数字中医中药

DOI:10.13193/j.issn.1673-7717.2025.07.003

基于 Neo4j 与 GDS 算法的中医药治疗冠心病知识图谱构建与隐性知识规律发现方法研究

张迪[1],何丽云[1],王鑫[2],孙雨辰[3],文天才[2]
(1.中国中医科学院中医临床基础医学研究所,北京 100700;2.中国中医科学院中医药数据中心,北京 100700;
3.北京中医药大学东直门医院,北京 100700)

摘要:中医药知识体系内容丰富、类型多样、关系复杂,知识图谱(Knowledge graph,KG)作为知识发现的重要工具,其承载的信息量大,同时可以保留信息之间的关系,能够有效地描述和挖掘中医药领域的隐性知识。KG 自引入到中医药领域后发展迅速,已应用在名老中医经验挖掘、用药规律分析、智能问答等多个方面。但是,目前 KG 的数据挖掘方法较为单一,主要以查询检索为主,限制了中医药隐性知识的发现。基于文献构建的中医药治疗冠心病知识图谱,对 KG 常用的算法,如中心性、社团发现、相似度和路径查找等进行测试分析,寻找适合中医药数据挖掘的算法,为中医药领域知识图谱的隐性知识发现提供助力。

关键词:知识图谱;Neo4j;GDS;知识发现;冠心病
中图分类号:R259.414 **文献标志码**:A **文章编号**:1673-7717(2025)07-0010-07

Research on Construction of Knowledge Graph of Traditional Chinese Medicine Treatment of Coronary Heart Disease and The Discovery Method of Tacit Knowledge Rules Based on Neo4j and GDS Algorithm

ZHANG Di[1], HE Liyun[1], WANG Xin[2], SUN Yuchen[3], WEN Tiancai[2]
(1. Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Science, Beijing 100700, China;
2. Data Center of TCM, China Academy of Chinese Medical Science, Beijing 100700, China;
3. Dongzhimen Hospital, Beijing University of Chinese Medicine, Beijing 100700, China)

Abstract: The knowledge system of Traditional Chinese Medicine (TCM) is rich in contents, diverse in types and complex in relationships. Knowledge Graph (KG), as an important tool for knowledge discovery, carries a large amount of information while preserving the relationship among the information and can effectively describe and mine the tacit knowledge in TCM. KG has been developing rapidly since its introduction to the field of TCM, and it has been applied in the mining of the experience of famous and old TCM practitioners, prescription regularity analysis, intelligent Q&A and many other aspects. However, the current data mining method of KG is relatively single, mainly focusing on query retrieval, which restricts the discovery of tacit knowledge in TCM. In this paper, based on the KG of traditional Chinese medicine treatment for coronary heart disease constructed from literature, it tested and analyzed the algorithms commonly used in KG, such as centrality, community detection, similarity, pathfinding and so on to find the algorithms suitable for data mining in TCM and to provide assistance for the discovery of tacit knowledge in the KG in the field of TCM.

Keywords: knowledge graph; Neo4j; GDS; knowledge discovery; coronary heart disease

知识图谱(Knowledge graph,KG)是一种知识表示和存储方法,用于记录开放世界中的实体和概念以及实体之间的关系。KG 的概念最早是在 2012 年由 Google 提出并应用在搜索引擎领域,提高了搜索效率[1]。KG 可以有效融合多来源、多类型的数据,进行语义表示和知识关联。同时支持智能检索、社团发现、神经网络、深度学习等新方法,广泛应用于医学领域的知识发现,比如预测药物靶点、预测药物相互作用、识别药物不良事件和解释药物毒性等[2-5]。中医药知识体系内容丰富、类型多样、关系复杂,KG 作为知识应用的重要工具,其承载的信息量大,同时可以保留信息之间的关系,能够更加有效地描述和挖掘中医药领域的隐性知识,为中医药发展提供新的助力。KG 自 2015 年应用于中医药领域后发展迅速[6],已经用于中医药的经典古籍可视化、名老中医经验挖掘、用药规律分析、

基金项目:国家自然科学基金面上项目(82374624);北京市自然科学基金面上项目(7232306)
作者简介:张迪(1990-),男,河北邯郸人,博士,研究方向:中医临床评价。
通讯作者:文天才(1977-),男,四川绵阳人,教授、高级工程师,博士研究生导师,博士后导师,博士,研究方向:智慧医学与中医真实世界。E-mail:wtcsnake@163.com。

知识检索和智能问答、文献计量分析等多个方面[7-14]，在中医药领域受到越来越多的重视。

目前关于中医药 KG 的构建已经基本形成规范的流程，虽然不同研究之间存在一定的差异，但整体流程上已趋于一致，首先构建本体框架，然后依次进行数据来源、数据提取、数据规范等步骤，最终构建相关领域的知识图谱[15-19]。但是，在构建中医药 KG 之后，后续的隐性知识挖掘则较为单一，主要以知识检索查询为主。如在构建痴呆痰瘀互结证 KG 后，通过检索痰瘀互结证临床要素，发现痰瘀互结证的药物治疗和治法[20]；或构建中医药治疗痴呆 KG 后通过查询脾虚证，推理此证的临床用药和加减[21]。Neo4j 是目前最常用的中医药 KG 构建工具，图形数据科学库（Graph data science，GDS）是 Neo4j 自带的图形算法合集，提供了大量的常见图形算法，还包括机器学习，用于训练预测监督模型以解决图形问题，比如预测缺失的关系。本文基于 Neo4j 平台，通过对 GDS 的常用算法进行介绍，并基于中医药治疗冠心病文献数据构建的 KG，对 GDS 算法进行测试分析，寻找适合中医药数据挖掘的算法，以促进中医药领域 KG 的数据挖掘等方面的发展，为中医药知识规律发现提供帮助。

