

基于本体的中医症状知识表示模型构建^{*}

孙 静 杨 帆

邓文萍 马 利

(湖北省中医院 武汉 430061)

(湖北中医药大学 武汉 430065)

〔摘要〕 介绍中医症状信息利用现状,以《中医诊断学》作为知识源构建症状知识分类体系,采用分类与编码技术构建症状知识表示模型。在电子病历中利用该模型采集症状数据并进行关联分析,得出数据分析结论与临床实际情况相符。

〔关键词〕 本体; 中医症状; 知识表达模型

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.02.011

Construction of TCM Symptoms Knowledge Representation Model Based on Ontology SUN Jing, YANG Fan, Hubei Provincial Hospital of TCM, Wuhan 430001, China; DENG Wen-Ping, MA Li, Hubei University of Chinese Medicine, Wuhan 430065, China

〔Abstract〕 The paper introduces the status of TCM symptoms information application, constructs the symptoms knowledge classification system by taking the *TCM Diagnostics* as the knowledge source, constructs the symptoms knowledge representation model based on the classification and coding technologies, and draws a conclusion that the data analysis result is consistent with the actual clinical situation based on collecting symptom data in the Electronic Medical Records (EMR) and correlation analysis using this model.

〔Keywords〕 Ontology; Symptoms of TCM; Knowledge representation model

1 引言

近年来医院信息化建设取得飞速发展,包括医院管理信息系统、电子病历在内的各类信息系统收

集并储藏了大量临床病例信息,客观记录了患者的诊疗经过、病情变化、治疗效果等情况,为探索病情发展和医疗规律提供了第一手资料^[1]。中医学是一门信息医学,与西医这一试验医学相比而言,通过对真实世界中产生的病例数据进行分析比试验数据分析更能体现中医从临床中来到临床中去这一研究思路^[2]。

症状是中医临床辨证的基本要素,能够真实地反映临床事实,是中医病例的核心数据之一。然而中医在认识疾病、诊断疾病的过程中采用自然语言记录诊疗信息使中医病例数据呈现出多样性,临床术语表达难以统一和规范,术语之间复杂的关联关系缺乏知识层次的数据表达,导致计算机无法从语义的角度识别、理解中医概念及其之间的语义关系,进而在计算机技术支撑下的中医病例信息共享及探索医疗规律这种深度信息利用难以实现。

〔收稿日期〕 2016-11-10

〔作者简介〕 孙静,助理工程师,发表论文 10 余篇;通讯作者:杨帆。

〔基金项目〕 国家中医药管理局中医临床研究基地业务建设科研专项课题“真实世界临床研究中医症状本体构建及其应用示范研究”(项目编号:JDZX2015169);国家中医药管理局中医药信息标准研究与制定项目“中医临床基本症状信息分类与代码”(项目编号:SATCM-2015-XXBZ〔037〕)。

2 中医症状信息利用现状

2.1 主要的数据挖掘内容

当前,在中医药领域展开与症状相关的数据挖掘主要包括药—症之间的关系研究^[3]、症—证之间的关系研究^[4]以及病—症之间的关系研究^[5]。在这些研究中,中医症状的语言描述存在过多使用感性语言、多种用语表达同义或相近含义的症状、用复合词描述症状等特点,导致症状数据存在信息冗余、语义模糊、缺乏量化、多词同义等问题^[6]。对数据挖掘而言,一些成熟的数据挖掘算法要求所处理的数据完整性好、冗余少、属性之间的相关性小,同时症状数据中的冗余信息将会严重影响数据挖掘算法的执行率,症状语义模糊、多词同义等噪声信息的干扰也将会影响数据挖掘的结果^[7]。因此为了消除这些影响,在进行相应的数据挖掘处理之前需要对症状数据进行预处理。

