

DOI 10.19656/j.cnki.1002-2406.20231005

信息工程

基于海洋中药知识图谱的数据挖掘 与知识补全研究

洪海蓝¹, 王书怡¹, 李文林^{1,2}✉, 杨丽丽¹

(1. 南京中医药大学, 江苏 南京 210023; 2. 江苏省中医外用药开发与应用工程研究中心,
江苏 南京 210023)

【摘要】目的:探索海洋中药数据挖掘和知识补全技术,挖掘海洋中药文献的隐藏知识,为古籍文献的中医药知识挖掘和知识图谱质量提升提供借鉴。方法:针对领域知识图谱的特点提出了一种融合知识图谱与数据挖掘的领域知识图谱补全方法。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。结果:检索出补肾类海洋中药51味,相关效方123首,以此为数据源进行数据挖掘。发现补肾类海洋中药大多味咸,性温,归肝、肾两经。筛选出83首补肾类效方,发现补肾常用海洋中药有海马、海狗肾、海参、石首鱼和4类核心药组。将挖掘结果以三元组的形式补充到海洋中药知识图谱中,可有效完善补肾类效方的核心配伍关系。结论:以知识图谱为数据源进行数据挖掘,并将挖掘结果补充知识图谱,可以使知识图谱更为完整,相关思路与方法可为中医药领域数据挖掘、中医药知识图谱质量提升提供参考。

【关键词】知识图谱;知识补全;海洋中药;数据挖掘

【引用格式】

洪海蓝,王书怡,李文林,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.

HONG H L, WANG S Y, LI W L, et al. Research on data mining and knowledge completion based on Chinese marine medicine knowledge graph[J]. Information on TCM, 2023, 40(10):27-34.

知识图谱是一种用于表达知识的模型,主要由实体和实体之间的关系组成,其中的三元组(实体、关系、实体)可以通过一系列的操作来表达知识^[1-2]。中医药知识图谱是基于中医药知识的专业背景予以构建,主要包括病证、证候、病因、病机、效方配伍等内容^[3]。知识图谱在中医药领域不断发展,已经逐渐实现中医药智能搜索、智能问答、诊疗经验挖掘、辅助决策等应用。知识补全可提升知识图谱的质量^[3-4]。知识图谱补全是指在知识图谱中,通过预测出三元组中缺失的部分,使知识图谱变得更加完整,该任务可以分为实体补全

和关系补全两类^[5-7]。根据三元组中的实体和关系是否在已有的知识图谱中,知识图谱补全可以分为静态知识图谱补全和动态知识图谱补全两种^[8-9]。静态知识图谱补全是补全已知实体之间的隐含关系或补全存在于知识图谱中的实体属性;动态知识图谱补全是能够建立知识图谱与外界的关联,从而扩大知识图谱的实体集、关系集以及三元组集^[9-12]。当前大多数中医药知识图谱实现了图谱的构建,并没有对知识图谱进行补全完善。为此,针对领域知识图谱的特点提出了一种融合知识图谱与数据挖掘方法的领域知识图谱补全

基金项目:江苏省中医药科技发展计划(ZT202101);江苏省研究生科研创新计划(KYCX22_2052)

第一作者简介:洪海蓝(1996-),男,2020级软件工程专业硕士研究生,主要研究方向:中医药信息分析平台构建研究。

✉通信作者简介:李文林(1973-),女,研究员,主要研究方向:中医药信息分析平台构建研究。

方法。并以海洋中药知识图谱(CMMKG)为数据源,获取数据源进行数据挖掘,将挖掘结果补充到海洋中药知识图谱中,以提升海洋中药知识图谱的质量。

1 数据与方法

1.1 海洋中药知识图谱的挖掘技术线路

本研究以海洋中药知识图谱为数据源,按照智能检索、数据清洗、数据挖掘、数据分析等流程实现海洋中药挖掘和海洋中药知识图谱补全。具体流程见图1。

智能检索:首先根据输入的检索词对海洋中药知识图谱进行智能检索。检索词可以是功效词、主治词、海洋效方类型等,判断检索词选择对应的nGQL语句检索海洋效方。其次根据从海洋中药知识图谱中检索出的其他概念类型的数据进行补充,将检索到的所有数据作为数据源。