1 研究方法

1.1 数据来源 以主题（“中医药”OR“中药”OR“中医”）AND（“冠心病”OR“冠状动脉粥样硬化性心脏病”OR“胸痹”OR“胸痹心痛”）检索中国知网（China national knowledge infrastructure，CNKI）数据库中中医药治疗冠心病文献，检索时间为 2021 年全年。纳入标准：①干预方法为中药、中成药或院内制剂；②研究的疾病为冠心病；③临床试验；④提取信息完整。排除标准：①重复文献；②学位论文、成果及报纸类文献；③综述或系统评价文献；④实验类文献；⑤临床经验类文献。共检索 1180 篇文献，首先删除重复文献，其次阅读文献题目摘要和全文，根据纳排标准筛选后，纳入 346 篇文献。

1.2 知识图谱构建 通过北京交通大学研发的“人机协同表型标注系统”，对 346 篇文献进行标注，主要标注的实体类别为“方名”“中成药”“西医疾病”“中医疾病”“西药”“生化指标”“物理指标”等 24 种类别。构建实体间关系“包含”“治疗”“证候表现”“生化疗效评价”等 23 种关系。最终构建实体节点 911 个，实体关系 11 208 条的中医药治疗冠心病知识图谱。

1.3 GDS 算法 GDS 是一种图形算法，用于计算图形、节点或关系的特征，可以提供图中相关实体的中心性、排名或者节点之间的社区结构等。主要包括中心性算法（Centrality）、社团发现算法（Community detection）、相似度算法（Similarity）和路径查找算法（Path finding）。见表 1。

图是由节点和关系组成，节点（Nodes）一般是指中医药实体，比如方剂、中药和症状等。关系（Relationships）是节点与节点之间的联系，比如方剂 - 包含 - 中药，“包含”就是节点“方剂”和“中药”之间的关系。本文章使用的 Neo4j Desktop 版本为 1.5.7，Cypher 处理程序（Awesome procedures on cypher，APOC）插件版本为 5.3.0，GDS 插件版本为 2.3.0。

2 研究结果

2.1 中心性（Centrality） 中心性算法用于确定图中不同节点的重要性，根据不同的需求，在中心性算法下又分出了度中心性（Degree centrality，DC）、特征向量中心性（Eigenvector centrality，EC）、中介中心性（Betweenness centrality，BC）、网页排名（Page rank，PR）和文章排名（Article rank，AR）算法。

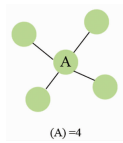
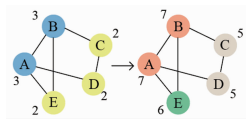
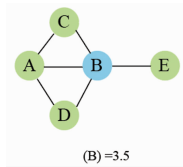
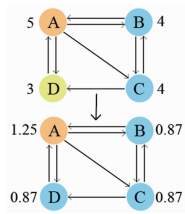
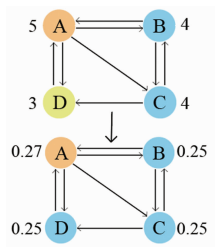
DC 可用于查找图中的流行节点，主要用于评估单个节点的重要性。在中医药知识图谱中，可以根据与单个节点相连的关系数量，确定并发现图中的核心中药、方剂、证候等。DC 在评估重要性时，将所有节点的重要性视为一致，但是在图中，不同节点的重要性存在差异，为了区分与之相连节点重要性对该节点的影响，因此引入了 EC。EC 的核心思想为，一个节点的中心性是相邻节点中心性的函数，也就是说，与节点相连的节点越重要，该节点就越重要。相较于 DC，能更全面地发现核心节点。BC 是检测节点对图中信息流影响程度的方法，通常用于寻找图中一部分到另一部分的桥梁节点。相较于前两种算法仅计算单一节点的中心性，BC 主要用于计算双节点或多节点之间的中心性，在中医药知识图谱中，可用于查找症状和疗效之间的核心处方，或症状和方剂之间的核心证候等。PR 是 Google 浏览器用来体现网页相关性和重要性的算法，被引进到知识图谱中。PR 与 EC 类似，都是依据相邻节点的重要性在评估该节点的重要性，但是特征向量中心性不能很好地处理有向图，并且中心性没有传递给传入边，例如证候诊断很重要，但是并非所有与诊断相连的节点都很重要，比如证候的治法对证候诊断影响就很小。PR 可以处理有向图，并且中心性也会影响相邻节点。AR 是 PR 的一个变体，PR 遵循的假设是，源自低度的节点比源自高度的节点具有更高的影响力，AR 通过降低每次迭代发送给相邻节点的分数来降低低度节点的影响。两种算法可以根据不同的临床需求单独和合并使用。中心性算法特征见表 2。在 Neo4j Desktop 软件对已经构建的知识图谱分别运行 5 种中心性算法，每种算法对 911 个节点的中心性均进行了计算，本文选取每种算法的前 10 个节点的算法得分进行对比展示，中心性算法结果见表 3。

2.2 社团发现（Community detection） 社团发现算法用于评估节点组是如何聚类或划分的，以及它们加强或分裂的趋势。

表 1 GDS 常用算法

算法	描述	算法类别
中心性	用于确定图形中不同节点的重要性	Degree centrality、Eigenvector centrality、Betweenness centrality、Page rank、Article rank 等
社团发现	用于评估节点组是如何聚类或划分的，以及它们加强或分裂的趋势	Triangle count、Local clustering coefficient、Weakly connected components、Label propagation、Louvain 等
相似度	根据节点的邻域或属性计算节点对的相似度	Node similarity、K - Nearest neighbors 等
路径查找	寻找两个或多个节点之间的路径或评估路径的可用性和质量	Shortest path、Single source shortest path、A* Shortest path 等

