

基于本体的医学知识获取^{*})

周肖彬 曹存根

(中国科学院计算技术研究所智能信息处理重点实验室 知识获取和共享课题组 北京 100080)

Medical Knowledge Acquisition: An Ontology-Based Approach

ZHOU Xiao-Bin CAO Cun-Gen

(Knowledge Acquisition and Sharing Group, Key Lab of Intelligent Information Processing Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

Abstract In this paper, we introduce an ontology-mediated method for medical knowledge acquisition and analysis. Using the method we establish an ontological structure and ontologies for the Medical Knowledge Base (or NKIMed). To check the consistency of the acquired knowledge, we use a set of medicine-specific axioms. These axioms are also used in knowledge inference, and interconnection between different medical concepts. Finally, two applications of NKIMed, i. e. intelligent teaching systems and speech diagnosis are illustrated.

Keywords Medical ontology, Ontology-mediated knowledge acquisition, Knowledge analysis

1 引言

知识是实现人工智能的必由之路。大规模知识处理致力于各学科专业领域的知识处理,从知识的角度来探讨智能的实现方式,并提供可共享、可复用的知识。医学知识库(NKIMed)是国家知识基础设施(NKI)的一个子集。

医学是人类最密切相关的科学,医学知识也当然成为人类最重要的知识,因此世界各国都在进行医学知识工程的研究。其中以美国国立医学图书馆的统一医学语言 UMLs 最为著名^[1]。但目前还没有一个大规模的、可共享、可复用的医学知识库。因此建立一个良好的医学知识库具有极重要的意义,它将为医学专家系统、医学信息检索、医学教育系统、自然语言理解等领域提供智能基础。

概念及概念之间的相互关系,是人类知识的表述方式,而概念的分类是人类组织知识的方法,这是基于工程的本体论的主要观点。它的应用受到越来越多的重视,很多有名的知识系统中本体论都有不同程度的应用,如美国 D. Lenat 教授领导的小组正在研制的一个大型的常识知识库系统 Cyc, Princeton 大学研制的语言知识库 WordNet,国内陆汝钊等研制的常识知识系统等。

本体论已成为知识工程中的一种常用工具,用于解决知识获取这个知识工程的瓶颈问题^[2~6,10,12,16],受到了广泛的重视和研究。它用以分析一个论域中的类(categories)、关系(relationships)以及它们所满足的限制(constraints)^[7,9~11]。需要指出的是医学本体体系并不只是一种描述语言,而是独立于任何语言之上的客观存在。同时对医学这样一个巨复杂系统来说,我们所描述的医学本体结构并不是唯一的。

我们基于形式本体的理论建立了 38 个医学形式本体,并以此为基础进行了基于文本的医学知识获取和知识分析。

2 医学本体体系概述

2.1 医学本体论

医学本体论从知识工程的角度来说,指的是一个医学本体结构以及对医学概念的本体知识描述。在这里本体知识既可以在概念层次上进行定义也可以在语言层次上进行描述,它包含了概念的所有相关知识。也就是说我们在描述概念时,按照曹存根教授的提法,既要“言之有理”(知识必须符合医学领域,与医学领域矛盾),又要“言无不尽”(完整地 and 全面地认识概念域,对隐式的或不确定的内容进行形式刻画)。

在本体设计中,基本元素是类(category),表示专业领域中的归属集体,每一个类中都有相应的个体(individual)及实例(instance)。类间的基本关系是继承(inheritance)和实现(implement),为了描述类及类的个体及实例,我们使用槽(slot)的描述符,并将槽分为属性(attribute)及关系(relationship 或 relation),为了对槽名及槽值进行限制,同时定义了侧面(facet)。下面是关于类“疾病”的描述(要说明的是,限于篇幅,本文所有关于本体的类描述都只是引用部分而不是全部,实例知识也是如此)。

Defcategory 疾病 继承 医学本体,和 名称本体 实现 治疗属性,和 并发症属性

```
{
  英文名:disease
  属性:症状
    :类型 字符串数组
    :注释 “患病病人病人主观上的异常感觉,如‘恶心’”
}
```

2.2 医学本体结构

医学是个复杂的学科。我们从知识工程的角度出发,在参考众多分类法的基础上^[9,14,15],形成了医学本体体系。在设计结构时,既要考虑到医学专业的通用性,同时还必须考虑到知识库处理的方便性、可共享性、扩展性以及知识库知识的再提取,因此我们设计了如图 1 所示的医学本体结构(注意:在医学本体中不存在虚本体,即只包含槽名,而不包含槽值的本体)。

3 医学本体设计及医学知识获取

3.1 本体描述语言

^{*})本文中的工作受到国家自然科学基金资助(课题号 # 20010010-A)、科技部重大基础研究专项资助(课题号 # 2001CCA03000)、中科院二期创新工程资助。