数据清洗:将从海洋中药知识图谱中检索到的数

据源进行清洗。如删除效方配伍中中药的单位属性(g、mL、适量、个、具)等;对异名同义的中药进行术语规范和其他类型术语规范;剔除用药仅为一味中药或一种提取物的海洋中药相关效方等。

数据挖掘:将筛选出来的最终数据利用关联规则算法,根据不同的挖掘需要,设置支持度、置信度进行数据挖掘。如找出具有某种功效或主治某种病症的海洋中药相关效方的药物配伍关系中隐含的规律,或者具有某种功效和治疗某种疾病的海洋中药的性味归经、营养成分。

数据分析:从中医药或者现代医学的角度对关联规则的挖掘结果进行分析,并将关联关系补充到海洋中药知识图谱库中,为海洋中药知识图谱提供更丰富的知识表示方式,形成更为完备的海洋中药知识图谱。

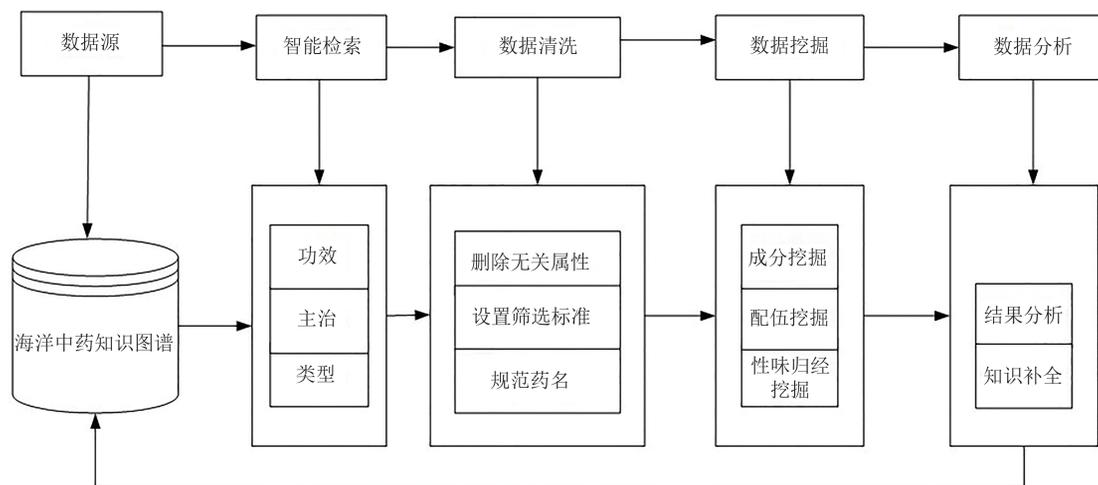


图1 海洋中药知识图谱的数据挖掘流程

1.2 数据来源

本研究以前期构建的海洋中药知识图谱为数据源,目前海洋中药知识图谱以海洋中药相关效方和海洋中药为核心,包含实体18 744个,关系30 664条^[13-14]。以补肾功效为例,以“壮阳补肾”“补肾壮阳”“暖肾壮阳”“补肾”“补肾壮腰”“补肾固精”“温肾”“补肾壮骨”“补肾生血”等40个补肾功效词语为检索词,检索补肾类海洋中药和海洋中药相关效方,并分析海洋中药的性味归经、营养成分、功效、主治和海洋中药相关效方的配伍、功效、主治。

2 海洋中药知识挖掘

知识图谱和关联规则是数据挖掘和人工智能领域的重要工具,它们的有效结合可以提供更丰富的知识表示,从而推断出更多的见解和理解^[4,15]。知识图谱可以用作关联规则挖掘的数据源,而关联规则算法挖掘出的关联关系可以补充知识图谱,并帮助

验证和修正已有的实体和关系。本研究将关联规则、聚类和知识图谱结合,在海洋中药知识图谱中从功效、主治、成分、配伍、性味归经等角度进行数据挖掘,并对挖掘结果进行分析和对海洋中药知识图谱的知识补全^[16]。