表 2 中心性算法特征

序号	算法	描述	图解
1	度中心性(Degree centrality,DC)	根据节点与其他相连节点的数量判断该节点的重要性	
2	特征向量中心性(Eigenvector centrality,EC)	根据与节点相连的其他节点的重要性判断该节点的重要性	
3	中介中心性(Betweenness centrality,BC)	根据节点占不同节点连线数量的比值判断该节点的重要性	
4	网页排名(Page rank,PR)	与 Eigenvector centrality 类似,但更擅长处理有向图,并传递中心性至相邻节点	
5	文章排名(Article rank,AR)	与 Page rank 类似,但降低了低度节点的影响	

社团的形成在中医药图谱中很常见,比如症状社团、中药社团等,识别图谱中的社团对于筛选核心方组和确定特定证候等均不可或缺。发现社团的一般性原则是,社团成员在群组内部的关系要多于其与群组外部节点的关系。目前常用的社团发现算法有三角形计数(Triangle count,TC)、局部聚类系数(Local clustering coefficient,LCC)、弱连通分支(Weakly connected components,WCC)、标签传播(Label propagation algorithm,LPA)和 Louvain 算法。

TC 用来确定图中每个节点的三角形数量。三角形是由 3 个节点组成的集合,其中每个节点与其他两个节点都有关系。TC 用来检测社团并测量这些社团的凝聚力,还可以用来确定图的稳定性,也被用来计算 LCC。LCC 用来度量一个图的聚类程度,一个节点的 LCC 也表明其邻节点之间互相连接的程度。某个节点的 LCC 越大,表明这个节点周围的节点之间的关系程度越高,这个群组聚集程度越高。WCC 是指在一个图中,某个子图的任意两点有边连接,并且该子图与剩余的任何节点都没有边连接。WCC 可以根据节点是否相连对图进行聚类,也可以对图的分布进行初步分析,判断图中是否有孤立的节点或者团体。LPA 是一个在图中快速发现社团的算法,在 LPA 算法中,节点的标签由它的直接邻居决定,其基本思想是节点的标签依赖其相邻节点的标签信息,影响程度由节点相似度决定,并通过传播迭代更新达到稳定。LPA 逻辑简单,时间复杂性低,在超大规模图中有优异的性能,适合做社团发现的基础算法。Louvain 是目前最常用的社团发现算法之一,使用模块度(Modularity)对社团进行划分。Modularity 常用来衡量节点分组质量的好坏,模块度越高说明所检测到的社团内部连接越紧密,而社团外部的连接则相对稀疏,即“内紧外松”特征,分组质量越好。Louvain 可用于大型图中发现社团,对于评估复杂网络的结构也很有帮助,尤其擅长发现多层级结构,比如可能会发现某个疾病下的辨证论治层级结构。社团发现算法特征见表 4。对知识图谱运行 Triangle Count 和 LCC 算法,分别得到每个节点的 Triangle Count 和 LCC 结果,选取结果前 5 的节点进行展示。见表 5。对知识图谱运行 WCC、LPA 和 Louvain 算法,其中 WCC 和 LPA 各产生社团 1 个,Louvain 产生社团 3 个。见表 6。

2.3 相似度(Similarity) 相似度算法根据节点的邻域或属性计算节点对的相似度,使用机器学习中常用的距离度量方法来计算相似度得分,可以用来发现中医药知识图谱中的证候 - 方剂辨证论治对,或者中药 - 中药药对等。

常用的相似度算法主要有两个,分别是节点相似度(Node similarity)和 K 近邻算法(K - Nearest neighbors,KNN)。Node Similarity 根据节点所连接的节点,对一组节点进行比较,如果两个节点共享许多相同的节点,则认为它们相似。Node similarity 根据杰卡德系数(Jaccard)或者重叠系数(Overlap)评估节点的相似性。Jaccard 定义为两个集合 A 和 B,那么两个集合的交集除以两个集合的并集即为杰卡德系数,Overlap 与 Jaccard 类似,为两个集合的交集除以两个集合中最小的集合。KNN 计算图中所有节点对的距离值,并在每个节点与其 k 个近邻之间创建新的关系,根据节点属性计算距离。比如一个需要确定中药 A 的剂量,参考附近 5 个方剂中中药 A 的剂量

表 3 中心性算法结果

排序	度中心性	分数	特征向量中心性	分数	中介中心性	分数	网页排名	分数	文章排名	分数
1	临床总有效率	570	冠状动脉粥样硬化	0.992	中医证候积分	6076	冠状动脉粥样硬化	49.35	冠状动脉粥样硬化	9.47
2	中医证候积分	506	不稳定性心绞痛	0.069	血府逐瘀汤	1603	稳定型心绞痛	23.89	稳定型心绞痛	3.93
3	LVEF	300	稳定型心绞痛	0.062	超声心动图指标	1488	心绞痛	19.26	心绞痛	2.92
4	不良反应	235	心绞痛	0.055	心电图指标	1437	不稳定性心绞痛	9.52	心力衰竭	1.90
5	胸痛	196	急性心肌梗死	0.049	血流动力学指标	927	心力衰竭	7.66	不稳定性心绞痛	1.67
6	胸闷	185	心力衰竭	0.038	炎症指标	923	PCI 术后	6.50	PCI 术后	1.59
7	心悸	180	PCI 术后	0.030	不良反应	710	焦虑	3.27	心律失常	0.96
8	超声心动图指标	177	瓜蒌薤白半夏汤	0.006	瓜蒌薤白半夏汤	666	心律失常	3.23	焦虑	0.90
9	心绞痛发作次数	171	涤痰汤	0.005	PCI 术后	634	抑郁障碍	2.35	血府逐瘀汤	0.80
10	TC	160	丹参饮	0.005	气虚血瘀证	614	胸痹心痛	1.93	抑郁障碍	0.70