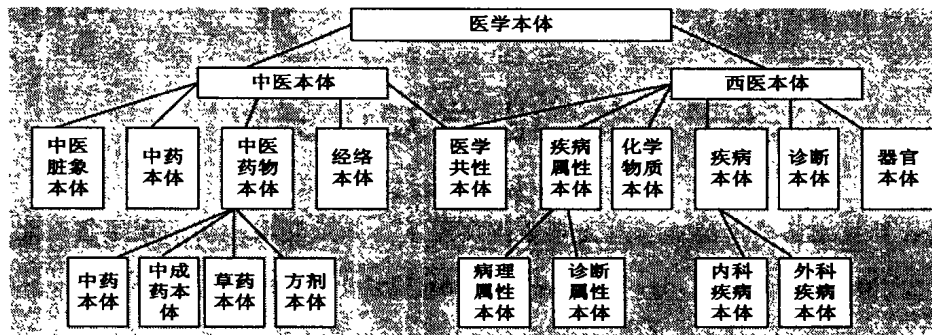


图 1

我们在描述医学本体时使用的是 NKI 本体语言^[1]。NKI 本体语言是一种框架语言(Frame Language),吸收了 Generic Frame Protocol[Farker et. al 1998]^[9]的设计思想。NKI 本体语言主要文法如下所示:

```

<NKI 本体描述> ::= <类定义> { <类定义> } { <公理定义> }
<类定义> ::= DefCategory | DefPrivateCategory <类名> [ <相关类> ]
{
    { <类知识描述> }
    { <槽定义> }
    { <类内公理定义> }
}
    
```

类定义的核心内容为槽及类公理。槽代表了本体的属性及概念相关的知识,是我们在医学本体中描述的“言之有理”“言无不尽”的重要组成部分。其描述文法如下图:

```

<槽定义> ::= <槽类型> : <槽名>
            类型 <槽值类型>
            [ : 值域 <值域> ]
            [ : 单位 <槽值的单位> ]
    
```

槽的槽类型可分为 5 种类型:属性槽(简称属性)、关系槽(简称关系)、属关槽(简称属关)、方法槽(简称方法)、聚类属性槽(将相同性质的属性归纳为一个虚属性)。

为了反映本体之间的关系,定义了继承和实现两种类间关系以处理类的相关类。如下所示:

```

<相关类> ::= 继承 <继承类序列> [ ; 实现 <实现类序列> ]
            | 实现 <实现类序列> [ ; 继承 <继承类序列> ]
<继承类序列> ::= <概念类> { , 和 <概念类> }
    
```

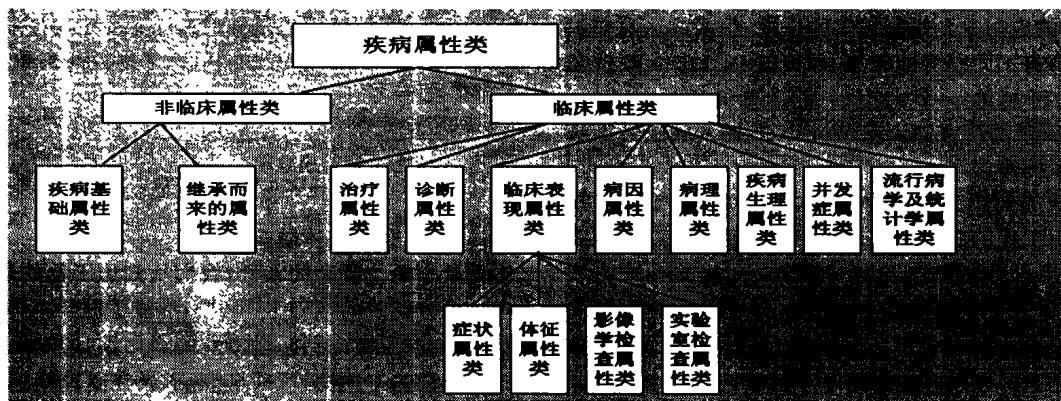


图 2

图 2 是疾病属性类的分类结构。描述了 NKI 中描述疾病类时所使用到的属性及关系集,由于属性之间固有的层次关系,疾病的属性及关系本身也可形成一个本体体系,称为疾病属性类。这个本体将不含有任何实例,而只能由疾病类来实现它,同时在这个属性本体中反映的类别知识也是类公理的一部分。

<实现类序列> ::= <聚类属性类> { , 和 <聚类属性类> }

一个类 C_1 继承另一个类 C_2 表示 C_2 的公共知识可以被 C_1 继承, C_2 中定义的槽可以用来描述 C_1 的知识; C_2 中的相关公理可以被 C_1 继承。

