南京中医药大学学报 2021 年 7 月第 37 卷第 4 期
J Nanjing Univ Tradit Chin Med Vol.37 No.4 Jul. 2021

基于知识元标引的《王旭高医案》逻辑数据及知识图谱探析

张泠杉[1]，王凤兰[1]，邢琛林[2]，王琨翎子[1]
（1.中国中医科学院中国医史文献研究所，北京 100700；2.北京邮电大学信息与通信工程学院，北京 100876）

摘要：目的 采用知识元理论的信息技术梳理、分析中医古籍内容，构建知识图谱探析隐含的逻辑关系以发现新知识。**方法** 以《王旭高医案》为例，在基于知识元理论与技术深度标引的基础上，首先利用 MS SQL Server 数据库将标引数据读取为逻辑数据并初步分析；其次以基于 neo4j 数据库构建的中医古籍知识图谱技术呈现出显性知识，同时探析其深层的逻辑推理关系，进一步发现隐性知识。**结果** 《王旭高医案》共有知识体 787 个，知识元 5 153 个，语义类型共有 1 149 个，语义关联共有 510 个。分析逻辑数据和机构化知识图谱可知，虚劳的知识元与语义关联最多，其中与肝脏、脾胃的语义关联最多。**结论** 《王旭高医案》整体来看特点在于重视对各类疾病证候表现的描述以及病因病机的分析，王氏诊治虚劳的经验较为丰富，从肝入手诊疗虚劳，尤其重视肝脾同病的病机。

关键词：王旭高医案；知识元标引；逻辑数据；知识图谱；古籍数字化
中图分类号：R249.2 **文献标志码：**A **文章编号：**1672－0482(2021)04－0592－05
DOI：10.14148/j.issn.1672-0482.2021.0592
引文格式：张泠杉，王凤兰，邢琛林，等.基于知识元标引的《王旭高医案》逻辑数据及知识图谱探析[J].南京中医药大学学报，2021，37(4)：592-596.

An Exploration of the Logical Data and Knowledge Graph of *Wang Xugao's Case Records* Based on Knowledge Element Indexing

***ZHANG Ling-shan*[1]，*WANG Feng-lan*[1]，*XING Chen-lin*[2]，*WANG Kun-ling-zi*[1]**
(1. *Institute of Literature in Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing*, 100700, *China*; 2. *School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing*, 100876, *China*)

ABSTRACT: OBJECTIVE To sort out and analyze the contents of Chinese medical classics by using the information technology of knowledge element theory, and to find new knowledge through exploring the implied logical relationships and constructing a knowledge graph. **METHODS** Through depth indexing of *Wang Xugao's Case Records* based on knowledge element theory and technology, we firstly read out indexing data as logical data and carried out preliminary analysis with MS SQL Server database. Secondly, we presented the explicit knowledge by using the knowledge graph technology of Chinese medical classics based on neo4j database, and explored the underlying logical relationships to further discover the implicit knowledge. **RESULTS** There were 787 knowledge bodies, 5 153 knowledge elements, 1 149 semantic types, and 510 semantic associations in *Wang Xugao's Case Records*. By analyzing the logical data and the structured knowledge graph, it can be seen that the knowledge elements and semantic associations of deficiency-consumption in the book are the most numerous, among which the semantic associations with liver, spleen and stomach are the most numerous. **CONCLUSION** It is concluded that *Wang Xugao's Case Records* is characterized by the emphasis on the description of syndromes and symptoms of various diseases and the analysis of relevant etiologies and pathogenesis. Besides, Wang had rich experience in the diagnosis and treatment of deficiency-consumption. His treatment of deficiency-consumption starts with the liver, with particular attention to the pathogenesis of the disease involving both the liver and spleen.