注:LVEF. Left ventricular ejection fraction,左室射血分;PCI. Percutaneous coronary intervention,经皮冠状动脉介入术;TC. Total cholesterol,总胆固醇。

表 4 社团发现算法特征

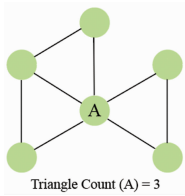
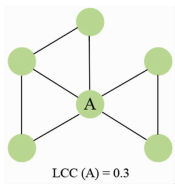
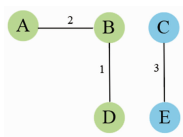
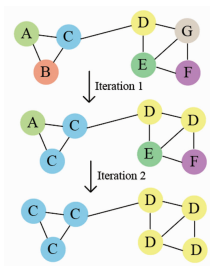
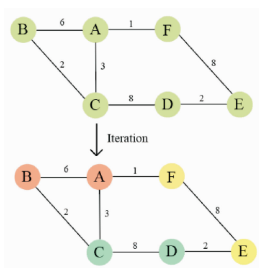
序号	算法	描述	图解
1	三角形计数(Triangle count,TC)	计算节点周围的三角形个数,检测社团凝聚力和图的稳定性,计算 LCC	
2	局部聚类系数(Local clustering coefficient,LCC)	判断图的聚类程度,节点 LCC 越大,其与周围节点聚集程度越高	
3	弱连通分支(Weakly connected components,WCC)	判断图中是否有孤立的节点或社团	
4	标签传播(Label propagation algorithm,LPA)	快速发现社团算法,逻辑简单,性能优异,适用于社团发现基础算法	
5	Louvain	发现社团算法,可用于评估复杂网络,尤擅于发现层级关系	

表 5 TC 和 LCC 结果

序号	节点	TC	节点	LCC
1	冠状动脉粥样硬化	2920	心悸	1
2	稳定型心绞痛	1739	斑块面积	1
3	心绞痛	1194	斑块性质	1
4	PCI 术后	1010	视物模糊	1
5	心力衰竭	997	肝郁气滞证	1

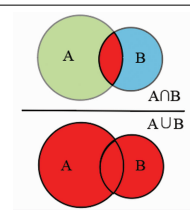
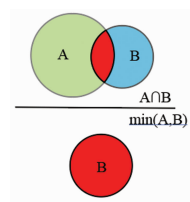
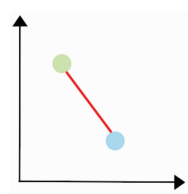
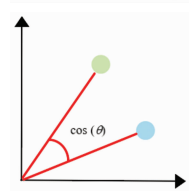
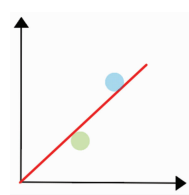
(k=5),最后选择出现最多的剂量定为中药 A 的剂量。KNN 使用的评估工具有欧氏距离(Euclidean)、余弦距离(Cosine)、皮尔逊(Pearson)、Jaccard 和 Overlap。Euclidean 计算两个节点之间的最短距离,是最常用和最多用的距离度量工具。Cosine 求得两个向量的夹角,并得出夹角的余弦值,夹角越小,余弦值越大,两个向量越相似。Pearson 用来衡量两个变量之间的线性关系,正值代表一个变量值增大另一个也跟着增大,反之亦是。Jaccard 和 Overlap 同上,距离评估工具特征见表 7。对知识图谱运行 Node similarity 中的 Jaccard 和 Overlap 算法,分别展示 5 对相似节点及其相似评分,见表 8;对知识图谱运行 KNN 中的 Euclidean、Cosine、Pearson、Jaccard 和 Overlap 算法,分别展示 5 个不同节点不同算法下的相似节点及其相似评分,见表 9。

表 6 WCC、LPA 和 Louvain 聚类结果

社团发现算法	聚类结果
WCC	血管内皮功能指标;3 个月;NO;ET1;真武汤;冠状动脉粥样硬化;心力衰竭;超声心动图指标;NT-proBNP;LVEDV 等 911 个节点
LPA	心绞痛;注射用丹参多酚酸盐;心电图指标;血流动力学指标;中医证候积分;全血黏度高切;np;血小板黏附率;HR;胸闷等 911 个节点
Louvain-1	心电图指标;心绞痛总有效率;硝酸甘油停减率;心电图疗效总有效率;3 个月;稳定型心绞痛;心绞痛程度;心绞痛积分;心痛宁免煎颗粒剂;气虚痰瘀证
Louvain-2	HR;不良反应;临床总有效率;冠状动脉粥样硬化;阿司匹林;不良反应发生率;8 周;早期心功能不全;宽胸气雾剂;TGFβ1
Louvain-3	心绞痛;力可生冲剂;银杏叶片;中医证候严重程度;血脂指标;心绞痛发作次数;心绞痛发作持续时间;硝酸甘油使用量;自拟益气通脉汤;曲美他嗪