2.1 基于知识图谱的海洋中药数据挖掘

每一味海洋中药都具有性味归经、功效主治、化学成分等属性特征。不同中药在配伍应用时,其属性特征之间也形成了多维的配伍关系。为多角度揭示海洋中药配伍应用时隐含的配伍规律,本研究对药物属性特征之间的关联关系进行了分析挖掘。使用nGQL语句共检索到鲛鱼、乌贼鱼蛋、海狗肾、鲍鱼肉、对虾、海牛、海马等补肾类海洋中药51味。结果见图2。

检索51味补肾类海洋中药的主治经统计分析发现,能够治疗178种疾病或症状。结果见图3。

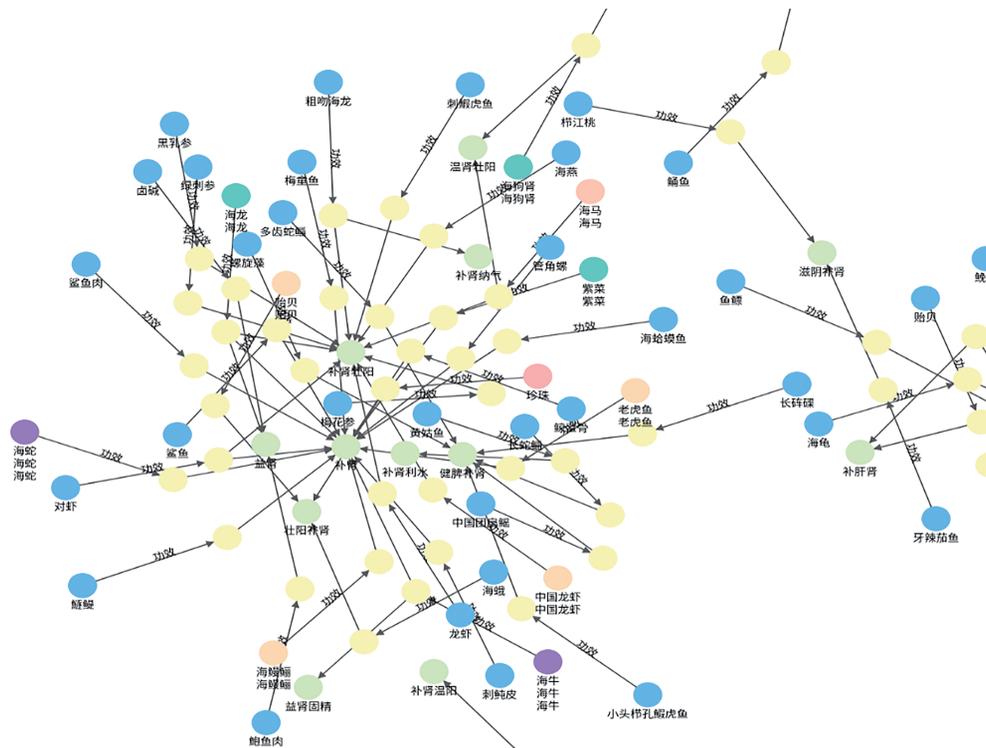


图2 补肾类海洋中药

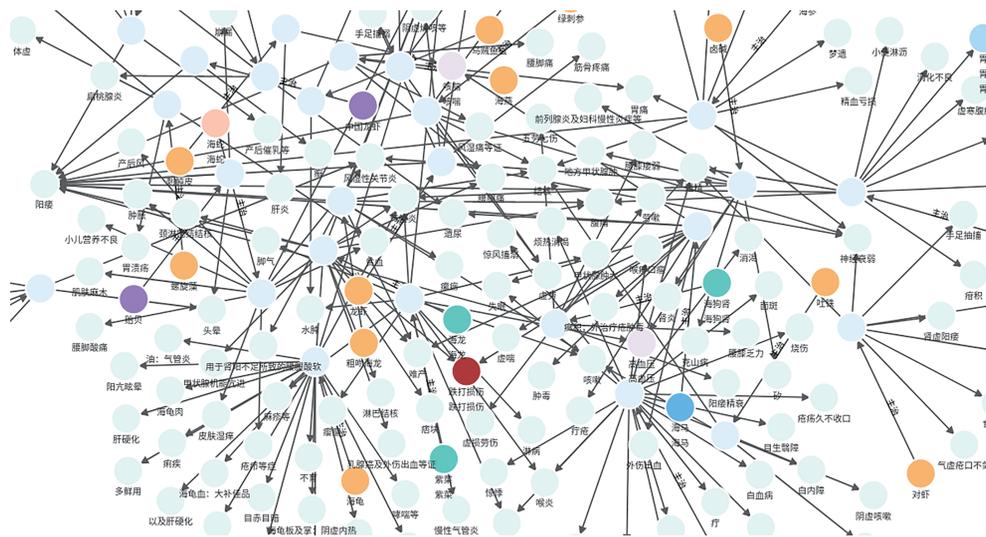


图3 补肾类海洋中药的主治

检索51味补肾类海洋中药的性味归经见图4。对这些性味归经进行数据清洗,一共获得21个性味归经词,见表1。对51味补肾类海洋中药的性味归经进行关联规则,设置支持度大于等于20%和置信度大于等于85%,结果见表2,将挖掘结果提交中医药领域专家评价,表明补肾类海洋中药以性温,味咸,归肝、肾两经为主。

2.2 基于知识图谱的海洋中药相关效方核心配伍挖掘

海洋中药知识图谱含有大量具有补肾作用的海洋中药相关效方,使用nGQL语句进行查询,从中检索到补肾类海洋中药相关效方123首,结果见图5。

