DOI:10.12174/j.issn.2096-9600.2025.02.12

中医古籍文本理论内容的标引方法及思考*

王于静,陈晗婷,王维广△,刘晓峰,翟双庆
北京中医药大学,北京 100029

[摘 要]对同一段中医古籍文本理论内容进行标引,并分析不同标引方式的优点及不足,指出“病脉证并治平台”自上而下的标引方法与“经典知识平台”自下而上的标引方法在知识框架、形成知识图谱、标引自由度、标引模板等方面各有优势,应根据实际需求选择相应标引方法。并得出如有明确研究目标或需要构建完整知识框架时,应选择“病脉证并治平台”标引方法;如需求为建立相应知识图谱或需要进一步进行知识挖掘时,则选择“经典知识平台”标引方法更为恰当。

[关键词]中医理论;古籍文本;知识标引;知识图谱
[中图分类号]R2-03 [文献标识码]A [文章编号]2096-9600(2025)02-0044-04

Methods and Reflections on the Citation of Theoretical Contents of Ancient Chinese Medicine Texts

WANG Yujing, CHEN Hanting, WANG Weiguang△, LIU Xiaofeng, ZHAI Shuangqing
Beijing University of Chinese Medicine, Beijing 100029, China

Abstract By indexing theoretical contents in the same ancient Chinese medicine texts, the advantages and disadvantages of different citation methods were analyzed in the paper, the top-down citation method for "treatment platform of disease, pulse and syndrome" and bottom-up citation approach for "classic knowledge platform" have their own advantages in knowledge frameworks, forming knowledge graphs, degrees of freedom for citation and citation templates, the corresponding citation method should be chosen according to the practical need. Citation method for "treatment platform of disease, pulse and syndrome" could be selected if there are clear research objectives or the complete knowledge frameworks need to be constructed; citation approach for "classic knowledge platform" are more appropriate if the need is to establish the corresponding knowledge graphs or further knowledge mining is needed.

Keywords TCM theory; ancient texts; citations; knowledge map

中医理论是中医学的基础,包括中医哲学基础和思维方式等[1],其实质是借用哲学框架将经验理论化。中医理论研究即是将这些从现实中归纳出来的理论进行判断、推理等[2]。中医理论目前的研究方法除了较为传统的思辨研究方法外,还有知识考古学、认知语言学、文化人类学和复杂性科学等,其研究目的主要为归真和创新,创新又包括对传统理论的再认识和对已有理论的再发展,而这些研究均需要以中医古籍文本为依据[3]。

中医古籍文本是中医研究的基础,随着现代科技的进步,中医古籍文本的整理研究逐渐由传统手工整理方式向数字化资源整理和建设方向转变。标引即是中医古籍文本数字化中重要的一个步骤,在中医药领域标引主要分为文献组织层次和知识组织层次[4]。文献组织层次的标引即标引信息资源的外部特征,如对文献标题的标引[5],目前文献组织标引的中心也逐渐从单纯实体标引转移到对实体间语义关系的抽取[6];知识组织层次的标引则是对古籍内部知识单元内容进行标引,早在2000年中国中医科学研究院就已开始研发中医文献临证平台,柳长华教授在研究中医古籍数字化过程中提出了基于知识元的中医古籍计算机知识表示方法,为文献整理数字化提出了一种新方案[7]。本研究主要对知识组织层面的标引进行讨论。

知识组织层面的标引使古籍文本更容易被理解,有利于推动中医古籍文本数字化信息化,提高检索效率,为中医理论研究提供数据支撑,为深层研究提供可能性。中医古籍文本数字信息化的目标实际上就是建立中医古籍文本数据库,数据库需要具备对大量文本信息的存贮、检索、考证、推理等功能[3]。计算机无法准确识别长段文本的主要内容,而标引则赋予文本知识检索标识,将自然语言转化为受控制语言,有利于计算机的识别,进而实现对知识的检索,推动信息服务转向知识服务[8]。数据库的建立有助于知识图谱的构建,有利于进一步进行知识挖掘、发现及智能检索[9]。