注:TGFβ1. Transforming growth factor-β1,转化生长因子-β1;NO. Nitric oxide,一氧化氮;ET1. Endothelin-1,内皮素 1;NT-proBNP. N-terminal pro-brain natriuretic peptide,氨基末端脑钠肽前体;LVEDV. Left ventricular end-diastolic volume,左室舒张末期容积。

表 7 距离评估工具特征

序号	算法	描述	适用算法	图解
1	Jaccard	两个集合的交集比上两个集合的并集	Node similarity、KNN	
2	Overlap	两个集合的交集比上两个集合中最小的集合	Node similarity、KNN	
3	Euclidean	计算两个节点之间的最短距离	KNN	
4	Cosine	节点之间夹角越小,节点越相似	KNN	
5	Pearson	衡量两个节点之间的线性关系	KNN	

2.4 路径查找(Path finding) 路径查找算法用来寻找两个或多个节点之间的路径或评估路径的可用性和质量,在中医药知识图谱中可以筛选临床指标,探索疾病治疗等。常用的路径查找算法有最短路径(Shortest path)、单源最短路径(Single source

shortest path，SSSP）和 A 星最短路径（A* Shortest path，A*）。

表 8 Node Similarity 算法结果

算法	节点 1	类型 1	节点 2	类型 2	相似度
Jaccard	气滞血瘀证	证候	心血瘀阻证	证候	1
	益气化瘀胶囊	中成药	加味补阳还五汤	方名	1
	血脂康胶囊	中成药	降脂灵	方名	0.96
	冠心平片	中成药	颈动脉斑块	西医疾病	0.96
	心悸	中医疾病	安神定志汤	方名	0.95
Overlap	血府逐瘀汤	方名	痰浊证	证候	1
	舒心饮	方名	瘀血阻络证	证候	1
	舒心饮	方名	气滞血瘀证	证候	1
	丹蒌片	中成药	温胆汤	方名	0.97
	冠心平片	中成药	益心舒片	中成药	0.95

Shortest path 用于计算一对节点之间的最短路径，可以找到节点对之间的最优路线，也可以直接探索特定节点之间的连通情况。Shortest path 首先查找起始节点到与其直接连接的节点的距离，并移动到最近的节点，针对当前节点执行同一计算，计算一条累计距离，总是选择距离最小的路径前进，直至到达目标节点为止。SSSP 与 Shortest path 相似，计算从根节点到图中其他各节点最短路径，当需求解从固定起始节点到其他各节点的最优路径时，可采用此算法。A* 是在 Shortest path 基础上进行了改进，查找最短路径速度更快，通过估计到达目标节点剩余路径的启发式代价，实现路径查找过程。但是，如果在启发式方法中高估路径代价，就可能跳过实际较短的路径，导致结果不准确。路径查找算法特征见表 10。对知识图谱运行 Shortest path，分别展示 5 种临床问题。见表 11。对知识图谱运行 SSSP 算法，以不稳定性心绞痛为根节点，分别计算距离根节点 1、2、3、4 距离的节点。见表 12。

3 讨论

知识图谱作为知识应用的重要技术，能够详细地展示、挖掘实体及实体间的关系特征，更规范、高效地存储各类资料。中医药领域知识体系规模庞大、内容复杂、关系多样，传统的知识存储工具不能充分地展示中医知识体系[22-24]。知识图谱自问世以来，很快便引入到了中医药领域，并得到广泛的应用。随着中医药人工智能和大数据分析等新兴领域的不断发展，知识图谱将越来越受到重视。虽然知识图谱在中医药领域发展迅速，知识抽取、知识融合和知识存储等流程已基本完善，但是对于知识推理挖掘等方面仍停留在基础的查询方面，未能充分地对构建的知识图谱进行深入的知识探索[25-26]。Neo4j 作为最常用的知识图谱构建工具，其自身提供了多种常用的图数据挖掘方法，比如中心性、社团发现等，但是这些数据挖掘方法并未在中医药领域得到应用。此外，Neo4j 提供的图数据挖掘方法，主要面向的领域为金融、社交、企业等，鲜有中医药领域的应用，对于这些方法是否适用于中医药知识挖掘，仍然存在疑问。因此，需要将图数据挖掘方法引入到中医药领域，同时，对每种方法进行了解测试，明确其是否适用于中医药知识图谱以及使用的条件。本文通过对 Neo4j 已经成熟的 4 种图数据算法进行介绍，并在中医药治疗冠心病文献知识图谱中进行运算，以寻找适合中医药数据挖掘的算法，并探索其应用的条件，为中医药现代化发展作出贡献。

知识图谱中，每个节点之间通过关系连接，评估某个节点在图中的重要性时，单单使用频率等传统指标只能评估该自身的重要性，并不能体现其在图中的重要性。中心性算法将主要依据节点与其他节点的联系程度来评估该节点的重要性，充分发挥了图数据库的特点。同时，GDS 提供了多种中心性评估算法，可以从节点之间的联系，或者节点在图中流通的重要性等多个方面评估节点的重要性，为我们分析中医药知识图谱提供了最基础的解读。从算法特征来看，DC 是中心性中最基础的算法，主要描述节点在图中与其他节点的连接情况。EC、PR、AR 三者类似，除了判断与节点的连接情况，还通过其相邻节点在图中重要性来判断节点的重要性。PR 和 AR 较 EC 更擅长处理有向图，AR 较 PR 则改善了低度节点对节点重要性的影响。BC 更看重节点流通之中的影响力，某个节点出现在多个节点连接路径的次数越多，BC 值就越高。从中心性结果来看，PR 和 AR 的结果大致相同，两者与 EC 的结果部分相同。PR、AR、EC 三者的结果趋势与 DC、BC 不同，同时，DC 和 BC 两者之间也不相同。综上所述，在中医药知识图谱中心性分析中，可以将 DC 作为最基础的分析，初步了解节点在图中的重要性。PR、AR 和 EC 3 种算法类似，可以作为 DC 的进阶算法，综合分析节点的重要性，如果某节点在以上 3 个算法中都重要，说明其是图数据库中非常有影响力的节点。BC 更看重节点在图流通性之中的重要性，应与以上 4 种算法区分来看，主要用来探索图中作为桥梁的重要节点。