一个类 C_1 实现另一个类 C_2 表示 2 个含义。第一, C_1 可以使用 C_2 中的词汇(即属性、关系、属关和方法); 第二, C_1 继承 C_2 中的所有公理。

NKI 类中另一个重要内容就是类公理,类公理本身也是知识,它是类概念知识之上的元知识,是更高层次的知识。它是对类中槽以及槽间关系的一种约束,它反映了知识内在涵义。医学公理是保证医学知识一致性的重要手段,它也是医学知识推理的必需前提,还可以为知识的联通提供一种方法。

根据上面给出的 NKI 本体语言,我们可以对医学本体进行描述。

3.2 医学本体的属性及关系

3.2.1 医学本体中的属性 医学是关于疾病防治及人类健康保健的科学,因此在医学知识中,最重要的以及最核心的知识就是关于疾病的知识^[8]。疾病是一个综合的医学概念,涉及到了病理,生理,解剖,实验室检查,治疗,统计,分类等各个方面,根据医学学科的范畴,可分为临床属性类和非临床属性类,如图 2 所示。

下面描述了流行病学与统计学属性本体作为例子说明具体的属性设计。流行病学与统计学是医学领域的一个分支,它是研究人群中疾病与健康状况的分布及其影响因素,并研究如何防治并促进健康的策略和措施的科学。从它的定义我们可以看出,它从群体研究各种各样疾病,方法则是从疾病的频率和分布出发。因此我们把它作为疾病一个属性类是很合理

的,它正好也表示了有关疾病的一些属性。

defcategory 流行病学与统计学属性 继承 医学本体

```
{
  属性:起病
  :类型 字符串
  :值域 急,和 慢,和 缓,和 缓慢,和 缓和,和 和缓,和 急骤,
  和 暴急,和 突然
  :注释“指疾病发生时的病变速度,如缓慢、急骤”
  属性:特好发人群
  :类型 字符串数组
  :同义词 尤好发人群
  :近义词 特易感人群
  :注释“指特别易患该疾病的人或人群”
  属性:发病率
  :类型 数量
  :同义词 发生率,和 病发率
  :近义词 患病率
  :侧面 范围,和 地区,和 人群,和 条件,和 时间,和 来源,
  和 根据
  :注释“指疾病发生的可能性,均指人群的平均可能性”
}
```

3.2.2 医学本体中的关系 医学本体中定义了一些关系,这些关系在医学本体中将医学概念与医学概念联结一起,具有很重要的作用。

表 1 医学本体中的若干典型关系

关系	注释
继发于	表示一种疾病由另一种疾病引起
并发	表示一种疾病引起另一种疾病
易误诊为	表示将一种疾病诊断时误认为是另一种疾病
多见于	表示疾病常常见于某个身体部位
原发于	表示疾病原发部位
鉴别于	表示疾病的鉴别关系

3.3 医学本体设计方法及知识获取

NKI 医学本体在设计时参考了 Gruber^[4]提出的本体构造的五个原则,即明确性和客观性,一致性(Coherence),可扩展性(Extendibility),编码偏好程度最小(Minimal encoding bias),本体约定最小(Minimal ontological commitment)。

我们在设计医学本体时提出了以下几个原则:

1. 具体类的下面不应该有抽象类。
2. 类的整体分布应该比较均匀。
3. 类的继承关系不能有循环,可以有重复继承,但不能有冗余继承。
4. 类的命名唯一。
5. 专业知识的描述词汇专业化。

具体设计本体中的槽时,我们采用的是从文本中直接获取知识的方法。对医学知识源要求知识可信度高(最好是国家标准或国际标准之类的知识源)、知识覆盖面要宽、知识不能陈旧过时。

设计医学领域的属性和关系时,原则上尽量采用医学术语,这样可以在从形式化还原到自然语言时,我们可以更自然地以医学语言的方式进行还原。但不可避免地会遇到形式化和医学术语表达之间的矛盾,比如流行病学与统计学属性类中的“特好发人群”,在正常医学术语中是没有这样的术语的,可为了表达“特别易患疾病的人群”的意思,我们在形式化表示和医学表示之间取中,得到“特好发人群”这样一个自然语言上可以理解的槽。下面是我们根据医学本体设计获取的关于糖尿病的知识。