表1 补肾类海洋中药的性味归经关联规则

序号	频繁项集	支持度/%	置信度/%
1	心 -> 咸	20.0	100.0
2	温 -> 咸	46.7	87.5
3	肝 -> 咸	40.0	92.3
4	无毒, 肝 -> 咸	23.3	87.5
5	温, 肝 -> 咸	26.7	88.9
6	肝, 肾 -> 咸	30.0	90.0
7	温, 肝 -> 肾	26.7	88.9
8	肝, 肾 -> 温	33.0	100.0
9	咸, 肝, 肾 -> 温	78.0	100.0

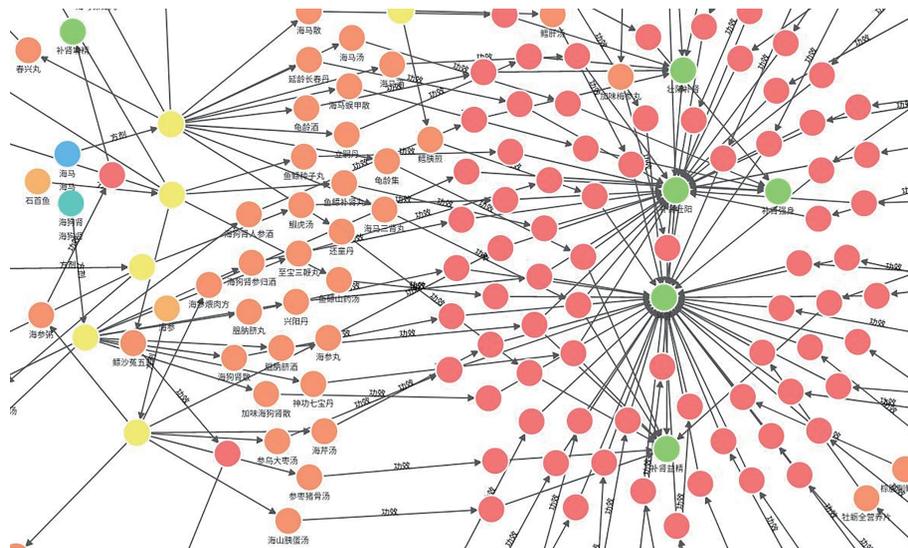


图6 补肾类海洋中药及相关效方

表2 相关效方的海洋中药表

海洋中药	数量	海洋中药相关效方
海马	13	龟龄酒, 延龄长春丹, 海马三肾丸, 海马保肾丸, 海马煎, 海马蜈蚣散, 还童丹, 春兴丸, 海马散, 海马汤, 龟龄集, 立嗣丹, 海马酒, 兴阳丹, 海狗肾散, 神功七宝丹, 脐脐酒, 温脐脐丸, 温脐脐酒, 《济生》温脐脐丸, 《圣济》温脐脐散, 海狗肾参归酒, 加味海狗肾散, 海狗肾散
海狗肾	11	
海参	8	参乌大枣汤, 参糖方, 参枣猪骨汤, 海山胰蛋汤, 海参煨肉方, 海参粥, 海参丸, 海芹汤
石首鱼	7	鱼鳔补肾丸, 黄鱼鳔胶丸, 鳔沙菟五煎, 鱼鳔山药汤, 鱼鳔散, 鱼鳔种子丸, 鱼鳔丸

