基于本体论的民族知识获取与分析*)

王丽丽[1] 曹存根[2] 顾 芳[2] 田 雯[2]

(中国科学院软件所 北京100080)[1] (中国科学院计算所 北京100080)[2]

Acquiring and Analyzing Knowledge of Nationalities: An Ontology-Based Approach

WANG Li-Li CAO Cun-Gen GU Fang TIAN Wen

(The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology Chinese Academy of Sciences, Beijing 100080, China)

Abstract In this paper, we introduce an ontology-based method for acquiring and analyzing knowledge of nationalities. To check for the consistency of the acquired knowledge, we introduce a set of domain-specific axioms of nationalities. These axioms are also used in reasoning and interconnecting with knowledge of different nationalities.

Keywords Nationality ontology, Ontological system, Chinese nationalities, Foreign nationalities, Knowledge acquisition, Knowledge analysis

1. 引言

民族学是一个非常重要的学科,它与语言学、宗教学、地理知识和历史知识等都有密切的联系.许多计算机应用系统都需要大量的民族文化、宗教、历史等知识.例如自然语言理解(词法和语法分析)、民族知识教学系统、民族知识的普及等等都离不开有关民族的基础和专业知识.但在人工智能中,民族知识的获取及知识表示一直得不到足够的重视.Harvard大学有一个ad2000的项目,其中包含了一千多个民族[1].但ad2000项目主要是为基督教的传播服务的,它虽然在尽可能地收录民族,但它对民族的描述主要是针对宗教,并没有给出关于民族的完整的描述.因此,我们迫切地需要建立一个真正意义上的世界民族知识库,尤其在国内,建立一个完善的民族知识库体系更是从事民族学的研究人员极为关注的问题.

一直以来,知识获取都是知识工程的一个公认的瓶颈问题,因此知识获取受到了广泛的重视和研究[2~6,15].知识获取的途径主要有两种,一是从学科专家处获得专业知识,二是从文本或数据库中直接获取.但是,由于专家的研究领域和研究精力的局限,很难给出完整的学科体系,而且据统计90%以上的知识可以从文本中获取,因此,对从事大规模知识获取的人员来说,从文本中直接获取知识无疑是一种更可取的方法.另一方面,由于文本都是以自然语言组织而成的,而自然语言的理解在现阶段仍是计算机科学中的一个难题,所以,要想由计算机自动获取基本上很难实现,因此,这里采用人机交互的半自动的知识获取方法[17,18].

近年来,本体论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,国内陆汝钤等研制的常识知识系统等.

本体论(Ontology)原是一个哲学名词.在哲学意义上,一个本体是指解释一定世界现象的一个特定的系统[7,8].在工程研究中,从知识共享的角度来说,本体论这个名词是对客观存在的概念和关系的明确刻画[7,9~11].

基于以上这些,我们在处理民族知识的时候,引入了本体论的观点,建立了基于本体论的民族知识获取和分析的方法.这里引进本体论是想从学科领域的概念和关系、属性集出发,建立一个便于理解和分析的民族知识结构,并支持满足一致性的民族知识库的开发[10,11].实践证明,这是一套行之有效的方法.

具体来说,工作主要分为四步:(1)建立民族本体;(2)根据民族本体整理文本知识;(3)知识编译和检查;(4)知识分析及知识链接.前面两部分的工作可以由知识工程师在专家的协助下完成,后面的工作则可由计算机处理.具体实现步骤如图1所示.

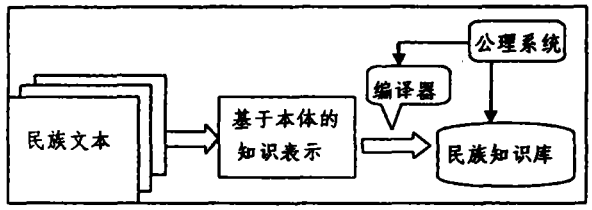

图1 基于本体的民族知识获取方法

下面,文章将对民族本体、知识获取以及知识的一致性分析做概括性的介绍.