defcategory 糖尿病 继承 疾病

```
{
  英文名:diabetes mellitus
```

```

  中医名称:消渴,和 消瘴
  西医病因:糖尿病家族史,和 病毒感染,和 自身免疫
  中医病因:饮食不节,和 情志不调,和 劳逸失度
  症状:多饮,和 多食,和 多尿,和 疲乏,和 消瘦 等
  多发人群:老年人
  发病率:0.67%
  :地区 中国
  临床分型:I型糖尿病,和 I型糖尿病
  :根据 胰岛素依赖
  体育治疗方法:适当体力活动,和 耐力运动
  饮食治疗方法:控制饮食
  :准确度 100%
  治疗药物:甲苯磺丁脲,和 氯磺丙脲,和 优降糖,和 达美康,和
  糖适平,和 克糖利,和 美吡达,和 降糖灵,和 二甲双
  胍,和 葡萄糖苷酶抑制剂,和 胰岛素制剂
  属性:胰岛素依赖治疗
  :类型 布尔
  :注释“用于表示是否在治疗时要依赖胰岛素”
}
```

defcategory I型糖尿病 继承 疾病

```
{
  英文名:insulin-dependent diabetes mellitus
  英文缩写:IDDM
  正式称呼:胰岛素依赖性糖尿病
  是一种:糖尿病
  起病:急
  胰岛素依赖治疗:是
```

defcategory I型糖尿病 继承 疾病

```
{
  英文名:non-insulin-dependent diabetes mellitus
  英文缩写:NIDDM
  正式称呼:非胰岛素依赖性糖尿病
  是一种:糖尿病
  起病:隐潜
  胰岛素依赖治疗:否
  诱因:肥胖
```

3.3.1 本体中属性与属性间关系表示 由于在疾病类中所涉及到的属性及关系非常之多。这些属性及关系显然不可能是相互独立的,而会形成一个层次关系。这种层次关系反映了属性的类别知识。并随之成为医学公理的一部分。我们采用属性本体结构以及聚类属性的方式来表示这种属性关系间的分类关系。如下的表示将好发人种,少见人种,最高发病人种,好发人群,罕见人群,易感人群等具体属性通过人种,人群,发病群体形成聚类:

defcategory 流行病学与统计学属性 继承 医学本体

```
{
  聚类属性:发病群体
  :类型 字符串数组
  :值域 人种,和 人群
  聚类属性:人种
  :类型 字符串数组
  :值域 好发人种,和 少见人种
  聚类属性:人群
  :类型 字符串数组
  :值域 最高发病人种,和 好发人群,和 罕见人群,和 易感
  人群
```

4 医学公理:表示和获取

医学公理是医学知识之上更高层次的知识,是关于医学知识的知识,是医学元知识,具有很高的抽象性,我们在这种抽象性上可以对医学知识进行分析。

4.1 公理表示方法

医学知识分析建立在医学公理上面,医学公理使用一阶逻辑来表示公理。使用“X:(类)”表示 X 是一个类变量,X 的取值是类中的任意一个实例;用“X·S”表示 X 的某个槽 S 的取值。S(X,Y)表示 X 的在槽 S 上的取值为 Y。同时,我们定义了一些函数和谓词来表示公理,如表 2 所示。

医学公理分为类间公理和类内公理,类间公理约束一个类和另一个类,以及一个类的槽和另一个类的槽;类内公理约束类中的槽以及类中槽与槽之间的关系。下面我们对类间公理和类内公理的获取分别述之。

表 2 医学公理中使用到的谓词(部分)

谓词或函数	含义
先于发生(A,B)	谓词:A 发生在 B 之前
同时发生(A,B)	谓词:A 与 B 同时发生
isa(X,Y)	谓词:X isa Y
大于等于(A,B)	谓词:A 的值大于 B 的值
是实例(A,B)	谓词:A 是 B 的实例
小于(A,B)	谓词:A 的值小于 B 的值
发生(A)	谓词:A 出现或发生
Beginwith(x,y)	谓词:X 以 Y 开始
Cutbegin(x,y)	函数:将 Y 从 X 的开始删除,返回删除后的 X 串
Addbegin(x,y)	函数:将 Y 加到 X 的开头,返回增加后的 X 串

4.2 公理获取方法

4.2.1 类间公理获取 类间公理分为类与类公理和类槽与类槽间公理。我们在获取这一类公理的方法是对每个类都从医学领域的涵义上,找出该类与其它类的关系,再把这种关系用一阶逻辑的方式表示。

4.2.1.1 类与类公理 用来约束一个类与另一个类的关系。为一个类与另一个建立联系提供描述。表 3 是这一类公理的一部分实例。

表 3 类与类公理

公理 1	所有 X:疾病,存在 Y:症状[是临床表现(Y,X)]
公理 2	所有 X:疾病[beginwith(X,“原发性”) -> isa(X, cutbegin(X,“原发性”))]
公理 3	所有 X:疾病[isa(addbegin(X,“恶性”),X)]
公理 4	所有 X:疾病[简称(X,Y) -> 简称(addbegin(X,“原发性”),addbegin(Y,“原发性”))]