KEYWORDS: *Wang Xugao's Case Records*; knowledge element indexing; logical data; knowledge graph; digitalization of classics

知识元是表示、控制、管理和操作知识的基本单元，是为了解决以文献为单位的知识组织方式所包含的知识内容太少，而无法满足用户增长的知识需求而逐渐发展起来的[1]。将知识元理论运用于中医

收稿日期： 2021-05-10
基金项目： 国家重点研发计划(2019YFC1709200，2019YFC1709201)
第一作者： 张泠杉，女，博士研究生，E-mail：18945104401@163.com
通信作者： 王凤兰，女，研究员，博士生导师，主要从事中医古籍数字化的研究，E-mail：lfw733@126.com

古籍的研究中，不但可以大幅度提升传统中医古籍整理的效率，并且可以进一步细化古籍文献知识元。知识图谱是人工智能领域近年来迅速发展起来的新技术，可应用于中医数字辨证知识表示与推理研究[2]。将知识图谱技术应用于中医古籍的研究，不仅可以令显性知识一目了然地进行展示，还能发现知识元之间的隐性关系，更加便于学者对中医古籍知识进行运用与管理。

王泰林，字旭高，江苏无锡人，清代医家。初从事外科，后专致于内科杂病，临证用药甚为精当，为江浙地区著名医家。王氏著述计有六种：《西溪书屋夜话录》《医方歌诀串解》《环溪草堂医案》《医学入门》《选方约注》《伤寒一百一十三方歌诀》[3]。《王旭高医案》为其门人方耕霞搜集整理，分为四卷二十六门，其中又包含内、外、妇、儿各科病症，每门都缀有方氏按语。考察《王旭高医案》的研究现状发现，多数学者运用传统文献学方法对《王旭高医案》进行整理分析，未见有利用信息技术对其进行研究。基于此，本研究基于知识元理论将《王旭高医案》进行数字化加工，并运用构建知识图谱的技术与方法对其进行分析，以期能够发现新知识。

1　材料与方法

1.1　建立标引数据库

本研究使用的《王旭高医案》版本为清光绪二十四年（1898年）倚云吟馆本。文本由专业人员进行文献整理规范，并将规范后的文本导入中医古籍“病脉证并治”知识元标引系统。

1.2　知识体标引

知识体是知识系统中可以独立表达一个特定主题的不可再分解的知识单元[4]。首先将导入古籍标引平台的《王旭高医案》按照目录划分结构层次。《王旭高医案》共四卷，以病症归类各科医案，每卷下统数种疾病，如卷一为温邪、暑邪、痢疾、黄疸等病种；其次需精读文献，对其进行深度标引，根据具体内容来标引属于该内容的知识体。知识体共分为中医理论、诊法、病脉证并治、病症、本草、方剂、医案、预防调护、学术流派、针灸10种。

1.3　知识元标引

根据知识元理论的研究综合来看，多数学者认为知识元是不能分割的最小的独立知识单位[5]，中医古籍的研究中也遵循其基本概念。以卷一暑邪丁案为例，从姓氏起到结尾的方剂，可将其整体标引为“医案”这一知识体，并命名本知识体为“暑邪”，病案名即为“暑邪”知识元；“丁”为姓名知识元；“如体肥多湿之人，暑即寓于湿之内；劳心气虚之体，热即伏于气之中”为体质知识元；“暑乃郁蒸之热，湿为濡滞之邪。暑雨地湿，湿淫热郁，惟虚者受其邪，亦维素有湿热者感其气”为时令知识元；“暑雨地湿，湿淫热郁，惟虚者受其邪，亦维素有湿热者感其气。如体肥多湿之人，暑即寓于湿之内；劳心气虚之体，热即伏于气之中。于是气逆不达，三焦失宣”为病因病机知识元；“于是气逆不达，三焦失宣，身热不扬，小溲不利，头额独热，心胸痞闷，舌苔黄腻，底绛尖红，种种皆为湿遏热伏之征”为证候表现知识元；“舌苔黄腻，底绛尖红”为舌象知识元；“拟以栀豉上下宣泄之，鸡苏表里分消之，二陈从中以和之，芳香宣窍以达之，冀其三焦宣畅”为治则治法知识元；其所列方药整体皆标引为方剂知识元。