目前中医古籍文本理论部分知识组织层面的标引主要存在两种不同形式的标引方法,即自上

而下和自下而上。自上而下的标引方法是一种从整体逐级向下拆分的标引方式，其主要目标是形成知识框架；自下而上的标引方法则是一种从部分到整体逐层归纳的方法，其目的主要是尽可能保留古籍文本所有内容并从中提取有用信息。这两种标引方法在标引中医古籍文本理论部分时各有优缺点。本研究以《脾胃论》[10]中“今之酒病者，往往服酒癥丸，大热之药下之，又有用牵牛、大黄下之者，是无形元气受病，反下有形阴血，乖误甚矣。酒性大热，已伤元气，而复重泻之，况亦损肾水，真阴及有形阴血俱为不足，如此则阴血愈虚，真水愈弱，阳毒之热大旺，反增其阴火，是以元气消耗折人长命；不然，则虚损之病成矣。”为例，分别介绍这两种标引方法，并分析其优点与不足。

1　自上而下的标引方法

1.1　自上而下标引方法的思路与原则　“中医古籍‘病脉证并治’知识元标引系统”（下简称“病脉证并治平台”）的标引方法是自上而下标引方法的代表平台之一。此系统是在中国中医科学院研发的“中医药古文献知识库”（http://www.zywx.org.cn）标引平台的基础上发展而来，在原平台基础上增加了中医理论、临床诊疗、预防调护、学术流派等标引模块。这些模板由古籍整理专家以及中医各学科专家从知识元层面对古籍文献如《备急千金要方》[11]进行分析而确立与改进的。现共有10个知识体模板，包括中医理论、病脉证并治、病证、本草、方剂、医案、预防调护、学术流派、诊法和针灸，更符合中医理论的内在逻辑结构。

此平台采用的标引方法是基于柳长华教授创立的基于知识元的中医古籍计算机表示方法的知识元标引方法。知识元标引需要对文献中具有检索价值的概念进行全面深度标引[8]，其标引结构主要划分为知识体、知识元、语义3个层次。知识体是由1个以上的知识元构成的知识集合；知识元是知识系统中可以表达一个完整概念且不可再分解的最小知识单元，由多个语义构成，表达1项相对完整的知识；语义则是构成知识元的基本要素，即通常意义上所谓的概念[8]，其形式可为语词、词组或短语；当单独语义不能表达完整含义时，则需建立语义关联[11]。此标引是在模板的基础上进行标引，其方法按照“选择知识体模板”→“解析知识元”→“解析语义”→“建立语义关联”的顺序进行。面对一本古籍，首先要对这本古籍进行书目标引，在章节内容中将文献划分为不同主题的段落，依据段落内容选择一个最合适的知识体模板对其进行标引，根据知识体模板结合内容确定知识元，并对每个知识元中具有完备含义的词、词组或短语结合相应的语义模板标引语义，再选择相应模板建立语义关联。此标引方法可认为是框架式知识表示方法与语义网络知识表示方法的结合，其标引过程从整体再到部分，是一个拆分的过程，即自上而下的过程。

1.2　自上而下标引方法的应用举例　以上文提到的《脾胃论》中的段落为例进行标引。这一段可划分为一个知识体，虽有中医理论相关内容，但总体而言更符合“病证”模板，因此将这一段标引为知识体“酒病（病证）”。其标引得到的知识结构见图1，其中第一层为知识体，第二层为知识元，第三层为语义。而误治、预后等知识元提取的语义不能表达完整含义，需要建立语义关联。

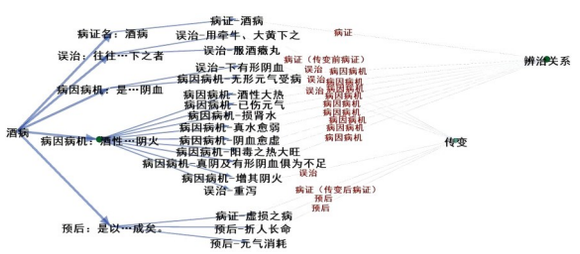

图1　标引知识结构

按照上述标引方法，对整本古籍进行标引，可得到一个完整的知识框架，即书名下包含各目录知识体，目录知识体下包含多个知识体，知识体下又可包含知识体以及多个知识元，知识元之下又细标为多个语义，语义之间可互相建立联系。此框架结构清晰明了，方便快速提取书籍信息，语义的标引也为检索提供了便利。