GDS 提供了多种社团发现算法，其中 TC 和 LCC 用来评估知识图谱的稳定性和聚类程度，值越高代表图越稳定和聚类程

表 9 K－Nearest neighbors 算法结果

算法	节点	节点 1 AA 抑制率	节点 2 ACEI	节点 3 ADP 抑制率	节点 4 AF	节点 5 AIP
Euclidean	节点	腹痛	安神复脉汤	PAF	气阴两虚证	气虚痰瘀证
	相似度	1	1	1	1	1
Cosine	节点	养心安神	安神定志汤	替格瑞洛	痰瘀互结证	心悸
	相似度	1	1	1	1	1
Pearson	节点	平板实验评分	益气化瘀汤	强心利水汤	痰瘀互结证	降糖通络方
	相似度	1	1	1	1	1
Jaccard	节点	血栓指标	益气复脉汤	疏肝解郁汤	痰瘀互结证	秽浊痰阻证
	相似度	1	1	1	1	1
Overlap	节点	Mb	CGRP	CCS 心绞痛分级	气阴两虚证	冠心平片
	相似度	1	1	1	1	1

注：AA. Arachidonic ccid，花生四烯酸；ACEI. Angiotensin converting enzyme inhibitors，血管紧张素转化酶抑制剂；ADP. Adenosine diphosphate，腺苷二磷酸；AF. Angina frequency，心绞痛发作情况；AIP. Plasma atherogenic index，血浆致动脉硬化指数；CCS. Canadian Cardiovascular Society，加拿大心血管学会；CGRP. Calcitonin gene－related peptide，降钙素基因相关肽；Mb. Myoglobin，肌红蛋白；PAF. Platelet activating factor，血小板活化因子。

表 10 路径查找算法特征

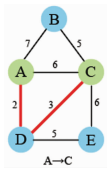
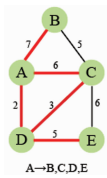
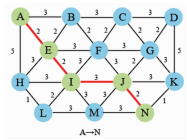
序号	算法	描述	图解
1	Shortest path	计算两个节点之间的最短距离	
2	Single source shortest path	计算一个节点到其他相邻节点的最短距离	
3	A * Shortest path	基于 Shortest path，通过启发式代价实现路径查找，算法更快，但是可能跳过实际最短路径	

表 11 Shortest path 结果

序号	开始节点	经过节点	结束节点
1	不稳定性心绞痛	血府逐瘀汤、氧化应激指标	eNOS
2	心力衰竭	PCI 术后、血管内皮功能指标	TXB2
3	PCI 术后	心律失常	PSQI
4	PCI 术后	补阳还五汤、丹参饮	METs
5	稳定型心绞痛	血府逐瘀汤	MLHFQ

注：eNOS. endothelial nitric oxide synthase，内皮型一氧化氮合酶；METs. Metabolic equivalents，代谢当量；MLHFQ. Minnesota living with heart failure questionnaire，明尼苏达心力衰竭生活质量调查；PSQI. Pittsburgh sleep quality index，匹兹堡睡眠质量指数；TXB2. Thromboxane B2，血栓素 B2。

表 12 Single source shortest path 结果

根节点	距离	节点	类型
不稳定性心绞痛	1	注射用益气复脉冻干	中成药
	1	1 个月	时间
	2	抑郁障碍	西医疾病
	2	气滞血瘀证	证候
	3	中药足浴	特色疗法
	4	化瘀祛痰	治法

度越高。WCC、LPA 和 Louvain 则是根据节点的相似度发现图中不同的社团，用来探索图结构下隐藏的规律。WCC 虽然称为聚类，但其判断的依据是节点之间是否有连线，因此更适合判断图中是否有孤立节点或社团。LPA 和 Louvain 则更适合发现图中的不同社团，前者可以作为社团发现的基础算法，后者作为补充，发现更复杂的社团或层级结构。从 TC 和 LCC 结果来看，中医药治疗冠心病文献知识图谱稳定性和聚类程度均较高。其中 LCC 的结果都是 1，考虑可能和构建的知识图谱有关，因图中每个节点都互相连接，因此可能构成了一个完全图，所以 LCC 均为 1。通过 3 种不同社团算法分析，共展示 5 种聚类结果。其中 WCC 和 LPA 各聚为 1 类，都包含了 911 个节点，即将图中所有节点聚为了 1 类，可能与完全图有关。Louvain 算法则将图聚成了不同的社团，每种聚类结果均包含疾病名称、治疗方法、生化指标和物理指标等，部分聚类结果尚包含中医证候、治疗时间、中医疗效评价和西药等。综上所述，TC 和 LCC 可以用来评估知识图谱的聚类程度，WCC 可以发现图中孤立的节点，从而找出构建图谱过程中遗漏的节点或节点间的关系。LPA 和 Louvain 是社团发现的主要算法，前者是基础，后者是进阶。需要注意的是，在运行算法时，需要考虑完全图的问题，如果 LCC 均为 1，即构建的是一个完全图，则 LPA 会将所有节点聚为一类。因此 LPA 不适用于完全图，Louvain 算法则不受此限制。