2.2 症状数据预处理

包括症状数据采集时预处理与症状采集后症状数据预处理两种方式。症状数据采集时预处理,即在症状信息录入时采用一组规范化的症状术语,用有无来判断是否有该症状的存在,如将症状“脘腹不适”作为字段,其对应的值为可选项“有”、“无”^[5,8]。症状采集后症状数据预处理,即对采集症状进行规范化处理,通常有 3 种处理模式:一是参考症状相关教材、标准和专著^[9];二是在第 1 类方法的基础上结合专家意见,当出现的症状不能被所参考的资料规范化时,听取专家意见进行相关的规范化处理^[8-9];三是采用关键词的模式规范症状信息^[3]。然而,无论采用哪种处理方式都是从字面上对症状进行标准化处理,没有从内涵上表达症状信息,症状相关信息如部位、性质、持续时间、程度、缓解或加剧因素等被丢失,而这些信息是中医临床辨证论治的重要依据^[10],是中医临床真实世界中客观存在的信息。

2.3 本体

是概念模型规范的形式化说明^[11],是一种结构

化的知识表示方式,其给出了某一领域的概念并形式化地描述了概念之间的关系,且在该领域内达成共识,是计算机处理领域知识的基础。因此,利用构建症状本体表达症状知识能够很好地解决上述问题。

3 中医症状知识分类体系构建

3.1 症状知识来源

3.1.1 知识及知识体系 知识是由多个概念集合在一起并且形成一定的关系^[12],知识体系是指相关知识互相联系构成的整体^[13]。中医症状来源于中医文献和各医家的经验^[14],《中医诊断学》是在中医理论指导下,研究如何通过诊察临床所得资料来识别疾病的一门学科,是基础知识与临床沟通的桥梁,其从四诊的角度系统阐述了症状知识,反映了症状知识体系。

3.1.2 来源 在“读秀学术搜索”上输入书名《中医诊断学》共搜索出相关教材 70 本,其出版年限为 1958-2013 年。随着中医药学的不断发展,年代过早的教材已不能满足并适应当前教学、临床、科研工作的需求,本研究以 1982 年全国高等中医药教材编审委员会成立为标志性事件^[15],筛除 1982 年以前出版的《中医诊断学》,1982 年后出版的教材共 65 本。对 65 本教材的主编进行分析,发现朱文锋主编的教材最多,因此本研究认为由朱文锋主编的《中医诊断学》教材较为经典,流传范围较广,见表 1。

表 1 朱文锋主编的《中医诊断学》教材

出版年	出版社
1995	上海科学技术出版社
1997	中国中医药出版社
1999	人民卫生出版社
2002	中国中医药出版社
2011	人民卫生出版社

本研究采用最新版本,即 2011 年由朱文锋主编、人民卫生出版社出版的《中医诊断学》作为症状知识体系框架的来源。分析《中医诊断学》教材体例,可知其按中医诊断知识体系分为“上篇 诊

法”、“中篇 辨证”、“下篇 诊断综合运用”3 大部分，其中每部分知识按照章节划分进行组织，每单元一般包括基本内容、补充阐发、古代文献、研究进展 4 个方面的内容，其中基本内容概括本单元的基本知识，为该学科中较成熟内容。本研究认为基于成熟内容得出的研究成果更有说服力，更能起到示范作用，因此中医症状知识体系来源于“上篇 诊法”基本内容中阐述的内容。

3.2 构建方法

3.2.1 技术路线 本研究以症状知识体系作为症状本体原型，即将描述症状知识的目录作为本体类原型，通过对症状知识的分析，向下细化症状本体类。同时依据症状术语整理情况，将症状术语作为最小概念从下而上聚类，分析《中医诊断学》诊法内容确定症状聚类方向、症状本体类划分粒度及其属性，见图 1。

