中医临床数据库及挖掘分析平台的研究与应用探讨*

□周雪忠 （中国中医科学院广安门医院 北京 100053
北京交通大学计算机与信息技术学院 北京 100044）
刘保延** 姚乃礼 陈世波 （中国中医科学院 北京 100700）
李 平 王映辉 张润顺 （中国中医科学院广安门医院 北京 100053）

摘 要：中医学是以临床医学为基础的学科，辨证论治诊疗实践产生的临床诊疗信息是重要的科学基础数据。本文探讨了旨在实现中医临床诊疗数据存储、分析、利用，并支持观察型中医临床研究分析的中医临床数据仓库及挖掘分析平台的构建和研究，并结合开发实现的平台，分别就中医临床信息模型、多维数据模型、抽取－转换－装载处理、数据规范化整理和挖掘分析平台集成等进行了分析论述。同时针对如何有效的进行挖掘分析应用的问题，探讨了数据预处理、分析目标问题确定和结果的阐释等几个关键环节。

关键词：中医临床数据仓库 数据挖掘 多维分析 中医临床研究

中医学是以观察型临床研究为基础的临床医学[1]，反复的临床实践和临床经验理论提炼形成了中医学发展的基本模式。数据是中医临床研究中首先需要把握和处理的对象，由于中医辨证论治个体化诊疗的特点，单个病例样本数据的多样性和复杂性是中医临床研究中面临的主要课题。基于现代医学临床研究的思路预先进行病例样本因素变量的限定是一种方式，但往往缺乏初期的变量结构关系及其影响因素的研究基础。目前基于现代医学临床研究的方法普遍存在方法学与中医临床研究问题“生搬硬套”和脱离中医临床实际诊疗过程的问题。因此，出现不能体现中医学的特点和本质规律的临床研究结果则成为一种必然。我们认为中医临床研究的首要课题是从保持辨证论治个体诊疗实践的临床中构建并发现中医学的理论框架，发现启发性或可验证性的变量结构及其关系知识，基于临床事实和数据，以及临床专家知识进行反复提炼修正，并获得体现中医临床诊疗规律的科学假设

组稿日期：2007-07-28

* 北京市科委重大计划项目（H020920010130）：中医药防治重大疾病临床个体诊疗评价体系研究，负责人：刘保延；国家973计划项目（2006CB504601）：中医辨证论治疗效评价方法基础理论研究，负责人：刘保延；国家博士后科学基金（2005037106）：中医学，负责人：周雪忠。世界中医药联合会临床疗效评价专业委员会成立大会暨首届国际学术交流会论文。

** 联系人：刘保延，本刊编委，主任医师，国务院政府特殊津贴专家，中国中医科学院副院长，主要研究方向：中医临床研究方法，针灸学，中医信息学，Tel：010-64014411，E-mail：liuby@mail.cintem.ac.cn。

及变量结构关系知识。由此获得的知识和结果能够直接指导临床实践,或者再经现代医学临床研究方法进行有效验证,最终获得可靠的研究结果。因此,从临床实际诊疗过程出发,积累数据,并通过后期分析方法探索发现中医诊疗规律是一个重大的科学研究课题,而日渐成熟的数据仓库技术为我们提供了基本条件。

数据仓库[2]是实现海量数据存储、组织和利用的综合技术平台,已经成为商务智能应用的基础。数据仓库为大量中医临床诊疗数据的积累和利用提供了成熟的解决方案。基于规范的中医临床术语和结构化数据采集,我们在相关项目的支持下进行了中医临床数据仓库、数据挖掘、多维分析(OLAP)平台的研究和构建,并进行了面向中医临床研究的多种分析和应用研究[3]。中医临床数据仓库以实际的临床诊疗过程采集获得的结构化数据为基础,当前该系统的研究和实现以北京市10余家医院相关科室中医临床的三大疾病(糖尿病、冠心病和中风)住院数据和20余位名老中医的门诊数据为基础,已经形成2万余例临床病例的大型数据集。本文就中医临床数据仓库及其挖掘分析平台的实现,以及该平台的分析应用进行阐述和探讨。