1.4　语义类型标引

语义类型宏观地表示了概念的语义场，藉此可精确地匹配不同来源词汇系统的概念，区别词型相似而含义不同的概念[6]。落实于具体操作上，语义类型提取需在知识元标引完成后方能进行。如卷一暑邪“丁案”中，将“于是气逆不达，三焦失宣，身热不扬，小溲不利，头额独热，心胸痞闷，舌苔黄腻，底绛尖红，种种皆为湿遏热伏之征”标引为证候表现知识元后，再将“气逆不达，三焦失宣”标引为病因病机语义类型；“湿遏热伏之征”标引为病证语义类型；“舌苔黄腻，底绛尖红”标引为舌象语义类型；“身热不扬，小溲不利，头额独热，心胸痞闷”标引为证候表现语义类型。

1.5　语义关联标引

语义关联是概念之间关系的抽象概括，它描述了概念间的内在联系，是构成知识体系的基本元素[7]。医案知识元包含的语义关联有辨治关系、证因关系、证象关系、方药关系等18种语义关联。仍以卷一暑邪“丁案”为例，勾选“气逆不达，三焦失宣”（病因病机），“湿遏热伏之征”（病证），“舌苔黄腻，底绛尖红”（舌象）；“身热不扬，小溲不利，头额独热，心胸痞闷”（证候表现）4个语义类型，点击“创建语义关联”，选择语义关联为“证因关系”，则语义关联生成。此标引系统中对于证因关系的规则定义为：病症↔证候表现↔脉象↔舌象↔病因病机。包含两种及以上元素的语义集合均属于“证因关系”语义关联。

1.6　数据分析

1.6.1　构建逻辑数据　通过标引软件系统的数据导出功能，导出《王旭高医案》的全部数据。存储在 MS SQL Server 数据库中的标引数据，通过数据导出程序，按其逻辑数据模型读取并保存为 Excel 格式的文件。此文件分为 5 个工作表，其中 4 个工作表分别为知识体记录、知识元记录、语义记录和语义关联记录，1 个工作表为各类知识体、知识元、语义和语义关联的统计信息。在此基础上对该古籍知识元、知识体、语义和语义关联的类型、数量、出现频率等进行统计分析。

1.6.2　构建知识图谱　在《王旭高医案》知识图谱构建过程中，数据处理流程为：标引数据（XML）→ json 半格式化数据→三元组数据→构建为知识图谱。

从标引平台中导出的数据是包含完整内容和层次关系的，通过‘＜＞’添加标签表达内容属性的 XML 格式数据，使用 xmltodict 库将 XML 字符串转换为嵌套字典类型，再使用 json.dumps() 将字典形式的数据转化为 json 字符串，处理得到的 json 数据中每行数据都对应某种标引模板，由此可以根据相应标引模板的标引逻辑将其解析为三元组形式。此外，为处理语义关联列表中的部分一对多、多对多的逻辑类型，引入虚拟节点的概念，将实体进行分组连接。最后，将解析后的实体列表与关系列表分别存储为 csv 文件，这里使用了 python 的 csv 包进行数据写入，再通过 neo4j 命令行 import 指令进行数据导入，从而生成知识图谱。流程图见图 1。