2. 民族本体

在建立民族本体之前,必须弄清楚一个概念,就是究竟什么是民族?汉语中"民族"这个词出现得较晚,1903年中国近代资产阶级学者梁启超把瑞士-德国的政治理论家、法学家J.K.布伦奇利的民族概念介绍到中国来以后,民族一词才在

*)本文中的工作受到国家自然科学基金资助(课题号#20010010-A)和科技部重大基础研究专项资助(课题号#2001CCA03000).王丽丽 硕士研究生,主要研究方向:民族知识获取和分析.曹存根 博士生导师,主要研究方向:人工智能.顾 芳 博士研究生,主要研究方向:人工智能中的本体论.田 雯 博士研究生,主要研究方向:常识知识获取和分析.

中国普遍使用[12]。很多人将民族、氏族、部落、种族、国家混为一谈，实际上，民族与这几个概念均不相同。

表1 民族与氏族、部落、种族、国家的区别

民族与氏族、部落的区别	-氏族、部落是以血缘关系为纽带的人们共同体 -民族是以地缘关系为基础的人们共同体
民族与种族的区别	-种族(人种)属于生物学范畴 - 民族属于历史范畴
民族与国家的区别	-国家不一定必须有共同语言，如多民族国家 -民族必须有共同语言 -除了阶级的因素以外，单一民族国家，在语言、地域、经济生活和心理素质方面，民族与国家是基本一致的

根据《中国大百科全书》中关于民族的论述，表1给出了民族与氏族、部落、种族、国家的区别[12]。对于民族的定义不同的人有不同的观点[12]。

1. 布伦奇利认为：民族有8种特质：①其始也同居一地；②其始也同一血统；③同其肢体形状；④同其语言；⑤同其文字；⑥同其宗教；⑦同其风俗；⑧同其生计(经济)。
2. 孙中山认为：形成民族有五个力，第一血统、第二生活、第三语言、第四宗教、第五风俗习惯。
3. 斯大林认为：民族是人们在历史上形成的一个有共同语言、共同地域、共同经济生活以及表现于共同文化上的共同心理素质的稳定的共同体。

图2 民族的不同定义

图2给出了学术界对民族的一些定义，综合各种观点，以斯大林的定义为基础，现在大家比较公认的关于民族的定义是[12]：

民族(nation)：是人们在历史上形成的一种具有共同语言、共同地域、共同经济生活以及表现于共同民族文化特点上的共同心理素质的稳定的共同体，同任何历史现象一样，有其发生、发展、消亡。

图3 民族定义

其中民族的共同语言指不同民族各自使用的语言，是民族内部最重要的交际工具；民族的共同地域指一个民族居住和生活的地域，在地理上联成一片，没有被巨大的自然界线(如海洋、高山等)所分隔，在政治上基本统一，没有长期被国家或其他政治区划所分割；民族的共同经济生活指民族内部的经济联系，是民族形成的决定性条件；民族的共同心理素质亦称“民族性格”，指各民族在形成和发展过程中凝结起来的表现在民族文化特点上的心理状态。民族共同心理素质通过民族的物质文化和精神文化的特点表现出来[12]。

实际上，就像定义中指出的那样，民族有其发生、发展、消亡的过程和规律，尤其在现今全球移民、民族融和等情况的影响下，民族的共同地域，共同语言往往已被打破，单单用共同语言、共同地域、共同经济生活、共同心理素质这四个共同已不足以描述一个民族，即这种定义并没有给出明确的判定一个集团是否是民族的方法，或者说这个定义本身在某种程度上是含糊的，我们希望知道民族有哪些属性、关系，而建立民族本体则可以给出尽可能多的民族方面的属性，以及民族之间的关系，民族与其他对象之间的关系等等。

2.1 民族本体体系

民族本体有两个特别重要的部分，一是本体体系，二是类。首先要给出民族本体的体系，在对民族这个概念进行分类时，有很多方法，如根据地域，根据语言系属、根据人种、根据种族变迁等，这里，我们采用大家最容易接受的一种，即根据民族所处的地域来建立本体体系，其本体形成的树结构如下：根节点为民族；民族的子节点为亚洲民族、非洲民族、欧洲民族、美洲民族、大洋洲民族；各个洲民族的子节点为各个国家民族；各个国家民族的子节点为具体的民族本身，如朝鲜人、波兰人、阿拉伯人等等。(在具体的民族中，可能会出现交叉的情况，如中国的朝鲜族与朝鲜民主主义共和国的朝鲜人及大韩民国的朝鲜人实际上为一个民族，也就是说这里的本体体系并非一个完全的树结构。)