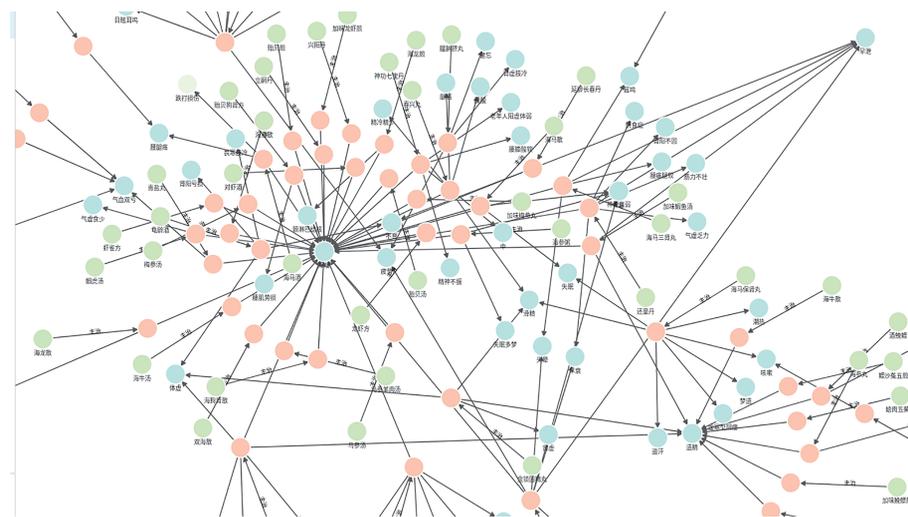


图7 补肾类海洋中药相关效方的主治

表3 中药名称规范化记录

序号	原始名称	规范后名称	序号	原始名称	规范后名称
1	酒牛膝/怀牛膝	牛膝	11	酒苁蓉/苁蓉	肉苁蓉
2	上沉香	沉香	12	巴戟	巴戟天
3	上肉桂	肉桂	13	毕澄茄	萆澄茄
4	大海马/海马对	海马	14	母丁香	丁香
5	大蛤蚧	蛤蚧	15	怀山药	山药
6	大小茴	小茴香/大茴香	16	生黑附子	黑附子
7	仙灵脾	淫羊藿	17	温脐脐	海狗肾
8	生黄芪	黄芪	18	酒当归/全当归	当归
9	大蛤蚧	蛤蚧	19	破故纸	补骨脂
10	杭菊/甘菊花	菊花	20	枸杞	枸杞子

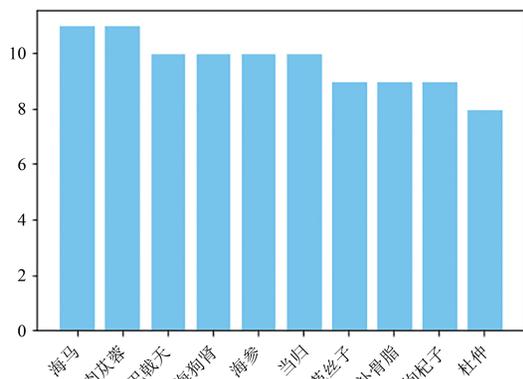


图8 补肾类海洋中药相关效方的前10味中药统计图

对83首补肾类海洋中药效方中出现频次 > 5 的21味中药进行层次聚类分析,并生成树状图,结果见图9。图中纵坐标代表中药之间的距离,横标则代表相对应的中药,中药聚为同一集合的距离越短,表明其关联性就越强。当纵坐标坐标距离取2.1时,得到