一、中医临床数据仓库及挖掘分析平台的构建

数据仓库及分析平台的构建是一个需要不断循环优化的复杂系统工程。为了实现可扩展的基础技术平台,需要中医临床信息模型、抽取-转换-装载(ETL)、数据的规范化整理、数据挖掘平台集成和多维分析系统开发等方面进行系统性的研究工作,下面对此进行简要介绍。

1. 中医临床信息模型研究

在医学信息领域中,HL7参考信息模型(RIM)*是具有广泛接受度和权威性的医学信息模型,HL7 RIM遵循HL7面向信息交换,而不是信息存储的原则。HL7 RIM最上层的类为实体(Entity)、角色(Role)和动作(Act),强调医疗业务过程中的一定角色的实体参与下的行为信息规范。其直接目的是实现医疗过程中各业务信息系统(如HIS、CIS、LIS和EMR等)交互的信息规范。

本文的中医临床信息模型面向中医临床研究,并以实现中医临床数据仓库和数据分析为目的。通过对中医临床信息模型的研究,为能满足中医临床研究应用分析需求的中医临床数据仓库的数据模型设计提供全局、可扩展的参考信息模型。因此,不同于HL7 RIM的信息规范内容,我们注重中医临床研究中关注的信息元素及其关系,注重对临床诊疗过程产生的具备研究价值的数据进行分类规范。在临床数据分析和对中医理论要素总体认识的基础上,我们形成了对中医临床诊疗数据中主体内容的认识,如图1所示。中医临床诊疗过程是医生对患者症、病证动态把握和治疗的过程,在中医理论知识的基础上,通过对症的理解和判定形成对患者疾病状态的认识和处方治疗效果的评价。其中处方治疗过程包含着医生的临床经验,是一种行为性的事件,时间信息和处方结果(如汤药、成药、针灸处方等)是治疗的主要信息内容;而医生对患者病证的判定是一种主观的疾病状态认识,相比较而言,临床诊疗过程中的症信息则是一种具备主观认识和描述的客观现象,是患者疾病状态的实在表现。显然,医生、患者和药物等是中医临床诊疗信息中的物理性实体;而阴阳、虚实、寒热、表里、证候、疾病、药性、功效、归经等则是中医临床诊疗信息中的概念性实体。

因此,在对中医临床信息要素框架认识的基础上,结合实际的临床诊疗数据内容和特点分析,根据中医临床研究的普遍信息粒度和层次,形成了如图2所示的中医临床信息模型。该模型注重中医临床数据以事件为核心的基本特点,对诊疗事件的分析是中医临床分析的主要内容。事件本身是人文和认知性的,世界上并不实际包含事件,事件是一种旨在被采用对变化的有用或相关模式进行分类的方法[4]。我们认为,在中医临床信息模型中,事件包含了医疗实体参与下,在一定时空发生的行为或动作的信息内容。因为,从本质上说,中医临床研究的内容不外乎研究不同事件之间或事件中实体的关系,时间以及时序信息是中医临床研究注重的重要内容,而时间是事件的基本属性。

* HL7 Reference Information Model, http://www.hl7.org/Library/data-model/RIM/modelpage_mem.htm (Accessed: 30 July 2007)

因此，我们把事件作为中医临床信息模型的最上层核心类。根据对事件信息内容的侧重点不同，我们把中医临床诊疗中的事件分为现象（Phenomenon）和活动（Activity）两个大类。现象表示与疾病或治疗相关的表现或状态，如临床表现、生理状态、疾病状态、季节等，现象是活动进行的信息基础或结果，典型的临床表现包括症状体征、理化指标、试验测量指标、疾病史、家族史等等，过去性的时间和时段信息是临床表现的特点。活动则表示引起临床诊疗过程中主客体状态或者信息量改变的行为或者动作过程，在中医临床数据中主要包括就诊、诊察、器械性的理化检查、诊断、治疗和随访等内容，这些活动是产生中医临床诊疗信息的基础。时间属性和引起状态改变的信息是其主要内容，结合现象信息和活动的时序性和时间性，研究事件中的实体关系或事件之间的关系是中医临床研究的主要内容。