2.2 本体的描述语言

上面已经给出了民族本体的体系，在这个体系中除了叶子节点以外的各个民族集合，都可以作为类(category)。本体描述语言，主要就是对类给出描述。我们的本体语言是一种框

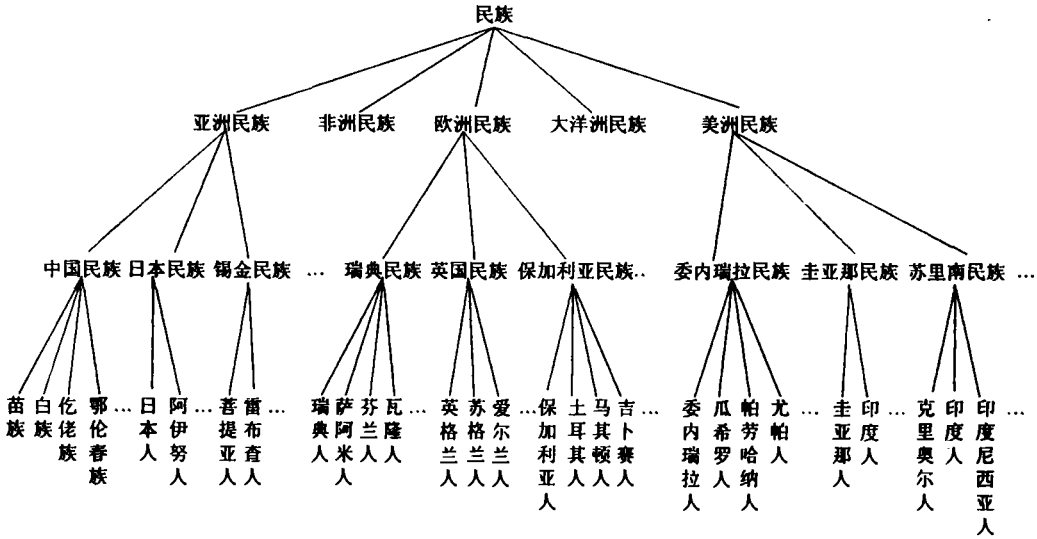

图4 民族本体体系

架语言(Frame Language)，吸收了 Generic Frame Protocol[13] 的设计思想。下面给出描述语言的具体内容：

〈类描述〉::=defcategory〈类名〉[〈相关类〉]

{

〈〈槽定义〉〉

}

在描述语言中，最重要的就是关于槽的定义，这是描述的中心。在图5中给出了槽定义的具体内容，这里主要来看一下槽类型和槽值类型的定义。

槽分为4个类型。

1. 属性槽(简称属性)。属性为名词。属性槽又分为布尔属性槽和非布尔属性槽。布尔属性槽对应于一元谓词。

2. 关系槽(简称关系)。关系为动词。

3. 属关槽(简称属关)。既可为名词又可为关系的槽。例如，“旧称”和“简称”等等。

4. 方法槽(简称方法)。

槽值的类型比较复杂，可以分为简单类型和复合类型。其中，简单类型有：整数、实数、分数、数量、比例数、字符串、时间，等。复合类型有：整数数组、实数数组、分数数组、数量数组、比例数数组、字符串数组、时间数组，等。

```
〈槽定义〉::=〈槽类型〉:〈槽名〉
        :类型          〈槽值类型〉
        [:值域         〈值域〉]
        [:不完全值域   〈值域〉]
        [:模糊值域     〈模糊值域〉]
        [:缺省值       〈值〉]
        [:单位         〈槽值的单位〉]
        [:例子         〈下位槽序列〉]
        [:同义词       〈同义词序列〉]
        [:近义词       〈近义词序列〉]
        [:反义词       〈反义词序列〉]
        [:逆           〈槽的逆〉]
        [:性质         〈槽的局部性质〉]
        [:侧面         〈用户定义的必要侧面序列〉]
        [:注释         〈槽的非形式化说明〉]
```