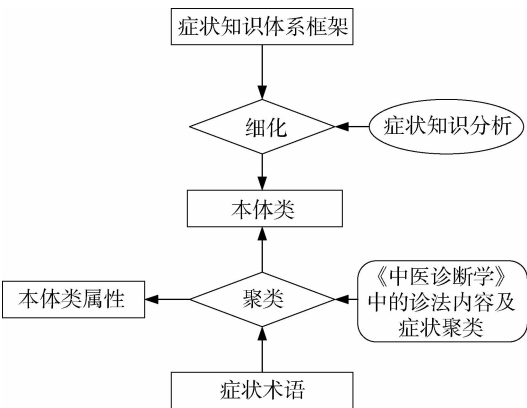


图 1 症状知识分类体系构建技术路线

3.2.2 步骤 症状知识来源于中医四诊，而《中医诊断学》中除了描述症状知识外还有一些诊法知识，整个描述症状知识的目录呈现 3 种情况：一是仅描述诊法；二是从诊法细化到症状分类，如“局部望诊”方法一直可细化到“头形”诊法下的症状分类；三是细化粒度为子诊法。根据以上情况，本研究分 3 步确定症状本体类架构，即删除仅描述诊法知识的目录，自上而下确定诊法分类架构；根据症状术语、症状知识体系中症状聚类及诊法具体内容情况，自下而上聚类症状；整合前两步结果，确定症状本体类及其属性。

3.3 症状知识分类体系模型

根据症状术语整理、症状聚类情况可知症状之间的关系分为两类。(1) 症状概念之间的关系：同义关系、上、下位关系。(2) 症状概念内的关系：症状主体概念、量、色、色泽、质地、位置、形态、气味、动态、病症、次数、排出感、周期、性质、部位、程度、发生因素、加重因素、缓解因素、人群。本研究选用应用相对广泛、界面友好、较好支持中文的 Protégé 作为本体编辑工具，见图 2、图 3。

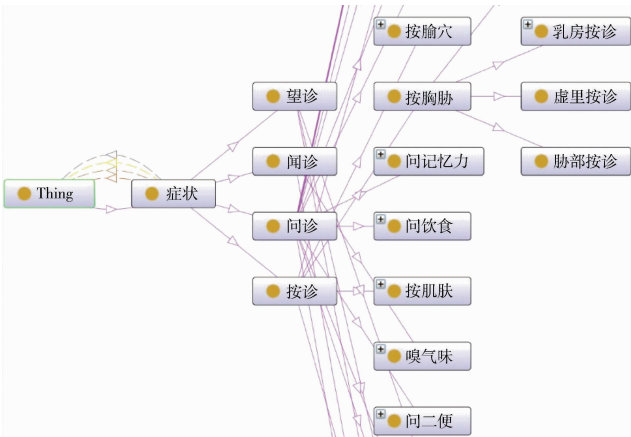


图 2 症状本体类

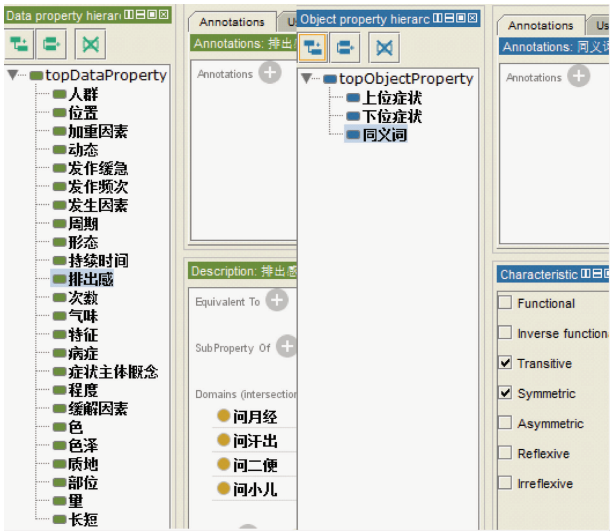


图 3 症状本体类属性

4 中医症状知识表示模型构建

4.1 构建技术

症状之间的关系为同义关系，上、下位关系。当症状的所有属性及其值域相同时为同义关系；当两个症状的主体概念相同，其中在同一个属性上其属性值存在上下位关系时则可以认为在本属性轴上症状具有上、下位关系，体现在属性值分类中。本研究采用多轴分类思想构建症状知识表示模型，即对不同的属性进行分类与编码，通过编码组合的方式表达症状知识。

