学学报(自然科学版), 2020, 42(7): 168-176.
- [22] 王友发, 周圆圆, 罗建强. 近 20 年智能制造研究热点与前沿挖掘 [J]. 计算机工程与应用, 2021, 57(6): 49-57.
- [23] 蔡兰英, 朱一邦. 赤水市金钗石斛产业发展现状、机遇与对策 [J]. 耕作与栽培, 2017(5): 59-60, 74.
- [24] 宋向文, 朱富成, 陈存武, 等. 六安市中药产业发展现状与存在问题研究 [J]. 皖西学院学报, 2019, 35(5): 20-23.
- [25] 王枫, 石红青. 我国铁皮石斛产业发展研究 [J]. 中国林业经济, 2019(3): 88-90, 135.
- [26] 杨燕婷, 陈荆菡, 许蕾, 等. 物联网背景下县域特色产业可持续性发展的路径探析——以安徽省霍山县石斛产业为例 [J]. 中国产经, 2020(4): 127-128.

责任编辑 汤振金