图5 槽定义及其侧面

另外，在类描述中提到了相关类。这里的相关类主要分为两种，一种是继承关系的相关类，一种是实现关系的相关类：

〈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〉}

〈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3个含义。第一，C_1 可以使用 C_2 中的词汇(即 C_2 中的属性、关系、属关和方法)；第二，C_1 遵循 C_2 中的所有公理；第三，若 P 是 C_1 中的概念，则 P 必然是 C_2 中的概念，例如：上面给出的民族本体体系中，亚洲民族类继承民族类，中国民族作为亚洲民族类中的概念，同时也必然是民族类中的概念。

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。

上面给出了一个本体的描述语言，依据这个语言，可以对本体中的类进行描述。

2.3 民族本体的属性和关系

以上给出了一个比较完整的关于民族的本体体系和本体描述语言，下面主要说明民族的类中所包含的属性和关系，以及对它们进行分析。

2.3.1 *民族及其属性* 民族是一个社会学概念，它涉及的内容比较广，与很多学科都有很密切的联系，因此，它的属性包括人口、分布地、语言、文字、宗教信仰、经济、服饰、饮食等很多方面[12~14,16]。在这里，为了研究的方便，我们把民族的属性分为外延型属性和内涵型属性。其中民族的外延型属性主要指人口、分布地、人种特征等，是描述民族整体和个体的共同的外部性质的属性；民族的内涵型属性主要指民族语言、文字、衣食住行、宗教信仰、婚姻、节日、文化等，是描述民族的整体的内在性质的属性。

(1)民族的外延型属性

民族的外延型属性是用来说明民族的一些基本的外在性质，主要包括如图6所示的几个方面的内容。

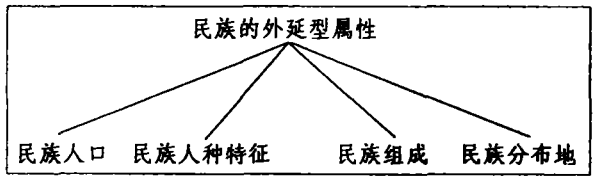

图6 民族的外延型属性分类

在描述任何一个民族时，首先涉及的就是民族的人口数。据统计，全世界人口百万以上的民族共三百多个，约为民族总数的十分之一，但这三百多个民族却占全球总人口的96%以上[14]。因此，人口数多的大民族无论是在经济还是在文化乃至社会生活的各个方面都占据着重要的地位，由此可见，人口数是对民族大小的一种度量。然后，我们会关心民族的外貌特征，即民族属于什么人种。实际上，外貌特征是人种的一种表现形式，可以由外貌特征判定一个人的种族。并且由于人种的分布有一定的规律，如蒙古人种主要分布在亚洲东部，欧洲绝大多数都属于欧罗巴人种等，因此，知道一个民族的人种特征可以相应地推导出民族很多的其他属性。人种特征是民族的一个最明显的外在表现。其次，分布地是民族的另一个外在属性。一个民族生活在哪一州，哪一个国家及居住地环境等方面很大程度上决定了民族的进化、经济文化发展等方面，因此，分布地是对民族的一种空间度量。再次，通常一个民族内部总会由一些分划，比如民族是由一些支系、部落、部族等组成，了解一个民族的组成结构对描述一个民族来说是很重要的。民族组成是对民族的一种结构的度量。从以上分析可以看出，民族的外延型属性是对民族整体的度量，从某些方面来测定一个民族。因此，可以在民族类的基础上，将这些方面的属性抽取出成为民族类的下位类，由民族类来实现。如图7给出了民族分布地属性类的一部分。

```
Defcategory 民族分布地属性
{
  属性:洲属
      :类型  字符串数组
      :注释“民族生活在哪些洲”
  属性:生活国家
      :类型  字符串数组
      :注释:“民族生活在哪些国家，一般情况下用‘主要生活国家’和‘其他生活国家’来说明民族分布在不同国家的人数的多少。若未说明是否‘主要生活国家’，则用‘生活国家’”
  属性:主要生活国家
      :类型  字符串数组
```