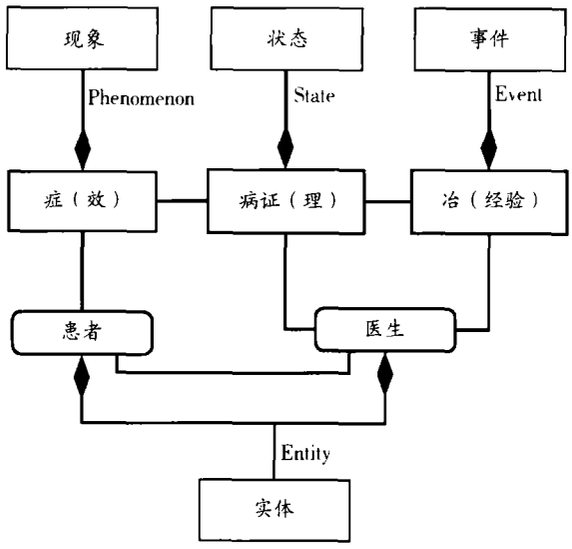

图1 中医临床信息要素框架图

实体则是中医临床信息模型的另一个最上层类，该类定义临床诊疗信息中概念性和物理性的实体信息。物理实体是那些有空间，有质量的东西。这个层次包含了医生，病人，结构，活的有机体，器械，药物、穴位等等。概念实体则规定中医临床诊疗数据中概念性的要素信息，包括中医理论概念和现代医学概念。根据分析粒度的不同，可以对概念性实体的类别进行不同程度的细化，但当前中医临床研究中到疾病、证候，治则治法、药物功效、归经，病因，病机等层次就已经满足分析的粒度要求。实体规定了临床诊疗事件中的主体或要素的层次关系及标准定义，相应于中医临床数据仓库的多维应用分析时，可以具体化为标准或规范的字典维表如中药字典表、归经字典表、中医疾病字典