4.2 表示模型

症状概念的内涵从 24 个方面进行表达，形式化表达结构用方图表示，见图 4。

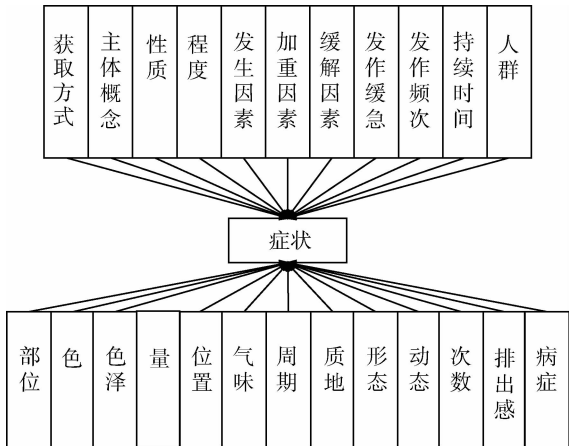


图 4 症状本体形式化表达结构

5 症状知识表示应用验证

5.1 验证方法

本研究以病—证之间的关系为应用研究点，选取中医确认诊断为黄疸的住院病历 114 份为研究对象。住院病历数据来源于某三甲医院肝病科电子病历系统，其中症状信息来源于入院病历中的“主诉”和“刻下症”两部分，邀请中医临床类学生利用中医症状知识表达模型对 114 份病历症状信息进行二次症状数据采集（即症状后结构化处理），然后由临床医生对采集数据进行二次确认。对采集的结构化症状信息进行数据挖掘分析，通过挖掘结果与临床实践进行对比验证模型的可行性。

5.2 结果

5.2.1 症状数据采集结果 将采集到的症状数据导出到 Excel 中，其数据表达形式为一份电子病历对应多个症状数据，每一症状数据的内涵由症状属性组合表达，属性之间用“*”隔开，如症状“目黄 1 月”其症状属性为“望目*目部症状*持续时间长*白睛*黄色”，即症状“目黄 1 月”获取手段为望诊中望目，主体概念为目部症状，持续时间为持续时间长，部位为白睛，色为黄色。

5.2.2 病—证关联分析 依据症状属性内涵对症状进行聚类处理，对聚类后的症状数据进行关联分析。设置支持度分别为 20%、40% 和 60%，得到症状组合网络，见图 5。

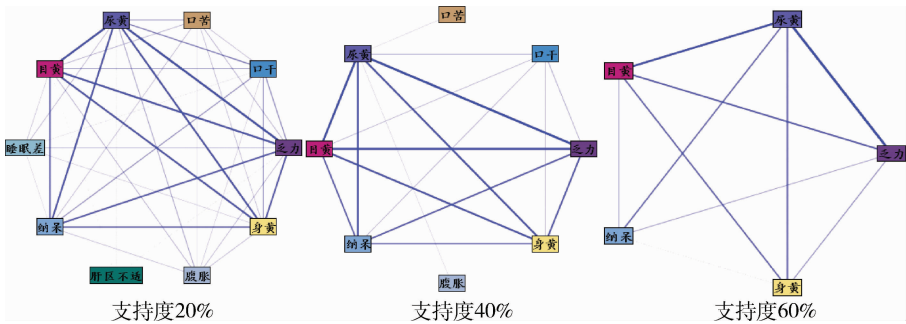


图 5 症状组合网络

5.2.3 病—症部位关联分析 对部位进行关联分析，设置支持度阈值 = 20%、置信度阈值 = 90%，得到黄疸高频症状部位及其组合、关联规则，见表 2、表 3。

表 2 黄疸高频症状部位及其组合

部位	频数	部位	频数
皮肤 = T	104	口部 = T, 白睛 = T	72
白睛 = T	96	口部 = T, 皮肤 = T	72
白睛 = T, 皮肤 = T	96	口部 = T, 白睛 = T, 皮肤 = T	72
口部 = T	84	腹部 = T	60