:注释"民族所生活的主要国家,相对于'其他生活国家'而言"

属性:其他生活国家

:类型 字符串数组

:注释"民族生活的国家,相对于'主要生活国家'而言"

属性:分布地

:类型 字符串数组

:同义词 居住地

:注释"民族的分布地,一般情况下包括'聚居地'和'散居地'"

属性:聚居地

:类型 字符串数组

:注释"民族集中居住的地方"

属性:散居地

:类型 字符串数组

:注释"民族分散居住的地方"

属性:原住地

:类型 字符串数组

:同义词 旧分布地,和原分布地

:注释"民族以前的分布地,现一般已经改变,并且不包括发祥地"

属性:民族发祥地

:类型 字符串数组

:同义词 发祥地

:注释"民族的发祥地,指民族最开始居住的地方"

……

}

图7 民族分布地属性类(部分)

图7给出了9条属性,是民族分布地属性类的一部分,完整的民族分布地属性类包含32个属性,是关于民族分布地的一个比较完整的描述。相应地,我们也建立了民族人口属性类、民族人种特征属性类和民族组成属性类,用来对民族的外延型属性进行分类描述。

(2)民族的内涵型属性

对任何一个民族来说,经济、文化、心理素质等是更高层次的内容,是属于民族属性的内在部分。它们对分析民族知识及研究整个世界民族的发展演变都有非常重大的意义,因此,如何很好地描述民族内涵型的属性就显得更加重要。同样对民族的内涵属性进行分类,具体内容如图8。

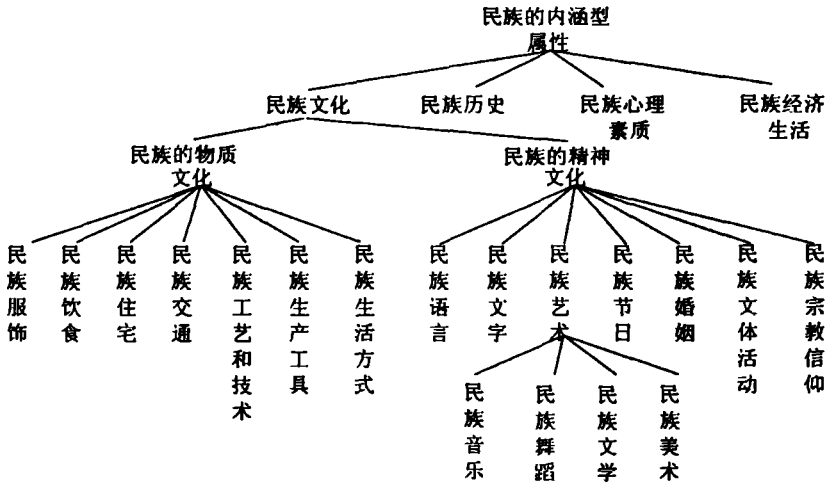

图8 民族的内涵型属性分类

民族的内涵型属性由22个部分组成,每部分都代表了民族的一个重要的方面。在上面给出的关于民族的概念中,民族是由四个共同来描述的,而这里,语言,经济,心理素质都属于民族的内涵型属性,足可见民族内涵型属性的重要性。但是,这22个部分并不是完全并列的,它们之间有一定的层次关系,为了更好地对民族的内涵型属性分类,我们将其表示为一个树的结构,具体如图8。需要说明的是,在图8中,"民族的内涵型属性"、"民族物质文化"、"民族精神文化"和"民族艺术"是虚节点,没有具体内容。由图8可以看到,民族的内涵型属性分为四个部分,即民族文化、民族经济生活、民族心理素质和民族历史,这是四个并列的部分,从四个方面对一个民族的内在属性作出了描述。这种划分是根据民族的概念并加上一些适当的整理,将民族概念中的语言归入文化,又将民族历史单独列出;由于文化可以分为物质文化和精神文化两部分,因此民族文化被分为"民族物质文化"和"民族精神文化";另外,作为生活中基本组成部分的衣食住行是物质文化的最基本表现,被划入"民族物质文化",同时民族工艺和技术、民族生活方式和民族生产工具也都是民族物质文化的基本组成;这里,民族语言和民族文字被归入"民族精神文化",作为精神文化的重要组成部分;另外,一个民族的艺术、节日、婚姻、文体活动和宗教信仰也都是"民族精神文化"的组成部分,它们从各个不同的方面对民族的精神文化作出描述;最后,民族艺术主要包含文学、音乐、舞蹈、美术四个部分。