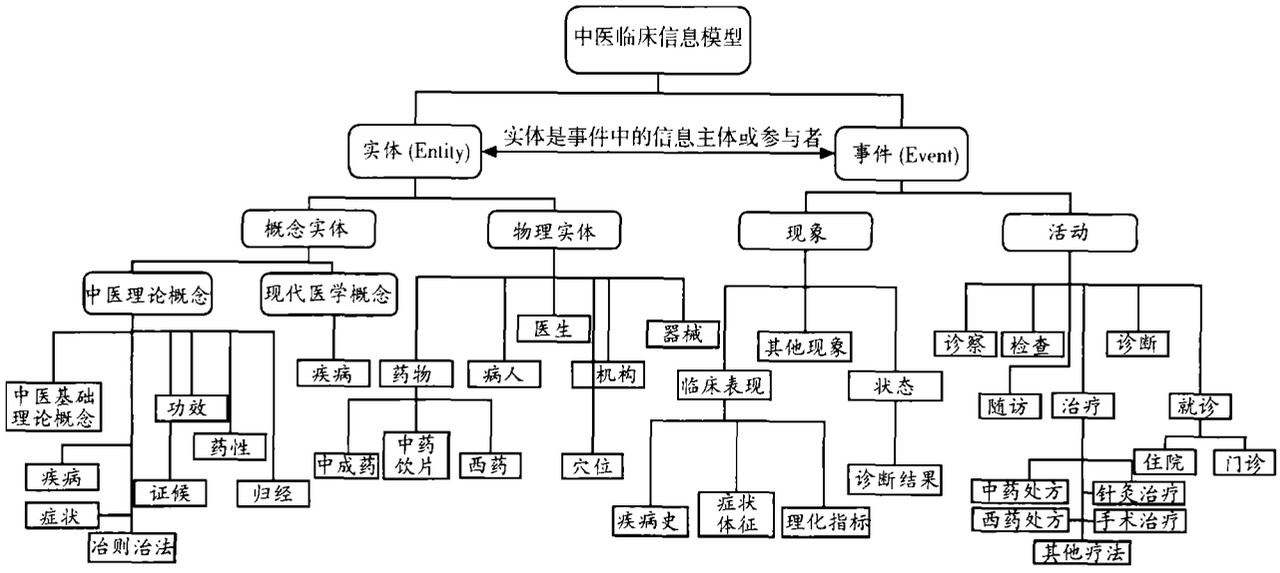

图2 中医临床信息模型

表、证候字典表和医生、病人字典表等。这些字典表之间可以根据各自的语义关系进行关联，如中药字典表与归经、性味、功效等之间具有相应的对应关系，从而实现多层次、多角度的关联分析需求。

UMLS[5]的语义网络(Semantic Network)被认为是著名的医学术语本体。其对术语的154种语义类型定义中把实体和事件作为两大最上层类，且把实体分为概念性实体和物理性实体。UMLS对医学术语概念的语义类型定义启发了我们对中医临床信息模型的认识。中医临床信息模型中实体及其下位类在名称上与UMLS的语义类型一致，但作为信息模型的层次，其具体定义更接近于HL7 RIM的实体类，而与UMLS的语义类型并无关系。中医临床信息模型的事件类也是如此，如HL7 RIM对动作类的定义，中医临床信息模型的事件类是临床诊疗信息的类别，而不是术语的类别。本文在概念框架上初步研究了中医临床信息模型，但具体的模型实现(在多维主题物理数据模型中已经有初步的设计思路体现)，则有待后续的研究完成，目前，我们仅以此概念性的框架指导中医临床数据仓库开发过程中的数据模型设计。

2. 中医临床多维数据模型的研究

多维数据模型是Web OLAP分析的数据源，以事实表和维表形式表达。本文基于ROLAP多维数据模型设计了面向主题应用的数据集市雪花型数据模型。本节以临床复方药物主题数据模型为例，分析中医临床数据仓库多维模型的设计问题(临床复方药物多维模型见图3)，临床复方药物多维模型旨在为从多个角度和维度分析临床处方用药规律提供高效、逻辑明确

