

# 基于框架和产生式表示法的病历知识库研究

## Study of medical record repository based on framework representation and producing representation

沈亚诚<sup>1</sup>, 舒忠梅<sup>2</sup> (<sup>1</sup>广东药学院, 广东 广州 510240; <sup>2</sup>中山大学, 广东 广州 510275)

**摘要:**病历知识库是开发电子病历和临床辅助诊断系统的基础,病历大都以带格式的形式出现。采用框架表示法和产生式表示法相结合,可实现病历的表示和存储。病历知识库结构由综合数据库、病历规则库、知识获取模块和人机接口组成,将每一病历规则拆分成条件部分和结论部分,并分别占据一条记录。病历知识获取是知识库建立的关键和难点,可选取规范病历,借助知识编辑器来实现。模糊匹配算法是诊断推理的主要依据,对病历中非文本型数据的匹配计算,如影像,可采用数组转换的方法解决,病历模糊匹配主要通过病历内容比较,阈值计算、加权平均来实现。

**关键词:**电子病历;知识表示;模糊匹配;知识库

**中图分类号:**G271 **文献标识码:**A **文章编号:**1673-4254(2006)10-1467-04

病历知识库是采用人工智能技术,将大量的临床医学知识有机地组织起来的知识库系统,是开发电子病历的基础,也是开发临床辅助专家系统的基础,应用前景广阔,病历知识库的研究在国内正处于起步阶段。

国内现阶段的医院信息管理系统大都以面向经济管理为主,国内一些公司也正在开发简单的病历系统,这些病历系统主要是采用文本型模板及直接打字输入的方式书写病历,检查申请与结果报告实现了无纸化传递,加快了结果的回报速度。但病历书写速度慢,生成的病历文件为文本型,没有校对和检索功能,没有病案质量检查功能;同时基于文本型模板的书写方式,不能与诊疗过程有机地结合起来,易受模板内容的影响而发生出错现象,不利于临床思维活动,效果不够理想。采用智能化病历知识库,可以提示医生进行必要的检查或处理,避免遗漏,能辅助医生确立医疗方案,规范医疗操作,提供医疗违规警示,避免医疗错误。联动专业数据库,象药品数据库,诊疗方案数据库等,供医生查询参考,有助于提高医疗质量。

国外临床应用系统发展较快,有不少较先进的系统,但国外的病历书写规范和表达方式不同,系统价格昂贵,同时其病历系统也不尽完善等因素,如缺少图表、影像资料等,阻碍了国外临床病历系统在中国的推广应用。

### 1 病历知识及其表示方法

病历是记录病人在诊断治疗全过程的原始记录,

包括首页、病程记录、检查检验记录、手术记录、护理记录等,其中既有结构化信息,也有非结构化的文本信息,以及图形图象信息等,这些记录大都是陈述性和推理性的。大部分表述具有规范性和确定性,但也有不少表述是模糊的,甚至是不确定的。

病历知识是病人多个信息的关联,病历知识表示就是病历知识的符号化和形式化的过程。从综合的角度来说,病历知识具有如下特性:

(1)相对正确性。在病历书写基本规范要求中,病历的内容一般是正确的。如病人的身体检查报告,医嘱处方等。

(2)不确定性。病历知识是病人有关信息关联在一起形成的信息结构,“信息”与“关联”构成病历知识的两个要素。由于病人个体的复杂性,检测手段和水准的差异性,获取的信息可能是精确的,也可能是不精确的、模糊的;关联可能是确定的,也可能是不确定的。使得病历知识并不只有“真”与“假”两种状态,而是在“真”与“假”之间还存在许多中间状态,即存在为“真”的程度问题,如:病人肚子痛,原因可能是吃了不卫生的食物。病历知识的这一特性就是其不确定性。