4组补肾类海洋中药相关效方的核心中药组合。组1:黄芪、山药、海狗肾;组2:当归、龟甲、牡蛎、海参、粮酒;组3:杜仲、枸杞子、菟丝子、巴戟天、补骨脂、牛膝;组4:肉苁蓉、黑附子、丁香、人参、鹿茸、海马、熟地黄。

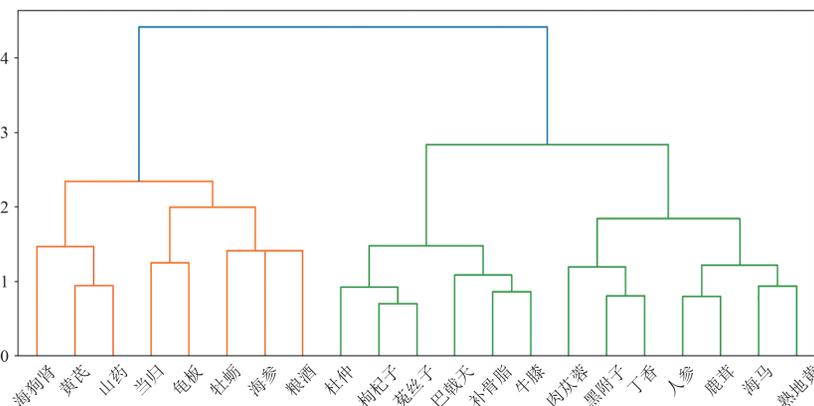


图9 药频次大于5的层次聚类分析树状图

设置置信度大于80%,支持度大于等于4.5%,对补肾类效方的配伍进行关联规则挖掘,结果显示:海洋

中药知识图谱中共隐含有50项补肾类关联项集,其中置信度为100%的15项关联项集见表4。

表4 补肾类效方的中药关联规则

序号	关联项集	支持度/%	置信度/%
1	丁香 -> 肉苁蓉	7.2	100
2	蛤蚧 -> 海马	6.0	100
3	丁香,巴戟天 -> 肉苁蓉	4.8	100
4	丁香,海马 -> 肉苁蓉	4.8	100
5	丁香,补骨脂 -> 肉苁蓉	4.8	100
6	丁香,黑附子 -> 肉苁蓉	4.8	100
7	巴戟天,枸杞子 -> 菟丝子	4.8	100
8	枸杞子,牛膝 -> 杜仲	4.8	100
9	淫羊藿,肉苁蓉 -> 海马	4.8	100
10	海马,淫羊藿 -> 肉苁蓉	4.8	100
11	熟地黄,肉苁蓉 -> 海马	4.8	100
12	海马,熟地黄 -> 肉苁蓉	4.8	100
13	肉苁蓉,蛤蚧 -> 海马	4.8	100
14	肉苁蓉,鹿茸 -> 海马	4.8	100
15	蛤蚧,鹿茸 -> 海马	4.8	100

将83首补肾类效方关联规则结果转换成三元组补充到海洋中药知识图谱中,2项关联关系以丁香->肉苁蓉为例,转换成1个三元组(丁香,补肾核心配伍,肉苁蓉)。3项关联关系以丁香,巴戟天->肉苁蓉为例,转换成3个三元组[(丁香,补肾核心配伍,巴戟天)(丁香,补肾核心配伍,肉苁蓉)(巴戟天,补肾核心配伍,肉苁蓉)]建立补肾核心配伍关系,形成更为完整的补肾类海洋中药核心配伍关系,结果见图10。结合层次聚类的分析结果,表明补肾类海洋中药相关效方是多以海马、丁香、肉苁蓉;海马、肉苁蓉、熟地黄、鹿茸和

杜仲、枸杞子、菟丝子、巴戟天、牛膝等核心药为中心进行药味的加减扩展。

3 小结

知识图谱作为一种能够使领域知识有效管理和应用的新技术,在中医药领域有着广阔的应用前景。当前中医药领域构建了规模不等的领域知识图谱,但大多数研究成果仅实现了图谱的构建,并没有进行知识图谱补全。而对知识图谱进行分析挖掘,知识补全,不仅可提升知识图谱的质量,也可为基于知识图谱形成的语义化数据源有效开展知识服务提供支撑。