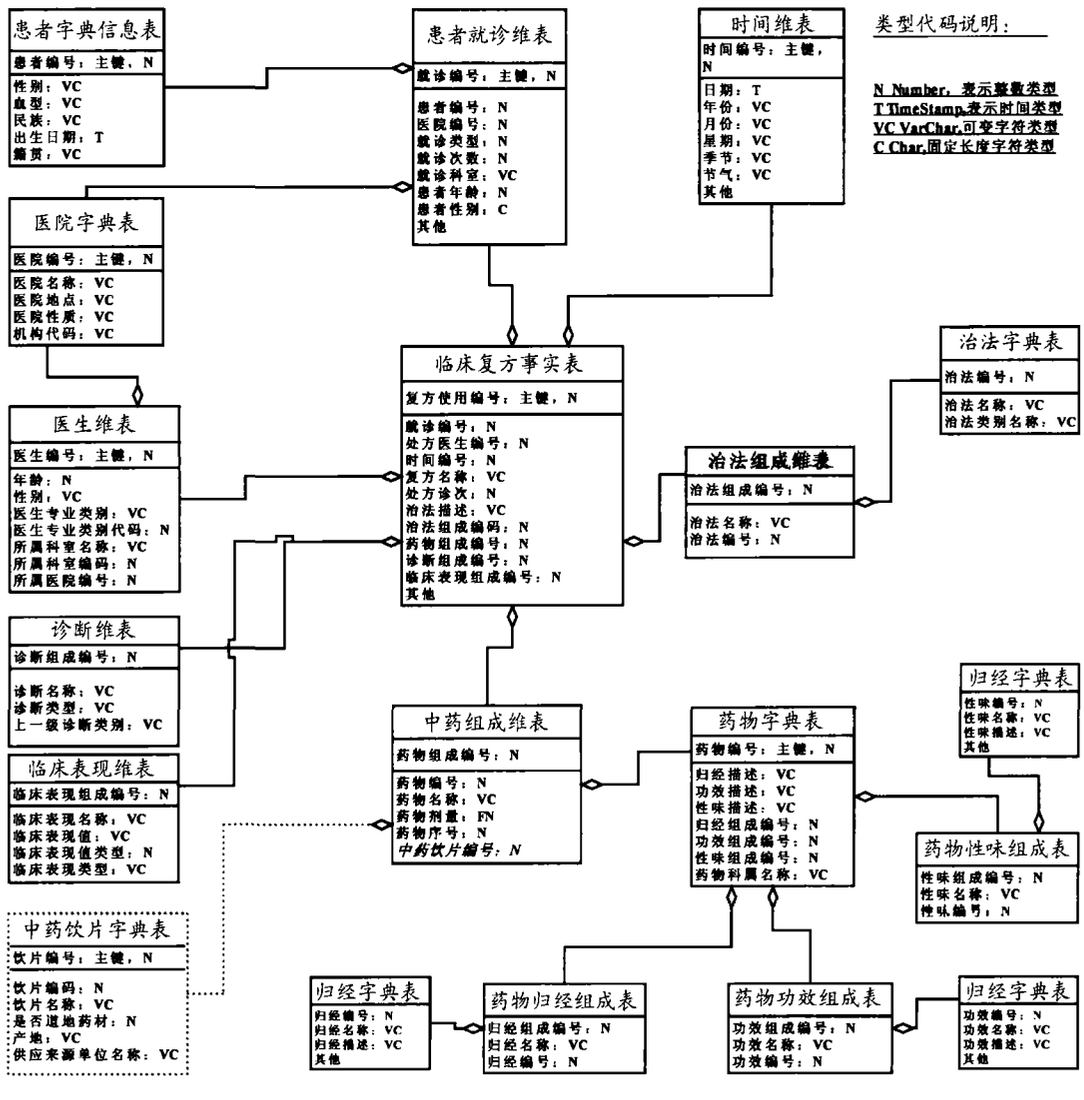

图3 中医临床复方药物主题的雪花型多维模型设计

的模型基础。包含一个事实表－临床复方事实表和多个分层的维表如治法、药物、诊断、临床表现、患者和医生等，多层的维表结构体现了雪花型多维模型的特点，其结构相对复杂。基于以上事实表和维表设计结构，能够支持多维度的临床复方药物主题分析应用。可以从不同的维度如医生（名老中医）、患者、诊断、临床表现和时间等分析临床复方药物的使用情况，可以治疗某病证的处方用药经验或规律。临床复方多维事实表的内容原则上应该直接包含药物组成和结构化的治法信息，但鉴于数据存储容量和效率的考虑，我们在该多维模型中把中药组成和治法组成从事实表中拆分出来，以维表的方式存储。在 OLAP 主题开发和设计时，可根据需求，创建新的物理视图/视图把以上两个维表实体化为临床复方药物度量的内容。