相似度算法根据两个节点共享相同节点的数量，来判断它们是否相似。如果两个节点共享相同数量的节点越多，它们也就越相似。相似度算法主要分为两大类，一种是 Node similarity，一种是 KNN，前者主要用于判断随机两个节点的相似度，后者则是选定一个节点，计算与其相邻节点的距离，从而判断相邻节点中与其相似的节点。从算法特征来看，Node similarity 可以直接运算，而 KNN 则需要提供节点属性值（Node property），一般使用权重作为节点属性值。但是中医药知识图谱构建过程中，节点的权重默认均为 1，所以运行 KNN 算法，需要先计算图中所有节点的权重，权重可以用 DC 的结果，也可以用节点出现的频率等，可以根据分析需求选择。相似度的取值主要基于节点之间的距离，目前 GDS 提供的距离评估工具主要有 5 种，Jaccard、Overlap、Euclidean、Cosine 和 Pearson。前两者仅适用于 Node property，KNN 则 5 种算法均支持。从结果来

表 13 GDS 算法应用总结

序号	算法	特征	应用	
1	中心性	Degree centrality	计算时将所有节点视为一致	计算知识图谱中重要节点，如中药、证候节点等
2		Eigenvector centrality	计算时区分相邻节点重要性对节点的影响	
3		Article rank	依据相邻节点重要性评估该节点重要性	
4		Page rank		
5		Betweenness centrality	检测节点在知识图谱中的流通影响	寻找两个节点间的桥梁节点，如证候和中药之间的治法
6	社团发现	Triangle count	计算每个节点的三角形数量	用来评估构建的知识图谱的稳定性
7		Local clustering coefficient	计算该节点相邻节点之间互相连接的程度	判断知识图谱的聚类程度
8		Weakly connected components	某子图任意两点有连接，但与剩余节点无连接	①通过聚类发现中药群或症状群等 ②判断是否有孤立节点或团体
9		Label propagation	通过传播迭代将节点标签相似的归为一类	快速聚类发现中药群或症状群等，适合基础挖掘
10		Louvain	依据模块度高低将社团紧密连接到一起	通过聚类发现中药群或症状群等，是最常用的社团发现方法
11	相似度	Node similarity	根据节点连接的节点，对一组节点比较，判读是否相似	可以用来探索知识图谱中药对、证候对等
12		K - Nearest neighbors	根据与该节点相邻的 k 个节点的属性计算相似度	可以用来探索方剂的效应指标、证候对、方剂对等
13	路径查找	Shortest path	计算节点对之间最短路径	可用于探索病 - 治疗 - 评价指标等关系
14		Single source shortest path	计算固定节点到其他节点的最短路径	用于探索疾病的合并疾病、证候、治法、治疗等
15		A * shortest path	通过启发式方法计算最短路径	由于需要经纬度属性，因此不适用于中医药知识图谱计算

看,Node similarity 用来初步探索图中相似的节点,可以发现药对、证候对、中成药对等。KNN 则更侧重于选定节点,判断与选定节点相似的节点,并根据相似度评分,进行排序。前者更适用于探索,后者则在确定目的节点后,进行精确探索。相似度算法不同于社团发现的多节点聚类,而是发现节点与节点之间的节点对,其作用类似于药对,但是可挖掘的隐性知识要比药对更丰富。

相似度算法只能计算两个节点之间的关系,如果要探索多个节点的关系,可以使用社团发现算法,但是社团发现的聚类结果是随机的,不能满足特定的知识发现。路径查找可以计算多个节点之间的关系,并可定义起始节点和结束节点,从而满足不同的临床需求。目前较为成熟的路径查找算法较少,主要分为两类,一类是限定两个节点,计算它们之间路过的其他节点;第二类是限定根节点(Start node),探索其他节点距离根节点的距离。第一类是指 Shortest path 和 A* Shortest path 算法,可以探索两个节点之间是否存在联系,及联系过程中涉及的节点。A* 算法需要提供经纬度属性(Latitude and longitude property),中医药知识图谱是一种抽象的知识体系,不存在经纬度属性,因此 A* 算法不适用于中医药领域的数据挖掘。第二类算法是指 SSSP,选择一个根节点,探索与其存在联系的节点,并计算与根节点的距离。从结果来看,Shortest path 可以用来探索两个节点之间的联系,比如疾病与指标之间,并且挖掘节点之间的其他相关节点。SSSP 则更适合探索兴趣节点的相关节点群,并根据距离,判断相邻节点对兴趣节点的影响程度。相似度算法弥补了社团发现和相似度算法的不足,拓宽了知识图谱隐性知识挖掘的范围。基于 Neo4j 的 4 种 GDS 算法具体应用总结见表 13。

目前中医药知识图谱构建已经较为成熟,已经构建多个名老中医、古籍和疾病等中医药知识图谱。但是,对于构建知识图谱后的应用和隐性知识挖掘方面却很欠缺。本文基于 Neo4j 中已经成熟的图数据库挖掘方法 GDS,对中医药知识图谱的应用进行探索,并以构建的中医药治疗冠心病文献知识图谱为例,对 GDS 算法进行运算和测试,探索适合中医药数据挖掘的算法。目前,尚未有系统介绍知识图谱应用的文章,也没有对数据挖掘算法适用性探索的文章,这是本文的特点和创新点。但是,由于纳入的文献较少,同时数据来源单一,主要是文献数据,因此节点间的关系和节点类型都较单一,GDS 算法的探索限于数据量不足,某些算法的探索不够充分。后续补充更多来源的数据,比如电子病历、名老中医医案和书籍等,更完整地寻找合适的算法,挖掘知识图谱隐形规律,拓展知识图谱的应用,为中医药的发展作出贡献。