表 3 黄疸症状部位关联规则

置信度	规则	置信度	规则
1	腹部, 皮肤 -> 白睛	0.92	腹部, 白睛 -> 皮肤
1	腹部, 口部, 皮肤 -> 白睛	0.92	腹部, 口部, 白睛 -> 皮肤
0.99	皮肤 -> 白睛	0.91	白睛 -> 皮肤
0.98	口部, 皮肤 -> 白睛	0.91	腹部 -> 白睛
0.94	口部, 白睛 -> 皮肤	0.90	腹部, 口部 -> 白睛

由表 2、表 3 可知，黄疸出现异常的部位为白睛、皮肤、腹部、口部等，白睛部位出现异常，常伴见皮肤、腹部、口部异常。根据行业组织标准 ZYYXH/T 32 - 2008《中医内科常见病诊疗指南——黄疸》中“黄疸主症：目睛黄染，身黄与尿黄。其中目黄是首要症状；若身黄尿黄而无目黄，则不属于黄疸病证。可伴有恶寒发热，或食欲不振、胃脘胀闷，或右上腹或右胁胀痛等类似感冒、胃脘痛或胁痛的症状”，可知目睛黄染、身黄、尿黄为核心症状，黄疸可伴有腹部异常，食欲不振等症状。该内容与关联分析出的“身、目、尿黄以及乏力、纳呆为核心症状”、“常见异常部位为白睛、皮肤、腹部、口部”等结果大致一致，故可认为症状本体模型基本可行。

6 结语

本文应用本体理论方法构建了症状知识表示模

型，基于该表示模型采集的症状数据较为完整，可从不同角度进行聚类处理，为症状本体在数据挖掘中的应用探索了一条有益路径。然而，本模型没有考虑阴性症状的表达，如在现有电子病历中存在“无胁痛”、“无口渴”、“精神可”、“面容端正”等阴性症状信息，在症状信息采集时该部分信息是否需要采集以及在症状本体中如何表达还需要进一步探索。

参考文献

1 丁丽萍. 大数据时代电子病历应用前景探析 [J]. 档案, 2015, 255 (2): 48 - 50.

2 刘保延. 真实世界的中医临床科研范式 [J]. 中医杂志, 2013, 54 (6): 451 - 455.

3 李仁泽. 基于数据挖掘方法的综合症 - 药物关系挖掘 [D]. 南京: 南京大学, 2013.

4 李文成. 基于文献评价与数据挖掘的亚急性咳嗽证候特征研究 [D]. 北京: 北京中医药大学, 2013.

5 刘少灿. 基于数据挖掘的张德英教授治疗高脂血症临证经验研究 [D]. 石家庄: 河北医科大学, 2014.

6 张菁, 陈涤平, 李文林. 中医症状的语言描述特征及规范思考 [J]. 中医药学报, 2011, 39 (2): 1 - 2.

7 马玉慧. 中医小儿肺炎辨证标准数据挖掘系统中的数据预处理技术 [D]. 沈阳: 东北大学, 2005.

8 边微. 名老中医周绍华教授中风病证治经验的数据挖掘分析 [D]. 北京: 北京中医药大学, 2012.

9 安然. 基于糖尿病结构化中医住院病历数据的糖尿病肾脏疾病病证结合诊疗规律探讨 [D]. 北京: 中国中医科学院, 2013.

10 杨留洪. 中医内科病历主诉书写体会 [J]. 江苏中医, 2000, 21 (12): 46 - 47.

11 Gruber T R. A Translation Approach to Portable Ontology Specifications [J]. Knowledge Acquisition, 1993, 5 (2): 199 - 220.

12 丁佩. 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.

13 林崇德, 王炳照, 王彬, 等. 中国中学教学百科全书 (教育卷) [M]. 沈阳: 沈阳出版社, 1990: 2.

14 黄碧群. 中医症状标准化的必要性 [J]. 中华中医药杂志, 2011, 26 (3): 429 - 432.

15 胡尚一. 高等医学院校医学专业教材编审委员会成立 [J], 中华医学教育杂志, 1982, 1 (1): 3.