3. ETL 软件的开发与实现

ETL 系统是数据仓库中数据整理阶段的主要工具，它与数据管理相结合，提供对不同的数据源的数据进行抽取、转换、清洗、装载的功能。用户从数据源中抽取所需要的数据，经过数据清洗，最终按照预先设计的数据模型规范，将数据装载到数据仓库中。它是构建和开发数据仓库成败的关键步骤。根据调查统计，数据仓库项目中往往有 60% 的精力花费在数据 ETL 的实现和实施上。由于中医临床数据的数据源电子病历结构复杂、诸多数据存在不一致性。另外，由于对原先文本型的病例数据进行结构化存储，因此，数据类型的核查十分重要。不同数据类型的转换等都可能造成 ETL 过程失败或者数据丢失。我们在中医临床数据仓库 ETL 的开发和实施过程中，主要解决了数据类型的匹配，数据格式的转换，异地数据表数据整合时的主键重复、数据规范整理和挖掘分析数据格式接口等问题。我们采用 Java 实现了称为 Medical Integrator 的自主版权 ETL 软件。

4. 数据规范化整理

大规模中医临床数据的规范整理是临床数据分析利用的重要步骤。由于临床术语使用和自然语言表述的多样性，中医临床数据的术语性整理是非常关键的环节。通过研究中医临床数据整理的共性规律，我们设计了基于规则的术语及数值数据整理方法，该方法基于一系列的规则库和 ETL 相应的功能模块。

我们对术语数据的规范整理总体上从同义词规范、相关概念信息扩展和术语的上位归纳等三个方面进行。同义词规范主要处理临床病例数据中的不规范数据如气滞、气滞证，我们都统一为气滞证，以及概念性的同义术语如关节痛、关节疼痛，我们统一为关节疼痛；相关概念信息扩展则对涉及的概念信息通过字典表进行必要扩展，如中药名称，我们通过增加包含中药规范名称、中药归经、性味、功效、药物分类等信息的中药字典表进行信息扩展，从而实现从归经、性味、功效和药物分类的维度对中药的使用情况进行分析的数据基础，西药的信息扩展也采用相同的机制；术语的上位归纳处理则对具体的概念性术语进行类别层次的上升处理，如小腿疼痛，归纳为下肢疼痛，同时保留原先术语。由此，实现了多层次和不同粒度的分析能力的数据规范基础。根据相关课题的临床研究分析的需求和临床专家对术语信息的理解，目前我们已经编辑形成了 10 万余规则库和词典库数据记录。

5. 数据挖掘平台的集成应用

在中医临床数据仓库数据模型设计和 ETL 系统数据挖掘接口的基础上，我们通过集成现有的数据挖掘软件平台如 Weka[6]，Oracle Data Miner* 等实现常规数据挖掘方法的联机应用，实现了海量临床数据存储和挖掘分析平台的无缝集成。数据仓库中的临床数据能够在数据挖掘平台中即时取样，并进行多种数据挖掘方法的应用分析。挖掘分析的结果同时又存储到数据仓库中，并结合商务智能平台进行挖掘知识的展现与进一步分析利用。利用集成的数据挖掘平台，我们针对中医临床中的处方用药经验、辨证经验和疾病人群分析等进行了多种方法的分析和研究。

6. 多维分析系统的实现

商务智能是数据仓库提供领域决策支持能力的重要手段和技术。在充分进行研究文献和商业产品调研的基础上，我们选用了业界领先的商务智能软件－

* Oracle Data Miner homepage, http://www.oracle.com/technology/products/bi/odm/odminer.html (Accessed: 30 July 2007)

Business Objects XI* 作为中医临床数据仓库 Web OLAP 分析的基础。基于中医临床数据仓库的细节及多维数据模型和前期一致性的海量数据存储，本文应用 Business Objects(BO)开发了面向名老中医经验传承、重大疾病的病证及临床表现要素关系等的主题分析集。这些主题分析集能够体现并分析中医临床中以人、以病证和处方治疗等不同角度的关系规律和经验知识，是对中医临床的总结归纳，能够有效辅助中医临床研究，为中医理论框架和科学假设的形成提供思路和启发，并以临床实际诊疗为基础，实质性的提高中医临床研究水平和临床研究能力。在中医临床数据仓库多维数据模型的基础上，基于 BO 实现多维分析的主要任务是实现多维数据模型到语义层(Universe)模型的语义映射，构建了面向主题的多维分析语义层之后，利用 BO 提供的分析任务设计软件如 Crystal Report，Web Intelligence 等就能便捷的实现分析任务的设计。