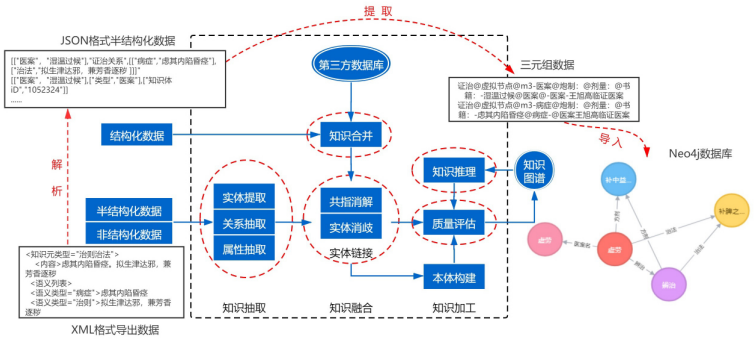

图 1　知识图谱构建流程示例

2　结果

2.1　知识元标引逻辑数据统计

《王旭高医案》共有知识体 787 个，全部为医案知识体。共有知识元 5 153 个，包括 25 种类型，其中前 11 种知识元占比较高，详见表 1。

表 1　知识元统计

知识元名称	数量	占比/%
证候表现	922	17.9
方剂	798	15.5
医案名	785	15.2
病因病机	558	10.8
姓名	530	10.3
治则治法	519	10.1
评按	205	4.0
脉象	202	3.9
舌象	130	2.5
预后	119	2.3
辨证	100	1.9
其他	285	5.5

语义类型共有 1 149 个，共涉及 9 种类型，详见表 2。

表 2　语义类型统计

语义类型	数量	占比/%
病因病机	421	36.6
证候表现	391	34.0
病证	119	10.4
脉象	95	8.3
舌象	42	3.7
治法	41	3.6
药物	25	2.2
方剂	13	1.1
治则	2	0.2

语义关联共有 510 个，全书涉及 5 种类型，详见表 3。

表 3　语义关联统计

语义关联	数量	占比/%
证因关系	403	79.0
证象关系	64	12.5
辨治关系	40	7.8
证治关系	2	0.4
方药关系	1	0.2

2.2　结构化知识图谱展示

下图为随机抓取的一张结构化知识图谱，不同的颜色代表不同种类的节点，具有连接作用。红色为医案名节点，粉色为虚拟节点。图 2 正中节点为知识体“湿温过候”，有 7 个下级知识体，分别为二诊到八诊。它们各自又有如证因关系、辨治关系、证象关系等多个不同的语义关联节点。

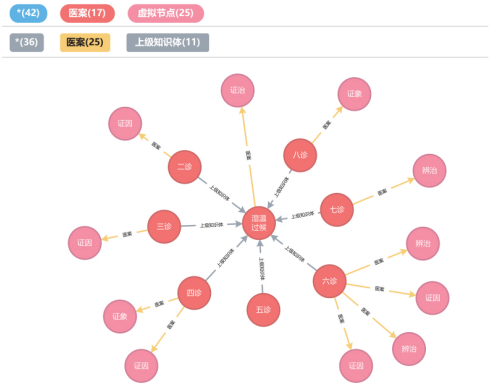

图 2　结构化知识图谱示例

3　讨论

3.1　逻辑数据分析

3.1.1　知识元逻辑数据分析　根据统计结果显示的前 11 种占比较高的知识元的类型可以看出，证候表现占比最高，说明《王旭高医案》记载病案的特点之一是重视对各类疾病证候表现的描述。通过统计发现医案名知识元中二诊及以上复诊率占全部医案名的 30.7%，说明本书载案有相当一部分是连续性的复诊，再结合预后知识元也有一定占比，说明本书的特点还在于重视疾病病程的发展及转归。医案类古籍的叙述方式众多，有详有略，对比其他篇幅较短、语出寥寥的医案类文献来说这一特点十分难得，使后学者更容易把握病程，从中吸取到临证经验。病因病机与治则治法占比较高且相当，说明本书长于对病理、治法等医理的论述。综合上述知识元的分析，可以看出《王旭高医案》对于各类疾病的证候表现包括舌象、脉象以及治法方药等治疗内容的描述充分而详细，并且重视预后，按语也颇丰。作为医案类古籍，其特点为医案各要素全面，病情记录详细，诊疗过程完备。

3.1.2　语义类型和语义关联逻辑数据分析　9 种语义类型中，病因病机与证候表现占比最高，而病因病机的占比高于证候表现，与知识元统计结果不同，表明《王旭高医案》着重于对病因病机关系进行阐述。对比脉象和舌象的占比发现脉象的语义类型占比远高于舌象，可见本书对脉象的关系性论述更为重视。治法、药物、方剂、治则的语义类型较少，说明本书较少有治法方药方面的关系性论述。