参考文献

[1] 眷红英, 窦华溢, 贾玉祥, 等. 基于多来源文本的中文医学知识图谱的构建[J]. 郑州大学学报(理学版), 2020, 52(2): 45 – 51.
[2] AISOPOS F, PALIOURAS G. Comparing methods for drug – gene interaction prediction on the biomedical literature knowledge graph: performance versus explainability[J]. BMC Bioinformatics, 2023, 24(1): 272.
[3] FENG Z Y, WU X H, MA J L, et al. DKADE: a novel framework based on deep learning and knowledge graph for identifying adverse drug events and related medications[J]. Brief Bioinform, 2023, 24(4): 228.
[4] ZHANG G, GAO Z J, YAN C K, et al. KGANSynergy: knowledge graph attention network for drug synergy prediction[J]. Brief Bioinform, 2023, 24(3): 167.
[5] HAO Y, ROMANO J D, MOORE J H. Knowledge graph aids comprehensive explanation of drug and chemical toxicity[J]. CPT Pharmacometrics Syst Pharmacol, 2023, 12(8): 1072 – 1079.
[6] 于彤, 刘静, 贾李蓉, 等. 大型中医药知识图谱构建研究[J]. 中国数字医学, 2015, 10(3): 80 – 82.
[7] 汤少梁, 贡悦, 冯雨莉, 等.《伤寒论》治疫方用药规律分析及核心药对治疗 COVID – 19 的作用机制研究[J]. 中草药, 2023, 54(1): 192 – 209.
[8] 李新霞, 陆建峰, 孟红梅, 等. 本体在中医脾胃病辅助诊断中的应用[J]. 江南大学学报(自然科学版), 2010, 9(2): 151 – 155.
[9] 刘琴,陈锋,朱智涛,等. 基于 CiteSpace 的中医药治疗腰椎间盘突出症研究热点可视化分析[J]. 辽宁中医杂志,2023,50(11): 20 – 24,254 – 255.
[10] 陈程, 翟洁, 秦锦玉, 等. 基于中医药知识图谱的智能问答技术研究[J]. 中国新通信, 2018, 20(2): 204 – 207.
[11] 方芳, 徐天馥, 沈同平. 糖尿病医案本体库的构建及应用研究[J]. 中医学报, 2016, 31(11): 1680 – 1683.
[12] 林芝颖, 郭静, 孙建华. 针灸促进腹部术后胃肠功能康复的研究现状——CiteSpace 可视化图谱分析[J]. 辽宁中医杂志, 2023, 50(9): 215 – 219, 256.
[13] 车梦竹, 成泽东. 针灸治疗动脉粥样硬化的知识图谱可视化分析[J]. 辽宁中医杂志, 2023, 50(4): 160 – 164, 226.
[14] 王松, 胡孔法, 杨涛, 等. 周仲瑛辨治原发性支气管肺癌经验知识图谱的构建及应用[J]. 中医杂志, 2023, 64(2): 132 – 138.
[15] 袁东超, 王彩霞, 崔家鹏, 等. 基于本体的脾相关疾病核心病机探讨[J]. 辽宁中医杂志, 2017, 44(11): 2304 – 2306, 2462.
[16] 刘金垒, 惠小珊, 张振鹏, 等. 基于中医诊疗指南的冠心病知识图谱构建[J]. 中国实验方剂学杂志, 2023, 29(7): 208 – 215.
[17] 刘凡, 王明强, 李凌香, 等. 名老中医临床经验知识图谱构建方法探索[J]. 中华中医药杂志, 2021, 36(4): 2281 – 2285.
[18] 刘兴, 王琦, 李炜, 等. 痰湿体质知识图谱构建研究初探[J]. 北京中医药大学学报, 2022, 45(9): 949 – 955.
[19] 秦微, 崔家鹏, 王彩霞. 中医药知识本体构建思考[J]. 辽宁中医杂志, 2013, 40(8): 1568 – 1570.
[20] 戎菲, 佟旭, 胡镜清. 中医药知识图谱应用现状分析及痴呆痰瘀互结证知识图谱构建探索[J]. 世界科学技术 – 中医药现代化, 2021, 23(7): 2454 – 2460.
[21] 戎菲, 屈尧, 张逸雯, 等. 基于 Neo4j 构建知识图谱分析胡镜清论治痴呆的思路与特色[J]. 世界科学技术 – 中医药现代化, 2023, 25(3): 826 – 834.
[22] 曾子玲, 张华敏, 于彤, 等. 知识图谱及其关键技术在中医药领域的研究与应用综述[J]. 世界科学技术 – 中医药现代化, 2022, 24(2): 780 – 788.
[23] 何丽云, 李新龙, 刘岩, 等. 中医师辨证论治知识图谱构建的思路与方法[J]. 中医杂志, 2017, 58(19): 1650 – 1653.
[24] 王松, 李正钧, 杨涛, 等. 中医药知识图谱研究现状及发展趋势[J]. 南京中医药大学学报, 2022, 38(3): 272 – 278.
[25] 李芊芊,付兴,杨凤,等. 基于“病脉证并治”诊疗思维的《伤寒论》知识图谱构建与应用[J]. 世界科学技术 – 中医药现代化, 2022,24(9):3613 – 3621.
[26] 徐安迎, 胡孔法, 杨涛. 基于 Neo4j 的肺癌中医诊疗知识图谱构建研究[J]. 世界科学技术 – 中医药现代化, 2023, 25(4): 1456 – 1461.