二、中医临床数据仓库平台的分析应用探讨

中医临床数据仓库及挖掘分析平台构成了中医临床研究的基础技术平台，该平台能够整合和整理结构化的临床诊疗数据，不断积累形成宝贵的中医临床数据资源。而且，基于规范化整理后的数据，该平台为中医临床研究中观察型研究阶段的变量关系的认识、科学假设和理论框架的形成等提供了分析方法和技术平台。在前期的多项分析应用如糖尿病及血管并发症辨证论治规律[7]、田从豁教授临床配穴经验整理研究[8]、急性冠脉综合征辨证论治规律分析[9]、2 型糖尿病代谢综合症证候分析[10]以及北京市 10 余位名老中医诊治经验分析[3]等工作中，充分体现了该平台的研究效率和科学价值。

数据挖掘是从大数据集中发现有效、创新、潜在有用和最终可理解的模式的非平凡过程[11]，是新兴的机器学习应用研究方向，为从海量的中医临床数据中发现和提炼中医临床经验和中医理论知识提供了有效的手段。但正确的认识和把握数据挖掘及分析过程中的关键问题是进行有效分析和应用的前提。下面我们简要探讨挖掘分析应用过程中需要注意的关键环节和问题。

1. 进行充分的数据预处理

数据挖掘或多维分析是从最原始的业务数据到知识发现的过程，原始数据必然存在信息缺失、噪音、错误或者不规范等情况，数据预处理是进行分析之前关键而耗时的步骤，有时需要反复进行。中医临床数据以术语型信息为主，因此，该步骤涉及的内容和处理更加繁琐。我们的平台软件中提供了数据整理功能，但仍然需要进行大量的人工处理如术语概念化、数据筛除、清理等。

2. 明确分析目标的科学问题

分析目标往往对应一个具体的科学问题，涉及到相应的数学模型和分析用数据。不同的挖掘分析方法在模型、算法特点和适用条件等方面一般各不相同，而且，不同的分析目标对应的数据变量情况也各具特点。因此，临床研究人员与分析人员进行充分交流，准确认识分析目标的科学内涵和模型问题，明确数据的特点(如高维性、变量类型等)是进行可靠的挖掘分析工作的基础。在此条件下，才能科学地选择合适的挖掘分析方法进行研究。

3. 按照使用方法的原理或条件进行结果的阐释

分析结果的专业阐释是数据挖掘的重要环节，解释人员需要充分理解使用方法的原理和结果的基本内涵才能赋予分析结果合理、合适的专业意义。不然可能导致错误的研究结果。如张等[12]对以往利用变量聚类方法进行基于症状样本的证候研究的文献进行分析发现，变量聚类方法存在与分析目标无法匹配的问题。对症状变量进行聚类，并试图获得证候学方面的知识的研究存在得到错误结论的风险，若最终获得的症状类不能达到证候分型判断时对阳性值的要求就会出现问题。虽然，这些可能导致错误的研究是极少出现的，但为保证科学可靠的研究结果，紧密结合方法学原理的分析结果阐释仍是挖掘分析过程中需要严格把握的关键步骤。

* Business Objects XI homepage, http://www.businessobjects.com/products/businessobjectsxi/default.asp (Accessed: 30 July 2007)

三、探讨及未来工作

中医学几千年的医学实践和研究模式是临床－理论－临床的不断循环，螺旋式研究及发展的过程。中医药理论为无需动物模型分析实验直接进行可靠的临床诊疗过程提供了保证，临床实践从多方面反映了人体疾病状态变化规律和复杂干预调节的动态决策规律和过程。基于中医临床诊疗实践，进行大规模的临床数据采集，规范化整理和计算机辅助自动挖掘研究，能够为新世纪中医学的理论和临床诊疗水平的提升提供可持续发展的技术和信息平台。本文介绍了相关课题中业已实现的中医临床数据仓库及挖掘分析平台的主要研究内容，并讨论了分析应用中的主要环节和问题。