4.2.1.2 类槽与类槽间公理 用于约束一个类的槽及另一个类的槽,为多个类之间的槽建立联系。我们在获取这一类公理时,将一个类与另一个类的同名槽或分类层次上的上下位槽进行比较,从医学意义上得出它们之间的关系。表 4 是这一类公理的一部分。

表 4 类槽与类槽间公理

公理 1	所有 X:糖尿病,所有 Y:人[大于(X·空腹全血血糖测定,Y·空腹全血血糖测定)]
公理 2	所有 X:糖尿病,所有 Y:糖尿病性肾病[大于(X·发病率,Y·发病率)]
公理 3	所有 X:恶性肿瘤,所有 Y:良性肿瘤[大于(转移速度(X),转移速度(Y))]

4.2.2 类内公理获取 医学公理中的类内公理包括类中关于属性医学涵义的公理以及类中属性之间关系的公理。是建立在医学本体之上的对本体中的槽及槽间关系的一种约束。我们在获取这一类公理时,采用对每一个槽进行分析,找到表示槽含义和约束的公理。我们把类内公理分为值域公理、语义类型公理、属性关系公理、属性关系内涵公理。

4.2.2.1 值域公理 用以对槽值进行限定,将槽的取值限定在一个合理的范围内,可以有助于进行知识的错误检查,以发现一些常识上的矛盾。如表 5 的公理 2 就是对疾病的发病率的取值作出限定。

表 5 两条值域公理

公理 1	所有 X:疾病[大于(X·复发率,0) ∧ 小于等于(X·复发率,1)]
公理 2	所有 X:疾病[大于等于(X·发病率,0) ∧ 小于等于(X·发病率,1)]

4.2.2.2 语义类型公理 用于对槽值的语义类型进行说明,实质上也是对槽的意义的一种解释。这有助于知识获取时对一个概念进行语义分析。如表 6 的公理 1 表明疾病的并发症一定是一种其它疾病。如前面的值域公理一样,如果槽值是可枚举的,我们也可以值域来固定语义类型。而不用再定义语义类型,比如流行病学与统计学属性中的属性“起病”,我们就可以不用再定义它的槽值的语义类型。需要说明的是,在语义类型公理的 isa(X,Y),其中 X 可能是一个概念集合。

表 6 语义类型公理(部分)

公理 1	所有 X:疾病 [并发症(X,Y)→isa(Y,疾病)]
公理 2	所有 X:疾病 [继发于(X,Y)→isa(Y,疾病)]

4.2.2.3 属性关系公理 用于对属性与属性之间,关系与关系之间的元关系进行限定,以便于在知识分析时进行诸如知识纠错查错、知识推理。如表 7 的公理 1 表明了最大发病年龄与最小发病年龄之间的关系。

表 7 属性关系公理

公理 1	所有 X:疾病 [大于(X·最大发病年龄,X·最小发病年龄)]
公理 2	所有 X:疾病 [小于等于(X·人群死亡率,X·发病率)]
公理 3	所有 X:疾病 [并发症(X,Y)→病因(Y,X)]
公理 4	所有 X:疾病 [预后(X,不好)→严重程度(X,高) ∧ 转归(X,不好)]
公理 5	所有 X:疾病 [小于(X·人群死亡率,X·死亡率)]
公理 6	所有 X:疾病 [易复发(X,是)→大于等于(X·复发率,20%)]
公理 7	所有 X:疾病 [大于(X·复发率,0)→可复发(X,是)]

属性关系公理、属性本体结构、聚类属性构成了整个医学体系中的属性本体体系,它在整个医学本体体系中占有很重要的位置。

4.2.2.4 属性关系内涵公理 是对属性及关系的内在意义的一种解释性公理,这是一种比较难以定义的公理,我们的目的在于把属性关系的内涵尽可能地用一阶语言公理的形式固定下来,如表 8 的公理 1 表明了疾病的最高发病年龄如果是 Y 的话,那么其它所有年龄在此疾病上的发病率都没有 Y 年龄高,它是对“最高发病年龄指疾病或疾患发生的最大年龄或最大年龄段”这句属性定义的形式化表示。

表 8 内涵公理(部分)

公理 1	所有 X:疾病 [最高发病年龄(X,Y)→isa(Y,年龄组) ∧ 所有 Z:年龄组 (小于(发病率(X·Z),发病率(X·Y)))]
公理 2	所有 X:疾病 [并发症(X,Y)→先于发生(X,Y)]
公理 3	所有 X:疾病 [并发症(X,Y)→小于(严重程度(X),严重程度(有新概念(X,Y)))]
公理 4	所有 X:疾病 [代表病种(X,Y)→所有 Z:疾病 [isa(Z,X) ∧ →等于(Z,Y)→大于等于(Y·发病率,Z·发病率)]]