从语义关联的统计结果可以看出，证因关系占比最高，方药关系占比最低，表明《王旭高医案》着重于对证因关系的阐述，方药关系不是其阐述的重点（这一点从语义类型上也可以验证）。此外，证象关系、辨治关系占比也相对较高，通过和语义类型的统计结果相参发现，本书在病因病机、证候表现、病症、舌脉等对疾病辨证分析性的阐述内容所占篇幅最多，其次为治法类的内容，而方药的分析性内容着墨最少，这说明本书的特点详于对医理的阐述，而对于方药的理论分析并不丰富。

语义关联数量排在前 3 位的分别为证因关系、证象关系、辨治关系。《王旭高医案》共涉 26 门疾病，为进一步分析这 3 种语义关联在本书所涉疾病中的分布特点，笔者统计逻辑数据，取频次最高的 3 种疾病，则证因关系中前 3 位的疾病分别为虚劳、水肿、肝风痰火；证象关系为伏暑、虚劳、疟疾；辨治关系为伏暑、疟疾、虚劳。从上述统计不难发现，这 3 种语义关联在全书所有疾病中，于虚劳所占比例最多，由此可知本书对虚劳的医理阐述及诊疗过程尤为精详，反映了王旭高在虚劳的治疗上十分有心得，经验丰富。

3.2　知识图谱分析

为进一步探索王旭高诊治虚劳的特点，将《王旭高医案》进行深度标引后，构建结构化知识图谱，以可视化的形式对相关内容进行不同角度的分析。虚劳的语义关联知识图谱展现的是部分节点辐辏于虚劳医案节点的关系，相关虚拟节点有 28 个，绝大部分为证因关系，也包括少量辨治关系（图 3）。从证因关系的节点中可以看出，虚劳主要以脏腑病机为主，涉及肝、脾、肾等脏腑，说明王旭高在治疗虚劳时较为重视脏腑间的功能关系。所有节点中涉及肝脏的节点有 8 个，分别为“土衰木横”“肝虚”“怒动肝木侮脾，土益受戕”“脾胃不足，肝木亢逆”“肝强脾弱”“虽为肝旺，亦属脾衰”“肝木乘中”“木横则虫动”。其中与土虚木旺相关的节点有 6 个，明显多于其他病机，说明王旭高从肝脾两脏入手虚劳诊治，认为此二脏为虚劳病理机制的关键脏腑。

此外，图谱中涉及脾脏的节点有 11 个，涵盖了证因关系和辨治关系，分别为“补其脾”“补脾之气也”“土衰木横”“脾气弱”“怒动肝木侮脾，土益受戕”“脾胃不足，肝木亢逆”“中土式微”“肝强脾弱”“虽为

肝旺，亦属脾衰”“肝木乘中”“脾虚气滞”。去除肝脾关系的节点，单纯涉及脾脏的节点有 5 个，其中 2 个为辨治关系节点，可见王旭高在虚劳的诊治中认为脾虚也是本病的重要病机，并且在治疗上十分注重补脾。

根据以上分析，可以总结出王旭高辨治虚劳的特点在于重视肝脾两脏间的关系，同时充分注意到脾虚的病机，治疗中也十分强调补脾。

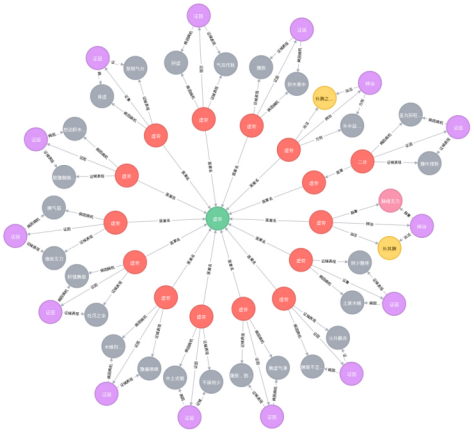

图 3　虚劳语义关联知识图谱

4　总结

通过对《王旭高医案》的深度标引，从知识元与语义关联逻辑数据的角度纵观《王旭高医案》全书，其特点为有关医案的各要素如证候表现、舌脉、病因病机、治法方药、预后等俱全，并且连续复诊率较高，能够清晰展示病情变迁过程与疗效。本书的撰写特点在于注重描述证候表现和丰富的病因病机分析，这是后学者研读医案文献以吸纳临证经验的重要抓手，更易于明确疾病诊断逻辑，同时也说明《王旭高医案》实为医案著作中的佳作。