我们认为，中医临床诊疗数据的结构化存储、整理、分析和应用是关乎以临床医学为核心的中医学继承、创新和发展的关键环节，是融合吸收现代化技术，并提升中医学研究水平的必要途径。海量的中医临床实际诊疗数据是中医学价值的最佳体现，也是中医学不断继承创新的知识源泉。因此，长期坚持以临床为核心的技术体系研究和应用，不断规范的持续积累临床诊疗数据，创建临床科研一体化的中医学临床湿干研究模式[1]应该成为中医临床研究的重要方向。

参考文献

1 刘保延，周雪忠. 中医临床研究方法的思考与实践－系统生物学湿干研究模式与中医临床研究. 世界科学技术－中医药现代化，2007，9(1)：85～89.

2 W. H. Inmon, Building the Data Warehouse (Third Edition), *John Wiley & Sons*, Inc. 2002.

3 周雪忠. 中医临床数据仓库构建及临床数据挖掘方法研究。中国中医科学院，博士后出站报告，2007年3月.

4 Allen J. F., Ferguson G., Actions and Events in Interval Temporal Logic, *Journal of Logic and Computation*. 1994, 4(5): 531～579.

5 Lindberg D. A. B., Humphreys B. L., McCray A. T.. The Unified Medical Language System. *Meth Inform Med*, 1993, 32: 281～91.

6 Witten I. H. and Frank E., Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

7 姜兆顺. 基于结构化住院病历采集系统对糖尿病及血管并发症辨证论治规律的研究，中国中医科学院，2005，博士论文.

8 张华. 田从豁教授临床配穴经验的整理研究. 中国中医科学院，2006，硕士论文.

9 高铸烨. 基于数据挖掘对急性冠脉综合征辨证论治规律的探索性研究. 中国中医科学院，2006，硕士论文.

10 陈世波. 2型糖尿病合并代谢综合征中医证候研究. 中国中医科学院，2006，博士论文.

11 Fayyad U. M., Piatesky－Shapiro, G., Smyth, P., and Uthurusamy, R., Advances in Knowledge Discovery and Data Mining, AAAi/MIT Press, 1996.

12 张连文，周雪忠，陈弢，等. 论证候研究中变量聚类结果的诠释. 中国中医药信息杂志，2007，14(7)：102～105.

Studies and Applications: TCM Clinical Data Warehouse and Associated Data Mining Platform

Zhou Xuezhong, Liu Baoyan, Yao Naili, Chen Shibo

(*China Academy of Chinese Medical Sciences, Beijing*, 100700, *China*)

Zhou Xuezhong

(*College of Computer and Information Technology, Beijing Jiaotong University, Beijing*, 100044, *China*)

Li Ping, Wang Yinghui, Zhang Runshun

(*Guanganmen Hospital, China Academy of Chinese Medical Sciences, Beijing*, 100053, *China*)

Clinical medicine is a core component of traditional Chinese medicine (TCM). Clinical information generated from clinical activities, featured with synthesized treatment based syndrome differentiation (STSD) is the data of scientific importance. This paper describes the construction of a clinical data warehouse for storing, analyzing, and utilizing TCM clinical data, in an attempt to support the analysis of observational clinical studies. We have built the platform, and discussed the main procedures, including TCM clinical information model design, multi－dimensional data model design, extraction－transform－loading development, data trimming, and data mining & analysis system integration. Furthermore, we analyzed several issues regarding applications of data warehouse and analysis platform in addressing TCM clinical analytical problems, including preprocessing, target problem analysis, and knowledge interpretation.

Keywords: traditional Chinese medicine clinical data warehouse; data mining; online analytic processing; traditional Chinese medicine clinical research

(责任编辑：王　瑀，责任译审：邹春申)