(3)可表示性和可利用性。病历知识可以用适当形式表示出来,如文字、符号、图表、影像、切片等。病历实际上就是病历知识的一种约定,一种按某种规则(如病历书写基本规范等)用于记录健康状况、临床表现和诊疗的数据结构。对病历知识进行表示的过程就是把病历知识编码成某种数据结构的过程。目前,用得较多的知识表示方法主要有:一阶谓词逻辑表示法、产生式表示法、框架表示法、语义网络表示法、脚本表示法、过程表示法、Petri表示法、面向对象表示法等。对同一知识,一般都可以用多种方法进行表示,但其效果却不相同。

产生式表示法的结构“if ... then ...”接近人类的

收稿日期:2006-02-03

基金项目:广东省医学科研基金项目(A2003339)

Support by Medical Science Foundation of Guangdong Province(A2003339)

作者简介:沈亚诚(1962-),男,硕士,副教授,E-mail:shenyacheng@126.com

通讯作者:舒忠梅,女,硕士生导师,E-mail:issszm@mail.sysu.edu.cn

思维和会话的自然推理形式，在医疗诊断领域中，其知识一般具有经验性、因果性的特点，因此适用于产生式进行表示。框架是一种描述所论对象(一个事物、一个事件或一个概念)属性的数据结构。一个框架由若干个被称为“槽”的结构组成，每一个“槽”又可根据实际情况划分为若干个“侧面”。一个槽用于描述所论对象某一方面的属性，一个侧面用于描述相应属性的一个方面。槽和侧面所具有的属性值分别称为槽值和侧面值。无论对于框架、槽或侧面，都可以附加上一些说明性的信息，一般是指一些约束条件。不同格式的病历，其内容的表示形式也不同，具有框架表示的特性。为充分表示病历的各种结构关系和内容，采用框架表示法与产生式表示法相结合。对各种格式的病历，采用框架表示法去定义建立；而对各种格式内病历的内容，采用产生式表示法去实现。

## 2 病历知识库结构模型研究

病历知识库的研究主要致力于将临床医学有关知识有机地组织到数据库中，使之能够被灵活调用，并像人类专家一样按某种规则进行重组和表示，所以，病历知识库的研究重点就在于知识的获取、表示和运用这 3 个环节。

病历知识库内部结构决定着以上 3 个环节的实现机制和执行效率，病历知识库内部结构建立应具有以下特性：

(1)可扩充性。要求系统的数据结构和存储程序必须足够灵活，使得不需要作任何硬件或控制结构上的修改就能对知识库进行扩充。

(2)明确性。所采用的知识表示方法能明确地表示各类知识，便于知识库的检查与调试，提高知识重组和表示的效率。

(3)清晰性。表示模式必须有利于知识的检索和推理。一旦知识表示模式被选定，其被调用的方法也就被确定下来。

(4)易理解性。便于知识库研制人员把专家的专门知识整理归纳及形式化，也便于知识库的设计、实现和改进。

目前，组成病历的格式有 20 多种，不同格式的病历，使用于各个就诊环节或多个不同的系统中，并产生相应的医疗信息。要将这些信息按照类别及发生的时间顺序有机地组织为一个整体，需要建立病历的描述结构。这些内容的信息量往往很大，若只用一个文件存储在计算机中，势必影响信息的交换速度和效率。因此，每一种格式的病历分别建立一个文件或数据表，并以住院号或门诊号为关键字，以提高数据的访问速度和效率，同时减少系统的开销。

病历知识库结构包括 4 部分：综合数据库、病历

规则库、知识获取模块、人机接口。各部分功能如下所述：

(1)综合数据库。综合数据库(有时也称事实库)是病历知识库的基础，由格式数据库和病历事实库组成。格式数据库用于存放病历格式的初始数据和结果，事实库包含器官图库、症状库、疾病名称库、药物库、连接词汇库等。这些库一般由文字、数字、符号、基本图形等组成。