民族宗教信仰属性类一共包括了与宗教信仰有关的26个属性,是对宗教信仰的比较详尽的说明,在图9中给出了其中的11条。

defcategory 民族宗教信仰属性

{

属性:宗教信仰

:类型 字符串数组

:注释"民族在宗教方面的信仰,若未说明是否'主要宗教信仰',则用'宗教信仰'"

属性:主要宗教信仰

:类型 字符串数组

:注释"民族的宗教信仰,若大多数人信仰的宗教,则为主要宗教信仰。相对于'其他宗教信仰'而言"

属性:其他宗教信仰

:类型 字符串数组

:注释"民族的宗教信仰,相对于'主要宗教信仰'而言"

属性:崇拜形式
:类型 字符串数组
:同义词 民族崇拜
:注释"民族崇拜的形式,主要崇拜形式为自然崇拜,图灵崇拜和祖先崇拜"
属性:崇拜对象
:类型 字符串数组
:注释"民族崇拜的对象,如自然,祖先,神灵等"
属简:信仰内容
:类型 字符串数组
:注释"民族的宗教方面的信仰内容,如对一些神灵的观点等"
属性:祭祀对象
:类型 字符串数组
:注释"民族祭祀的对象,如神灵、祖先等"
属性:民族祭祀活动
:类型 字符串数组
:注释"民族的祭祀活动,如阿伊努人的民族祭祀活动:熊祭,和鲑祭"
属性:宗教神职人员
:类型 字符串数组
:注释"宗教的神职人员,如巫师、祭司、教士等"
属性:宗教仪式
:类型 字符串数组
:注释"民族与宗教有关的一些仪式。如塔希提人的宗教仪式:祭祀仪式,和祈祷仪式"
属性:宗教习俗
类型 字符串
:注释"民族有关宗教的一些风俗习惯,如掸人的宗教习俗:小孩不分男女,至五、六岁时要到村庙学习"
}

图9 民族宗教信仰属性类

2.3.2 *民族及其关系* 民族的类中定义了一些关系,用来描述民族与其他概念之间的关系。关系通常表示一个命题或断言。关系在类中是很重要的一个部分,它将一个概念与其他的概念和对象联系起来,起到了一定的知识联通的作用。

表2 民族类中的关系

关系	注释
自称	本民族对自己的称呼
旧译	民族以前的译称
泛指	民族的更广泛的含义,一般相对于'特指'而言
特指	民族的更具体的含义,一般相对于'泛指'而言
嗜好	民族的特殊的爱好,通常指饮食方面。如拉祜族嗜好:饮烤茶
喜欢	民族的喜好,如像族喜欢:独家独院,应用范围比嗜好广
被誉为	对民族的誉称,如蒙古族被誉为:草原骄子
擅长	民族所擅长的行为,如艾马克人擅长:纺织毛毯
受影响	民族受某些文化或民族、国家、地区的影响。如岱人受影响:中国文化
相似于	一个民族在某方面相似于另一个民族
相同于	一个民族在某方面相同于另一个民族

表2中给出了现已找到的民族类中的一些关系,共11条。有的关系带有侧面和同义词,例如关系"相似于"的侧面为"方面"表示相似是在某些方面的相似,同义词包括"相近于"、"近似于"、"类似于"等。

另外,表2中的"自称"和"旧译"既可以是关系也可以作为属性,是我们前面提到的属关。

需要说明的是,在属性类中,包含了少量的关系,这些关系只适用在某一方面,例如"信奉"、"祭祀"类似的关系,只适用于宗教信仰,因此,这些关系也整理到相应的属性类中,上面所给出的关系不包括属性类中的关系。