2　自下而上的标引方法

2.1　自下而上标引方法的思路与原则　“中医经典知识挖掘与传播平台”（以下简称“经典知识平台”）提供的工具是自下而上标引方法的代表之一。此平台由北京中医药大学翟双庆教授课题组与电子科技大学联合建设完成，其主要功能是对上传至平台的中医古籍进行主题批注、建立批注间关系以及可视化，这些功能为建立中医经典理论体系、挖掘中医理论隐性知识提供支撑。其标引方法采用“实体-关系-实体”方式，对非结构化文字材料进行处理。在此方式中，“关系”多来自文本中的动词以及上下文联系。其关系最初由业内专家进行确定，在标引过程中补充新增关系，经过一段时间的标引后，最终确定为59种关系，这些关系列于专为此平台标引方法编写的《编码手册》中，用以确保标引具有一定规范性。

“实体”则多为文本中的名词。对“实体”的标引借鉴扎根理论，称为编码过程，编码根据其所含内容分为不同层级。一级编码为初始编码，用以

概括一个特定的字段，多以原文提取中医名词的方式实现，需要逐字逐句进行标引；具有共同主题的初级编码归类为二级编码，其共同主题即为二级编码名称；以此原则归类形成三级编码、四级编码等更高级编码，最终归类为不可再行归类的中医学核心概念。编码与编码之间还需依据《编码手册》建立合理关系，以构建成为一个互相联系的知识图谱。此标引方法依据从部分到整体的标引顺序，从初始原文逐级提炼归纳出核心含义，是一个从下而上的归纳方法。

2.2 自下而上标引方法的应用举例 以上文提到的《脾胃论》中的段落为例，应用此标引方式对其进行标引。首先，逐字逐句对全文进行一级编码，见表1。其中编码原文为原文内容，而编码内容则是与编码原文相对应的一级编码。其次，依据《编码手册》所列关系在一级编码之间建立关系，见表2。其中批注关系名称反应两编码之间的关系，当批注关系为名词时，表示为“批注1内容”是“批注2内容”的某关系，当批注关系为动词时，则表示为“批注1内容”某关系“批注2内容”。通过以上编码，最终形成的知识图谱包含了整段文字基本内容，具有丰富内涵，依据此图谱中各节点和关系，可以复述原文含义。见图2。

表1 初级编码举例

序号	编码原文	编码内容
1	酒	酒病
2	酒癥丸	酒癥丸
3	大热之药	大热之药
……	……	……

表2 关系建立举例

序号	批注关系名称	批注1内容	关系名称	关系箭头	批注2内容
1	治疗禁忌	酒癥丸	单向关系	→	酒病
2	治疗禁忌	牵牛	单向关系	→	酒病
3	治疗禁忌	大黄	单向关系	→	酒病
……	……	……	……	……	……