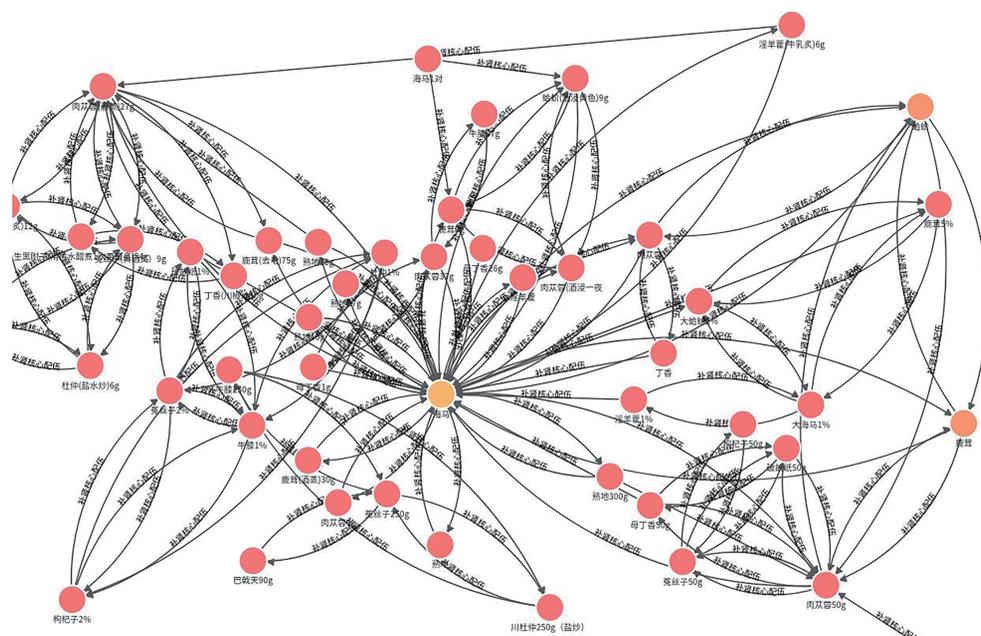


图10 肾壮阳类海洋中药相关效方的核心配伍

本研究以构建的海洋中药知识图谱为数据源,提出了一种针对领域知识图谱的特点提出了一种融合知识图谱与数据挖掘方法的领域知识图谱补全方法。并以补肾功效为例,通过从知识图谱中进行补肾类海洋中药及其效方的检索,对知识图谱中隐含的补肾类海洋中药及其构成的海洋中药相关效方进行了分析。基于这种思路,共从知识图谱中分析出具有补肾功效的海洋中药51味,以这些海洋中药为中心的补肾类海洋中药相关效方123首,进一步对这些海洋中药的成分、性味归经分析。通过对海洋中药相关效方中核心方药及其配伍关系的发现,挖掘补肾类效方83首、涉及相关中药一共275味。层次聚类和相关规则分析结果表明,83首补肾类海洋中药相关效方中存在4组核心药组合。将这些新发现的海洋中药相关效方、效方中的核心药组补充到海洋中药知识图谱中,有效提升了海洋中药知识图谱的完整性及知识图谱的质量,相关结果不仅为中医药领域的数据挖掘及知识图谱补全提供了参考,也为进一步促进海洋中药知识的传播与应用提供了更全面的知识平台。

【参考文献】

[1] 王昊奋,漆桂林,陈华钧. 知识图谱:方法、实践与应用[M]. 北京:电子工业出版社,2019:3-10.
[2] 田玲,张谨川,张晋豪,等. 知识图谱综述:表示、构建、推理与知识超图理论[J]. 计算机应用,2021,41(8):2161-2186.
[3] 张雨琪,李宗友,王映辉,等. 中医药知识图谱的构建与应用研究[J]. 世界中医药,2022,17(4):553-558.
[4] 孙郑煜,鄂海红,宋美娜,等. 基于大数据技术的医学知识图谱构