2.3.3 *侧面* 为了更精确全面地描述关系和属性,本体的类中定义了很多侧面,这些侧面是关系或属性的进一步补充。例如时间、可信度、依据、原因、结果、方面、范围等。与这些侧面相结合能够使关系和属性的刻画更完整,更加有效地提高知识的质量。侧面在知识的完整性维护方面有着重要的作用。如上面所列举的关系"相似于",就包含了一个侧面:方面,这就使相似于这个关系有了一个完整的表述。

2.4 民族的类

为了进一步说明本体设计的思想,下面给出了民族的类的部分定义。这里每一个属性都有其特定用法:例如含义是一个属性;它的取值的类型为字符串数组;侧面为语种;表示某个民族的名称在某个语言中所代表的意义。

在图10中给出了民族类的一部分。由于篇幅所限,没有办法将民族类给全。民族类包括民族称谓,风俗习惯,民族特点三个方面,实现了上面所讲的26个属性类。在整个民族类(包括各个属性类)中,共包括280条属性,42条关系,建立了一个关于描述民族的比较完整的本体体系。当然,并不能说民族本体没有遗漏,由于民族知识的博大精深,现有知识源的限制,和民族作为概念的可变性等,要想将民族的所有属性及与民族有关的所有关系找全是不可能的。但是,已经建立了一个很好的民族本体的基础(即民族本体体系和民族类中的属性关系集合),结合一个好的公理库,就可以不断地将民族本体补充完全。

defcategory 民族继承万物;实现民族人口属性,和民族人种特征属性,和民族历史属性,和民族分布地属性,和民族组成属性,和民族语言属性,和民族文字属性,和民族服饰属性,和民族饮食属性,和民族建筑属性,和民族交通属性,和民族生活方式属性,和民族经济属性,和民族宗教信仰属性,和民族婚姻属性,和民族节日属性,和民族文体活动属性,和民族音乐属性,和民族文化属性
{
属关:自称
:类型 字符串数组
:注释"本民族对自己的称呼"
属关:旧译
:类型 字符串数组
:注释"民族以前的译称"
属性:正式定名时间
:类型 时间
:注释"民族的正式定名时间,应晚于命名时间"
属性:族名意义
:类型 字符串数组
:注释"民族的名称的意义,即对民族名称的解释"
属性:族名来源
:类型 字符串数组
:注释"民族名称的来源"
属性:含义
:类型 字符串数组
:侧面 语种
:注释"民族族名在某语种下的含义"
属性:民族习俗

:类型 字符串数组
:注释"民族的一些风俗习惯"
属性:丧葬习俗
:类型 字符串数组
:注释"民族的丧葬习俗,即关于死者如何埋葬如何出殡等方面的习俗"
属性:民族忌语
:类型 字符串数组
:注释"民族所忌讳的话语,如京族的民族忌语:饭烧焦"
属性:记事方式
:类型 字符串数组
:注释"民族的记事方式,如基诺族的记事方式:刻竹木"
关系:嗜好
:类型 字符串数组
:注释"如拉祜族嗜好:饮烤茶"
属性:民族特点
:类型 字符串数组
:注释"民族的特点,在文化,社会结构等方面的一些基本特点"
}

图10 民族类

3. 民族知识的获取

由于这里采用的是从文本中直接获取知识的方法,文本的来源就显得尤为重要。现今社会,科技资讯发达,各个方面的知识源层出不穷,但由于大多数的文本都有这样或那样的缺点,或是因为知识覆盖面不够,或是因为时间的关系,有些知识已经过时,因此,不是所有的民族知识文本都可以作为数据源。为了保证知识的准确性,文本的选择极为重要。这里,主要以中国大百科全书民族卷为基础,结合中国五十六个民族知识库和《世界民族通览》等加以整理。