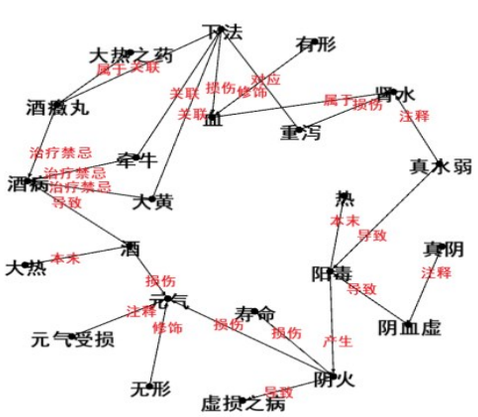

图2 编码知识图谱

3 两种标引方法比较

两种标引方法各有其优缺点，现将从知识框架、知识图谱、标引自由度、标引模板进行分析。

3.1 知识框架 从知识框架分析，“病脉证并治平台”的框架构成更为清楚。如图1所示，“病脉证并治平台”标引方法得到的知识框架从书名逐级分级为目录知识体、知识体、知识元、语义，成树状结构。以“酒病（病证）”为知识体依次分为病因病机、误治、预后等，清楚地将这段文字的结构展现出来，每一层框架都较清晰，方便研究者快速提取古籍内容。而“经典知识平台”形成的知识框架理论上接近“病脉证并治平台”，但其更侧重于将所提取的初级编码全部连接形成完整知识网络，框架体系相对不清晰，尤其在编码中医理论时，因某些讲述中医理论的段落过于繁杂，高级编码标准难以确定，只能通过关系建立来体现全文含义，无法形成清晰的框架结构。如前文“经典知识平台”标引形成的知识图谱可以确定“酒病”为本段文字中心，但图谱结构较模糊混乱，无法很好体现整体框架。

3.2 知识图谱 从知识图谱分析，“经典知识平台”标引得到的知识图谱更能反映原文含义。其一，从提取得到的语义看，“经典知识平台”较“病脉证并治平台”提取的节点更为详尽，其中“病脉证并治平台”提取得到的语义仅有17个，而“经典知识平台”得到的初级编码则有25个。此外，“病脉证并治平台”提取的语义可为词组或短语，一般较“经典知识平台”得到的初级编码长，初级编码越短，同一段落或不同篇章中越容易得到相同编码，有利于初级编码间形成联系，便于检索及深度挖掘。其二，从标引内容分析，“病脉证并治平台”采用的是从上而下的标引方法，其标引与模板相应的内容，即研究所需要提取的部分，非关注内容则舍弃。但中医古籍文本理论部分范围较广，难以用模板完全概括，存在有意义的内容标引不完全问题，有时需要模板嵌套使用。“经典知识平台”的标引方法为从上而下的标引方法，其标引更注重保留古籍全部内容，从中总结相关内容，从琐碎文本内容进一步提取理论，有利于知识挖掘及深层研究。其三，从节点关系分析，“病脉证并治平台”中“语义关联模板”的存在限制了语义间关系的建立，使很多原文有深层联系的语义无法建立关系，如病因病机“酒性大热”无法与病因病机“伤元气”建立联系，特别是中医理论部分涉及内容较多，很多知识点不能被提取为语义，可提取的语义之间存在因不符合语义关联模板而不能建立相应关系的问题，此时需要自定义语义和语义关联，但

自定义语义和语义关联无法由一般标引人员自主完成，其形成的知识图谱也需特殊处理。而“经典知识平台”的标引方法在中医理论部分的优势更为显著，尤其是“经典知识平台”确定了“本末”（用以标注体与用、象之间的关系，如方药的性味、归经、升降浮沉等）“产生”（表示两个节点之间的阴过关系）“注释”（用以补充明确原文所示内容含义）“阴阳”（用以连接两个形而上的实体，两个实体之间存在互根互用、互不统属的关系）“损伤”（用以表示两个实体之间是损伤的因果关系）等关系。这些关系的确定使一些难以表达的具有中医特色的关系得以在标引中体现出来，也使原文可以被拆分为更小的初级编码。如“酒性大热”在“病脉证并治平台”只能标引为一个语义，即“病因病机-酒性大热”，而在“经典知识平台”中可以拆分为两个初级编码，“酒”和“大热”，建立关系为“酒”-本末-“大热”。特别是当病因病机为一个动态过程时，如“肝郁乘脾”，在“病脉证并治平台”中，一般直接标引为语义“病因病机-肝郁乘脾”，而在“经典知识平台”中，则可拆分为“肝”“脾”“郁”，关系编码为“肝”-损伤-“脾”，“郁”-注释-“肝”，“肝”“脾”-属于-“肝木乘脾”。总之，“病脉证并治平台”提取到的原文内容相对不全面，“经典知识平台”提取到的细节更详尽，更能完整体现原文内容。