属性关系内涵公理是一种综合性公理,在内涵公理的形式化中,可以包含值域公理、语义类型公理、属性关系公理。

5 基于公理的医学知识分析

在医学本体及医学知识库建立之后,一个非常重要的任务就是对医学知识进行分析,医学知识分析内容包括知识一

致性分析和医学知识推理两个部分,一致性分析是分析知识中的“言之无理”的部分,也就是和医学领域知识矛盾的地方,这种矛盾可能来自于文本知识源的错误、人为的疏忽、对医学领域的理解不够深而导致的本体设计上的失误,还可能来自对某个问题域描述的缺失。推理分析是分析知识的内在涵义,从而得出知识库中没有的且和医学领域相一致的新知识。医学公理中的每一类公理都即可用于一致性分析,也可用于知识推理。

5.1 基于公理的医学知识一致性分析

公理库的每一类公理均可用于一致性分析。此处的一致性分析包括知识完备性检查、值错误检测、知识矛盾检测。下面以几个例子来说明公理在这些方面的应用。

例 1:表 3 的公理 1 告诉我们,一种疾病的知识中必然包含某些症状是该疾病的临床表现的知识,如果一种疾病的知识中没有关于症状的知识,那么关于该疾病的知识是不完备的。

例 2:假设我们有以下知识:

defcategory 糖尿病:疾病

```
{
  空腹全血血糖测定:5mmol/L
  继发于:青霉素
  发病率:5%
  人群死亡率:10%
  死亡率:8%
}
```

defcategory 青霉素:药物

```
{
  英文名:penicillin
}
```

defcategory 人:动物

```
{
  空腹全血血糖测定:4.4~6.7 mmol/L
}
```

我们来看看公理库可以告诉什么,表 4 的公理 1 可告诉我们的空腹全血血糖测定值可能有错。而表 6 的公理 2 则告诉我们糖尿病应该继发于一种疾病,所以它继发于青霉素的知识是错误的;表 7 公理 3、公理 6 则告诉我们,关于发病率、人群死亡率、死亡率的知识是不一致的。

5.2 基于公理的医学知识推理

公理库除了在一致性分析上的应用外,另一个重要应用就是在医学知识推理方面的应用。根据公理,我们可以从已知的概念出发,推理未知概念的一些知识;也可以从概念的已知属性出发,推理未知属性知识。

例如表 6 中的公理 1 告诉我们一种疾病 A 的并发症必然是另一种疾病 B,这样就算库中没有关于 B 的知识,我们也可知道 B 是一种疾病;表 7 的公理 6、公理 7 则告诉我们疾病 X 如果易复发的话,则可推出其复发率大于 20%,且该疾病一定是会复发的疾病。

由表 3 的公理 2、公理 4 则告诉我们,我们可以处理一些新出现的医学概念,如对于“甲状旁腺功能亢进症简称甲旁亢”这条知识,可以得到“原发性甲状旁腺功能亢进症简称原发性甲旁亢”并进一步得到“原发性甲旁亢是一种甲状旁腺功能亢进症”。

6 医学本体及医学知识库的应用

这样设计的本体及获取到的知识库我们可以用在很多方面,如医学专家系统、智能教学系统、语音识别、自然语言理解等方面。下面简要叙述一下我们根据以上的设计方法上获取的关于糖尿病的知识在语音识别、智能教学中的应用。

智能教学系统 基于本体的医学知识可以为医学教育提供很大的灵活性及互动性^[13],可以给教师提供编写教案的描述语言和面向本体的教学,减少了教师的重复工作,教师给出教学主题,由我们的教学系统根据教学主题所属的本体(如糖尿病:疾病本体),呈现给教师本体的内容,然后由教师设计面向本体的教学教案,最后由教学系统与知识库的交互获得教学主题的具体教学,经编译后生成教案。如内科教学教案。通过这样的方式把教学本体与知识库联系起来。同时,教师还可以根据不同的教学目的,根据本体中的属性关系本体将本体中的槽进行取舍以满足不同的教学需要。