通过对语义关联的逻辑数据和知识图谱的分析，不仅注意到王氏治疗虚劳有丰富的临床经验，更发现在对虚劳的认识上，王氏也有别于其他医家。虚劳最早见于张仲景《金匮要略·血痹虚劳病脉证并治篇》，仲景认为五脏俱虚，脾肾为本，故治疗上偏重补益脾肾，如小建中汤等，提出甘温扶阳，调补脾肾等治法，为后世医家治疗虚证奠定了理论基础[8]。唐代孙思邈重视从心肾两脏论治虚劳；宋代许叔微、明代张景岳和李中梓都提倡从脾肾论治；南宋严用和则强调从肾论治，并提出了“补脾不如补肾”的治疗原则；而金元时期李东垣、朱丹溪以及明代薛立斋和清代薛生白尤重以脾胃论治[9]，鲜有医家从肝论治虚劳。本研究充分体现了王旭高治疗虚劳长于从肝脾两脏入手，且病机尤重“土虚木旺”。王旭高素来以“治肝卅法”享誉杏林，其《西溪书屋夜话录》开篇即论：“病有肝气、肝风、肝火三者同出异名，其中侮脾乘胃，冲心犯肺，挟寒挟痰，本虚标实种种不同。”[10]并且在书中提出“治肝卅法”，其中有培土泄木法及泄肝和胃法，缓肝法也是针对“肝气甚而中气虚”者。可以看出王旭高在诊疗土木病机的疾病时，分脏腑从多角度进行辨治。这为虚劳的基础理论研究提供了新的切入点，也为临床治疗虚劳相关疾病提供了参考。

本研究利用知识元标引技术探索古籍文献内部，采用逻辑数据分析和结构化知识图谱可视化分析对《王旭高医案》进行文献挖掘与研究，根据诊疗思路及各类知识元之间的关联度，发现并选取书中的关联关系最为丰富的虚劳，总结出作者对虚劳的特色诊疗思想，为中医临床治疗虚劳提供了新的思路。此外，本次对《王旭高医案》进行的深度知识元标引和知识图谱构建，对于探索中医古籍数字化的研究也具有一定的参考价值。

参考文献：

[1] 索传军，盖双双.知识元的内涵、结构与描述模型研究[J].中国图书馆学报，2018，44(4)：54-72.
[2] 韦昌法，晏峻峰.从知识表示与推理方法探讨中医数字辨证发展[J].中华中医药杂志，2019，34(10)：4472-4473.
[3] 张洁瑜.王旭高医案研究[D].北京：北京中医药大学，2019.
[4] 严季澜，陈仁寿.中医文献学[M].北京：人民卫生出版社，2016：131-132.
[5] 高国伟，王亚杰，李永先.我国知识元研究综述[J].情报科学，2016，34(2)：161-165.
[6] 王勇，王凤兰.《黄帝内经》知识表示与标引研究[J].山东中医药大学学报，2020，44(5)：585-590.
[7] 朱玲，于彤，杨峰.基于关键动词的中医古籍概念实体间语义关系发现研究[J].中国数字医学，2016，11(5)：73-75.
[8] 王清华.《金匮·虚劳病》调补脾肾法浅析[J].湖北中医杂志，1998，20(6)：11-12.
[9] 戴梦，刘杰，盛昭园.中国历代医家对虚劳病的认识之回顾与分析[J].环球中医药，2021，14(4)：627-628.
[10] 王旭高.西溪书屋夜话录[M].北京：人民军医出版社，2012：1.

（编辑：叶亮）