3.3　标引自由度　从标引自由度分析，“经典知识平台”标引方法的自由度更高，而“病脉证并治平台”规范性较强。自由度指标引根据标引人员对文献主题的理解自行拟定标引词，规范性指标引依据一定标准，两种标引方法均属于有一定限制的自由标引。在“病脉证并治平台”中，只有符合显示模板的内容才能被标引，模板的存在虽然使标引具有一定规范性，但也有较多限制，如上文提到的原文内容标引不完全、语义关联建立不完整等问题。在“经典知识平台”中，《编码手册》对标引有一定限制作用，但《编码手册》的限制主要体现在关系建立上，相较“病脉证并治平台”，标引人员标引自由度更高，可以尽可能全面标引古籍内容，但其规范性稍差，编码内容较繁杂。

3.4　标引模板建立　从标引模板分析，“病脉证并治平台”在平台上建立模板，而“经典知识平台”的模板则体现在《编码手册》上。“病脉证并治平台”在平台层面框定模板方便标引进行，但此模板一般只适用于此课题，不利于其他课题进行相关研究。“经典知识平台”的标引模板在《编码手册》中确定，不同课题组在使用此平台时可以确定不同标引模板，使其为不同平台提供支持，提高了平台的利用率。

总体而言，两种标引方法的目的不尽相同，其优缺点也不同。“病脉证并治平台”的标引方法构建的知识框架较清晰，但形成知识图谱能否表达原文含义更依赖于模板的制定者，出现偏差后修改难度较高，其标引的自由度稍低，规范性较强；“经典知识平台”的标引方法给标引员的自由度更高，形成图谱的全面性、完整性更依赖于标引员的素质，出现偏差修改难度较低，框架性不足，而标引的自由度较高，规范性稍低。

4　小结

“病脉证并治平台”自上而下的标引方法与“经典知识平台”自下而上的标引方法在知识框架、形成的知识图谱、标引自由度、标引模板等方面各有优势。根据实际需求可选择相应标引方法，如有明确研究目标或需要构建完整知识框架，应选择“病脉证并治平台”标引方法；如为建立相应知识图谱或需进一步进行知识挖掘，则选择“经典知识平台”标引方法更好。

[1] 陈月，刘慧敏，张荣，等. 基于扎根理论的名老中医经验传承内容研究——以姚乃礼教授为例[J]. 西部中医药，2023，36(12)：12-17.
[2] 刘文平，王庆其. 中医理论研究方法论现状及策略[J]. 中华中医药杂志，2019，34(1)：23-28.
[3] 丁侃，柳长华，王凤兰，等. 面向临床的中医古籍数字化问卷调查与分析[J]. 中医文献杂志，2012，30(2)：36-39.
[4] 肖禹. 古籍索引数据应用研究[J]. 新世纪图书馆，2017，(5)：45-48.
[5] 方东行. 中医药文献标引·分类，规范化·自动化的初步研究[J]. 上海中医药大学学报，1996，10(Z1)：82-84.
[6] 李晓瑛，夏光辉，李丹亚. 主题标引文献的语义关系发现研究[J]. 现代图书情报技术，2016，36(Z1)：87-93.
[7] 丁侃. 基于知识元的中医古籍方剂知识表示研究[D]. 北京：中国中医科学院，2012.
[8] 许雯，柳长华. 知识元标引在中医古籍临证文献标引中的应用[J]. 国际中医中药杂志，2015，37(4)：296-298.
[9] 丁长林. 中医古籍文献语义标注技术的研究[D]. 沈阳：沈阳航空航天大学，2013.
[10] 李杲. 脾胃论[M]. 北京：中国中医药出版社，2019：73.
[11] 杨继红. 基于本体的中医古籍叙词表构建方法研究[D]. 北京：中国中医科学院，2008.

收稿日期：2024-08-12
***基金项目**：国家中医药管理局高水平中医药重点学科建设项目(zyyzdxk-2023252)；科技部科技基础资源调查专项(2022FY102000，2022FY102002)；中央高校基本科研业务费专项资金资助项目(2022-JYB-JBZR-011)；北京市社科基金规划项目(23LSC019)。
作者简介：王于静(1999—)，女，在读硕士研究生。研究方向：肾脏疾病的中医诊治。
△**通讯作者**：王维广(1987—)，男，助理研究员。研究方向：《黄帝内经》研究。Email：452786371@qq.com。