(2)病历规则库。病历规则库是整个知识库设计的核心，它由诊断推理规则库、病历格式规则库和病历书写规则库组成，分别包含疾病诊断推理规则、病历格式结构规则和病历书写规则。病历格式规则库描述病历格式的组成结构、组成元素，以及组成元素的属性。病历格式规则库采用框架表示法进行表示，为了进行病历的模糊匹配计算，病历的每个槽分别赋予相应的值，以区别它们的重要性或可信度。诊断推理规则库，主要包含条件部分和结论部分，每一个条件部分也可拆解开，分解为相关多个格式病历，为方便规则的增加、修改和推理控制，采用如下策略：拆解开条件部分和结论部分，数据库中每一条件和每一结论都分别占据一条记录。书写规则库的结构与诊断推理规则库类似。

(3)知识获取模块。这是病历知识库系统中能将临床医学领域内的事实性知识和领域专家所特有的经验性知识转化为计算机可利用的形式并送入知识库的功能模块。同时也负责知识库中知识的修改、删除和更新，并对知识的完整性和一致性进行维护。

(4)人机接口。人机接口负责把临床医师和病历专家等输入的信息转换成系统内规范化的表示形式，然后把这些内部表示交给相应的模块去处理，并以不同的组合形式输出；系统输出的内部信息也由人机接口转换成用户易于理解的外部表示形式显示给用户。病历知识库总体结构模型如图 1 所示，综合数据库和病历规则库是病历知识的存储形式，知识获取模块对输入的知识内容进行处理，分别送入综合数据库或病历规则库，并进行管理，而人机接口则是对人机交互的有关内容进行规范化处理。

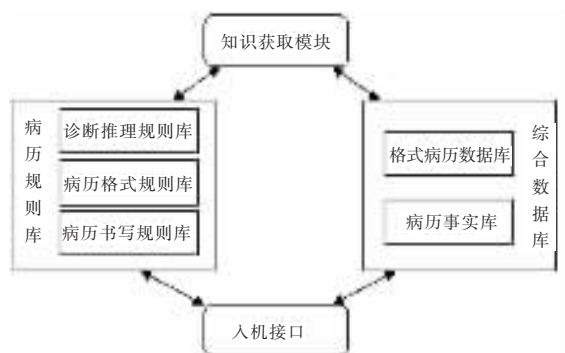


图 1 病历知识库总体结构模型

### 3 病历知识的获取

病历知识的获取是病历知识库建立的关键。按知识获取的自动化程度分,知识获取可分为非自动化知识获取和自动知识获取两种方式。自动病历知识获取是指病历知识库系统自身通过机器学习而获取病历知识的能力,目前大多数专家系统不具备这种方式。

非自动化知识获取分两个步骤:首先由知识工程师从临床规范病历和有关医学资料中获取知识,然后用病历知识编辑软件输入到病历知识库中,其工作方式如图 2 所示:



图 2 非自动化知识获取方式

目前非自动化知识获取比较普遍,它主要通过查阅专业领域的书籍、论文和专著,通过与领域专家对话,通过分析系统实际运行临床诊治记录,通过分析疾病的起因、现象、诊断和治疗等途径来获取知识。

采用非自动化知识获取方法获取知识,一般通过知识编辑器来辅助实现知识的获取,最简单的情况下,包括如下功能:

- (1) 辅助输入,编辑器动态提供相关词组和短语,供用户快速选择输入。
- (2) 规则检查,编辑器调用病历书写规则进行正确性检查。
- (3) 知识抽取,辅助用户将输入的内容进行抽取,并将抽取的新知识追加到相应的知识库中。

实践表明,知识的获取是病历知识库开发中最困难的过程,而且消耗的人力、财力也最多。病历知识获取就是把用于临床诊断和病历领域的知识从拥有这些知识的知识源中抽取出来,并转换为特定的计算机表示。通常,这些知识源以下面的表现形式存在:医学教科书、病历书写规范、示范病历、医药资料、临床检验数据和影像资料等。因此,病历知识的获取基本步骤如下:

- (1) 根据现存的各种规范病历,如首页、病程记录、检查检验单等,定义好病历的格式框架,添加病历格式规则库的内容;
- (2) 选取示范病历,分别更新不同格式病历对应的病历事实库;
- (3) 根据病历书写规则,抽取病历知识的表示规

则,建立病历书写规则库;

(4) 根据示范病历的格式规则及其病历数据库的内容,以及对应的临床诊断疾病关系,得出临床诊断规则,从而更新临床诊断规则库;

(5) 选取相关医学教科书,临床病历、医药资料,通过病历知识编辑器,并调用病历书写规则和病历数据库的内容,进行模糊匹配,确定可信度值,从而添加病历数据库的内容;

(6) 以某一病历格式和病历数据库为起点,选取相关疾病的规范病历格式和病历数据库内容,进行模糊匹配,计算可信度值,从而添加临床诊断规则库的内容;

(7) 重复以上步骤,并不断增加示范病历,扩大疾病的范围,使病历知识库的知识能满足某一专科使用。

病历知识的获取流程如图 3 所示:

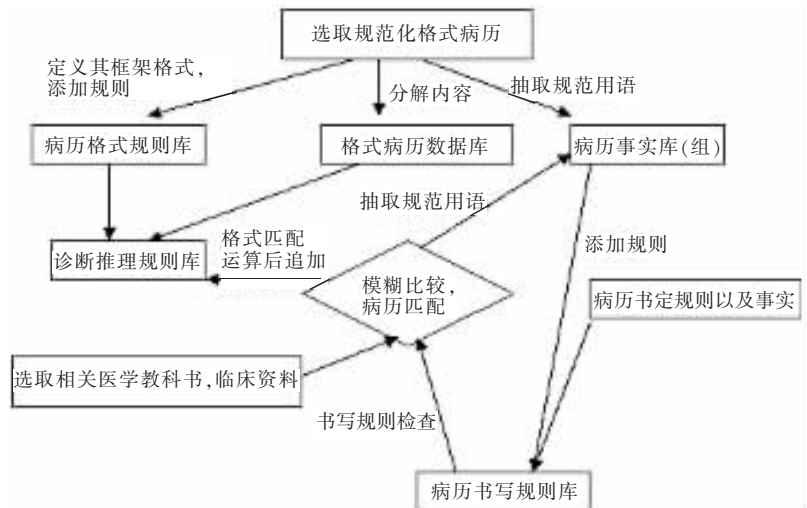


图 3 病历知识获取流程

医学的范围非常广阔,只能先取其一个分支开展研究,然后,逐步增加其他人体系统的病历知识内容,从而形成一个较完整的病历知识库,并生成软件包。

### 4 病历模糊匹配算法

某一疾病的病历由一系列不同格式的病历组成。各个不同格式的病历采用框架进行表示,框架表示法主要通过槽及其侧面来描述所论对象。尽管病历的组成元素除了文字、符号外,还包含图表、影像、切片等,但图表、影像、切片等资料,可通过转换,变为多维有界数组表示,如手术中病灶位置图,就可用如下方法表示:人体简图名称(病灶水平左坐标,病灶垂直上坐标;病灶水平右坐标,病灶垂直上坐标,病灶性质表示)。因此,病历框架的槽及其侧面的值一般是字符串、数、数组对等,其表达形式是用模糊数,模糊语言



值等,含有不确定性。

病历的模糊匹配度由组成病历的各个不同格式病历的匹配程度决定。先计算两个同一格式病历的匹配度,然后再计算病历的平均匹配度值,作为病历的模糊匹配度。可预先设定一个阈值,如 0.90,如两份病历模糊匹配度大于 0.90,可判断其疾病相近,模糊匹配度大于 0.98,可判断其疾病相同;反之,如两疾病相同,即其疾病模糊匹配度大于 0.98,则治疗过程部分的病历记录可基本相同。