基于知识的语音处理 在语音识别中会遇到大量模糊语

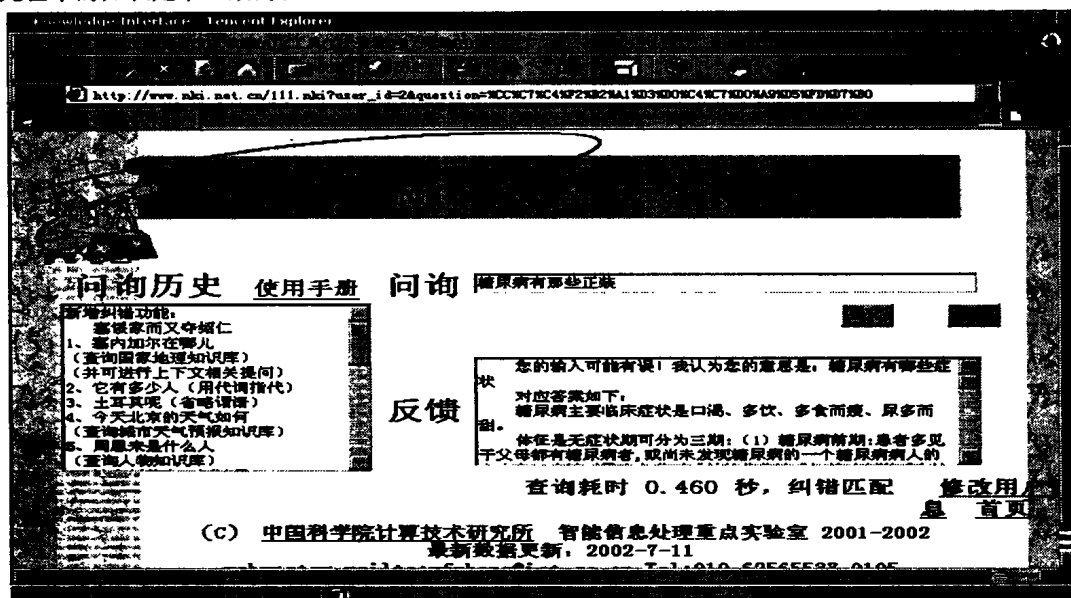


图 3

义,如“糖尿病有哪些症状”这句话,语音识别出来可能会成为“糖尿病有哪些正装”。如果没有知识作为支撑,我们无法从算法上处理这样的歧义。而从知识的角度出发,我们可以将“糖

尿病有那些正装”纠正为“糖尿病有那些症状”,并给出相应的知识。

(下转第 54 页)

$$i=1,2,3;k=1,2,3, \text{且 } k \neq i; \\ r=1,2,\dots,15 \quad (19)$$

每个 S_i 也构造出一个模糊集合 A_i , 由 A_i 和 A_r 构成了对特征空间 Ω 的模糊划分。根据上述算法进行雷达体制的分类实验, 算法中参数设定为: $P_m=0.001$, 进化的代数 = 50, $\beta=1$, $\delta=2$ 。用 75 个训练样本 140 个测试样本进行训练, 由雷达功能与雷达的技术特征参数之间的关系, 经训练提取了 18 条模糊分类规则。表 1 给出了利用上述算法与通常的模糊分类器分类效果的比较, 表 2 是基于 GA 训练的模糊联想分类器对 6 类目标的分类结果。由分类结果可见, 这个模糊联想分类器具有较好的分类性能, 比通常的模糊分类器要好, 并且有较高的识别率。

表 1 算法实验结果

每维的模糊划分数	正确分类率%			
	K=2	K=3	K=4	K=5
基于通常模糊分类器	90.33	92.00	95.12	96.22
基于模糊联想分类器	95.89	97.51	97.95	98.87

表 2 模糊联想分类器雷达辐射源的分类结果(k=3)

雷达种类	识别率%	拒判率%	误判率%
搜索雷达	97.8	0.8	1.4
跟踪雷达	97.3	0.7	2.0
火控雷达	97.4	0.9	1.7
轰炸雷达	97.7	0.9	1.4
导航雷达	96.9	1.0	2.1
气象雷达	98.0	0.2	1.8

(上接第 39 页)

图 3 是根据本体对语音错误的纠正例示, 并在纠正后给出相应的知识^[17]。

结论和进一步的工作 本文从本体论的角度出发探讨了医学知识的获取方法。迄今为止, 我们已建立了一个比较完善的医学本体体系, 并总结出了较多的一些医学知识获取模式以及医学知识获取主体, 形成了一套比较完整的医学知识表示方法和医学知识获取体系。由于医学是一门巨复杂的学科, 要想获取有关医学学科的完全和完整的知识是非常困难的, 这给我们在保证医学知识的完备性方面带来了困难。同时由于在进行形式化本体时由于人的理解以及获取知识的知识源本身都会不可避免地造成获取到的知识在知识库产生不一致, 因此, 我们目前虽然采用了公理来对一致性进行保证。但也很难避免出现由于本体设计时出现的不一致性, 这就给我们提出了如何分析本体的课题, 比如本体设计的合理性, 本体之间的相似性和等价性, 本体层面上的知识的相似性和等价性。