建方法[J]. 软件,2020,41(1):13-17.
[5] 丁建辉,贾维嘉. 知识图谱补全算法综述[J]. 信息技术,2018,12(1):56-62.
[6] 杨大伟,周刚,卢记仓,等. 基于知识表示学习的知识图谱补全研究综述[J]. 信息工程大学学报,2021,22(5):558-565.
[7] 吴国栋,刘涵伟,何章伟,等. 知识图谱补全技术研究综述[J]. 小型微型计算机系统,2023,44(3):471-482.
[8] JI S X, PAN S R, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications [J]. IEEE transactions on neural networks and learning systems, 2022, 33(2):494-514.
[9] 李凤英,范伟豪. 基于时序感知的动态知识图谱补全方法[J]. 计算机工程与应用,2022,58(15):202-209.
[10] 彭晏飞,张睿思,王瑞华,等. 少样本知识图谱补全技术研究[J]. 计算机科学与探索,2023,17(6):1268-1284.
[11] 王赫楠,杨昕悦,孙艳秋. 知识图谱关系检错和知识补全方法研究[J]. 计算机与数字工程,2022,50(7):1534-1539.
[12] 王子涵,邵明光,刘国军,等. 基于实体相似度信息的知识图谱补全算法[J]. 计算机应用,2018,38(11):3089-3093.
[13] 洪海蓝,李文林,杨涛,等. 基于知识图谱的海洋中药智能问答系统的设计与实现[J/OL]. 世界科学技术-中医药现代化:1-8 (2023-01-13) [2023-07-31]. <http://kns.cnki.net/kcms/detail/11.5699.r.20230112.1159.003.html>.
[14] 洪海蓝,李文林,杨涛,等. 海洋中药知识图谱构建及可视化研究[J]. 计算机时代,2023,41(1):50-54,59.
[15] ZHANG X L, MAO R, CAMBRIA E. A survey on syntactic processing techniques [J]. Artificial Intelligence Review, 2022, 56(6):5645-5728.
[16] 羊艳玲,李燕,钟昕好. 一种面向中医医案知识图谱的链路预测模型[J]. 中医药信息,2022,39(3):1-6,15.

(收稿日期:2023-05-12)

Research on Data Mining and Knowledge Completion Based on Chinese Marine Medicine Knowledge Graph

HONG Hailan¹, WANG Shuyi¹, LI Wenlin^{1,2✉}, YANG Lili¹

(1. *Nanjing University of Chinese Medicine, Nanjing 210023, China*; 2. *Jiangsu Provincial Engineering Research Center of TCM External Medication Development and Application, Nanjing 210023, China*)

【Abstract】 Objective: To explore the data mining and knowledge completion technology of marine traditional Chinese medicine, and to explore the hidden knowledge of marine traditional Chinese medicine literature, so as to provide reference for the knowledge mining and knowledge graph quality improvement of ancient Chinese medicine literature. Methods: According to the characteristics of domain knowledge graph, a domain knowledge graph completion method combining knowledge graph and data mining was proposed. The kidney – tonifying marine traditional Chinese medicine and related effective prescriptions were retrieved from the knowledge graph of marine traditional Chinese medicine. The association rules and clustering were used to analyze and cluster the correlation between the efficacy, indications, compatibility of effective prescriptions, and meridian tropism of marine traditional Chinese medicine, so as to complete the knowledge graph of marine traditional Chinese medicine. Results: 51 kinds of marine Chinese medicine for tonifying kidney and 123 related effective prescriptions were retrieved, which were used as data sources for data mining. It was found that most of the kidney – tonifying marine Chinese medicines are salty, sweet, warm, and belong to the two meridians of liver and kidney. A total of 83 kidney – tonifying effective prescriptions were screened out, and it was found that the commonly used marine Chinese medicines for kidney – tonifying were hippocampus, sea dog kidney, sea cucumber, *Sciaenops ocellatus* and 4 core drug groups. The mining results are supplemented to the knowledge graph of marine traditional Chinese medicine in the form of triples to effectively improve the core compatibility relationship of kidney – tonifying prescriptions. Conclusion: Data mining with knowledge graph as data source and supplementing knowledge graph with mining results can make knowledge graph more complete. Relevant ideas and methods can provide reference for data mining in the field of traditional Chinese medicine and quality improvement of knowledge graph of traditional Chinese medicine.

【Key words】 Knowledge graph; Knowledge completion; Marine Chinese medicine; Data mining