两个同一格式病历的匹配度计算步骤如下:

(1)分别求出两各病历格式对应属性的匹配度。设  $A_1, A_2, \dots, A_n$  是当前病历格式中各属性的模糊集,设  $B_1, B_2, \dots, B_n$  是规范化病历格式中各对应属性的模糊集。根据模糊集理论及模糊逻辑,则各对应属性的相似度为  $M_i = [A_i \cdot B_i + (1 - A_i \odot B_i)] / 2$

其中:  $A_i \cdot B_i = \mu(A_i) \wedge \mu(B_i)$  表示内积,  $\wedge$  表示取极小;  $A_i \odot B_i = \mu(A_i) \vee \mu(B_i)$  表示外积,  $\vee$  表示取极大。 $\mu(A_i)$  表示  $A_i$  属性模糊集与对应的标准属性模糊集的接近率,可采用集合元素比值法和数值比值法进行计算。

(2)求格式病历的匹配度  $M$ 。因为格式病历中,每个槽的重要性值  $Q_i$  不同,通过加权平均值,计算出病历的匹配度  $M = (\sum Q_i \times M_i) / \sum Q_i$

例如:某一简化阑尾炎手术记录,格式及规范病历的有关属性值设定如下:

术前诊断:(<疾病名称库>,重要性值 0.15)

术后诊断:(<疾病名称库>,重要性值 0.55)

手术名称:(<手术名称库>,重要性值 0.05)

手术经过:(重要性值 0.25)

假设有 A、B 两份病历,各槽的权值、槽属性与对应的标准属性接近率如表 1 所示:

表 1 权值及属性接近率表

槽名称	槽 1:术前诊断	槽 2:术后诊断	槽 3:手术名称	槽 4:手术经过
权值	0.15	0.55	0.05	0.25
A 与标准属性的接近率	0.90	0.85	0.78	0.90
B 与标准属性的接近率	0.95	0.80	0.80	0.87
对应属性的相似度 $M_i$	0.475	0.475	0.490	0.485

则对于槽 1 对应属性的相似度为  $M_1 = [A_1 \cdot B_1 + (1 - A_1 \odot B_1)] / 2 = [0.90 + (1 - 0.95)] / 2 = 0.475$ ,依此类推,求得其他槽对应属性的相似度为  $M_2 = 0.475, M_3 = 0.490, M_4 = 0.485$ 。则两格式病历的匹配度为  $M = (\sum Q_i \times M_i) / \sum Q_i = (0.15 \times 0.475 + 0.55 \times 0.475 + 0.05 \times 0.495 + 0.25 \times 0.485) / (0.15 + 0.55 + 0.05 + 0.25) = 0.4785 \approx 0.48$  即 A、B 两病历的匹配相似度约为 48%。根据阈值约定,可断定这两例疾病的治疗过程相差较大。

### 5 结语

病历知识规范化是病历知识库建立和推广使用的难点之一。目前国际上已形成的标准有:国际医学规范术语全集(SNOMED)、国际疾病分类标准:ICD-10、ICD-9-CM 等,国家卫生部已颁布执行若干

种医疗记录规范,可供病历知识采集使用。但大量的病历知识表示尚未规范化,如医疗卫生技术(服务)的术语、病人(门诊及住院)基本医疗数据术语等。要组织相关专家进行规范化的工作,根据具体情况确定病历格式中各个属性的重要性值,病历模糊匹配算法要根据不同类型的内容进行优化,以提高病历知识库在开发临床辅助系统的应用价值和效果。

### 参考文献:

[1] 沈亚诚. 病历知识库的构建探索与实践[J]. 广东药学院学报, 2006, 22(1): 110-3.

[2] 王永庆. 人工智能原理与方法[M]. 西安: 西安交通大学出版社, 1998: 57-200.

[3] 沈亚诚. 电子病历的研究与开发[J]. 中国病案, 2001, 2(2): 23-5.