同时, 为了加快医学知识获取的速度, 文本知识自动获取也是我们今后要进行的重要工作。

参考文献

- 1 美国国家医学图书馆. ULMs 项目. <http://www.nlm.nih.gov/research/umls/>
- 2 Bowden P R, Halstead P, Rose T G. Extracting Conceptual Knowledge from Text Using Explicit Relation Markers. In: N. Shadbolt, K. Ohara, G. Schreiber, eds. *Advances in Knowledge Acquisition. Lecture Notes In Artificial Intelligence*, Springer-Verlag, Berlin, 1996, 1076:147~162
- 3 Hahn U, Schnattinger K, Romacker M. *Automatic Knowledge*

小结 模糊联想分类器是通常的模糊分类器的扩展, 是将模糊联想记忆系统用于模式分类的形式构成的。它把规则的后件扩展为模糊集, 使得分类界面的变化范围更大, 因而分类性能要比通常的模糊分类器的性能好。本文采用遗传算法作为模糊联想分类器进化训练方法, 进行模糊分类规则的提取, 不但具有对解空间进行全局搜索能力, 同时解过程不依赖于种群初始值的选择。实验结果表明这样的结合可以得到较好的分类效果。遗传算法是通过优化分类规则的边界值来提高分类精度的, 由于它可以对解空间进行全局搜索优化, 优化后的边界值可以是最优的分类问题划分。

参考文献

- 1 Shin F B, Kil D H, Wayland R F. Active impulsive echo discrimination in shallow water by mapping target physics derived features to classifiers[J]. *IEEE Journal of Oceanic Engineering* 1997, 22(1):66~80
- 2 Ishibuchi H, Hozakik, Tanaka H. Distributed representation of fuzzy rules and its application to pattern recognition. *Fuzzy sets and Syst*, 1992, 52
- 3 黄建军, 等. 正则模糊联想记忆[J]. *电子学报*, 1997, 25(7):68~71
- 4 Simpson P K. Fuzzy min-max neural networks Part 1: Classification[J]. *IEEE Transactions on Neural Networks*, 1992, 3(5):776~786
- 5 Wolpert D H. Stacked generalization[J]. *Neural Networks*, 1992, 5(2):241~259
- 6 周明, 孙树栋. *遗传算法原理及应用*[M]. 国防工业出版社, 1999
- 7 Li Q, Tufts D W. Principal feature classification [J]. *IEEE Transactions on Neural Networks*, 1997, 8(1):155~160
- 8 Stewart C, Lu Y C, Larson V. A clustering approach for high resolution radar target classification [J]. *Pattern Recognition*, 1994, 27(4):503~513

Acquisition from Medical Texts. Text Knowledge Engineering Lab, 1996

- 4 Lu R Q. *New Approaches to Knowledge Acquisition*. World Scientific Publishers, 1992
- 5 曹存根. *面向专家的知识获取*. 北京: 科学出版社, 1998
- 6 Welty C. The Ontological Nature of Subject Taxonomies. In: *Proc. of the First Intl. Conf. (FOIS'98)*, June 6-8, Trento, Italy. 317~327
- 7 Guarino N. Formal Ontology and Information Systems. In: *Proc. of the First Intl. Conf. (FOIS'98)*, June 6-8, Trento, Italy. 3~15
- 8 《中国大百科全书》之医学卷. 中国大百科全书出版社, 1998
- 9 Chaudhri V K, Farquhar A, et al. The Generic Frame Protocol 2.0. [SRI International Technical Report]. 1997
- 10 Cao Cungen. Extracting and Sharing Medical Knowledge. *Journal of Computer Science and Technology*, 2002, 3
- 11 Cao Cungen, et al. Progress in the Development of National Knowledge Infrastructure. *Journal of Computer Science and Technology*, 2002, 17(5)
- 12 张德海, 曹存根, 张宇翔. 国家和城市知识获取与本体论分析. 见: 中国人工智能学会第九届全国学术年会暨中国人工智能学会成立 20 周年庆祝大会, 2001. 366~370
- 13 唐素勤, 曹存根. 智能教学系统: 综述与改进. 见: 中国人工智能学会第九届全国学术年会暨中国人工智能学会成立 20 周年庆祝大会, 2001. 1129~1132
- 14 《疾病和有关健康问题的国际统计分类》第十次修订本. 世界卫生组织
- 15 《医学主题词表》2002. 美国国家医学图书馆
- 16 曹存根. 大百科全书中医学知识的获取. *LNCS*. 2001, 2101:268~271
- 17 丰强泽, 曹存根, 田雯. 关于知识朗读的思考和结果. *CAAI-9*, 2001