得到可靠的文本之后,就是如何进行形式化的问题。前面已经建立了民族本体,根据本体即可以由知识工程师在专家的指导下对文本进行整理。由于文本知识的获取涉及自然语言理解,因此需由知识工程师手工工作,并且,这其中还涉及到民族学的专业知识的问题,故专家的指导也是必不可少的。对整理后的文本,可以用知识编译器根据给定的本体进行编译和检查,并进行知识分析和知识链接。下面是具体的形式化的例子:

```
defframe 巴塔克人:民族
{
  汉语拼音:batakeren
  英文:Bataks
  主要生活国家:印度尼西亚
  人口数:大约3400000
        :时间1978年
  主要分布地:苏门答腊岛的中部山地,和苏门答腊岛的北部山地
  主要聚居地:多巴湖的周围地区
  人种:蒙古人种马来类型
  起源:原始马来人的后裔
  民族体质特征:一般身体矮小,皮肤呈暗棕色
  本民族语言:巴塔克语
  民族分支:代里人,和帕克帕克人,和托巴人,和卡罗人,和曼代林人,和昂科拉人
        :根据方言,和文化
  民族分支数:5
  宗教信仰:伊斯兰教逊尼派,和基督教
  ……
}
```

图11 民族形式化举例(待续)

```
defframe 德昂族:民族
{
  生活国家:中国
  是一个:中国的少数民族
  是:西南边疆的最古老的民族
  旧称:崩龙族
      :时间段 新中国成立后至1985年
  正式定名时间:1985年
      :依据 德昂族人民的自愿
  人口数:15462
  杂居民族:汉族,和傣族,和景颇族,和佤族,和傈僳族等
  传统工艺:制造银器
  民族节日:德昂族泼水节,和关门节,和开门节
  饮食习俗:喜吃酸辣,和嗜喝浓茶
  宗教信仰:小乘佛教
  饰品:绒球,和腰箍
  青年男女首饰:银项圈,和耳筒,和耳坠等
  ……
}
```

图11 民族形式化举例

图11中给出了两个民族的例子,用来说明民族知识进行形式化的框架形式。其中下面给出的是一个印度尼西亚的民族,上面是一个中国的少数民族,利用本体的类中的属性和关系,对这两个民族做出了比较详尽的描述,从以上的形式化中,可以很清晰地得到一些感兴趣的民族的资料,如民族的人种,宗教信仰等。另外,在形式化的过程中,专家是必不可少的,只有在专家的适当的指导下,才能保证我们形式化的结果。

世界上一共有多少个民族,一直是令民族学家比较困扰的问题,根据多数学者的估计,当今世界上约有大小民族两三千个,其中,人口上百万的大民族有三百多个[14],由于资料的限制,只能对其中的一些民族进行形式化。到目前为止,我们已经形式化了大约800个民族,整理出了将近一万二千条的知识,建立了一个比较大的民族知识库。

4. 知识的一致性分析

在建立了民族本体和民族知识库之后,还有一个非常重要的任务,就是对知识进行检查和推理,而这些仅依据现有的本体是无法实现的,这就需要建立一个满足一致性的公理库,再根据这些公理对知识库中的知识进行检验和推理。实际上,这里讨论的本体还应该包含公理集合。一个本体建得好不好很大程度上取决于它的公理库建得如何。

严格地说,公理本身也是知识,而且应该是更高层次的知识。公理主要是为了保证知识之间以及知识库与本体之间的一致性,这里所说的一致性就是没有矛盾。另外,还可以利用公理对知识进行推理和利用公理实现知识之间的联通。当然,后面的两个方面从本质上讲也是为了保证知识之间的一致性。这里,公理中所使用的语言是经过定义的 NKI 一阶语言,公理中用到的各种符号也经过了严格定义。在从以上三个方面对公理进行举例并说明之前,先对 NKI 一阶语言中的函数

和谓词举例并说明。

表3 NKI一阶语言中的一些函数和谓词

谓词或函数	含义
上取值(A)	函数:取小于或等于A的最大整数
下取值(A)	函数:取小于或等于A的最小整数
加(A,B)	函数:求A和B的和
基数(A)	函数:本A的取值集合中含有的元素的个数
百分比(A,B)	函数:求A占B的百分比
交集(A,B)	函数:求集合A和B的交集
大于(A,B)	谓词:A的值大于B的值
等于(A,B)	谓词:A的值等于B的值
属于(a,A)	谓词:a属于A
!属于(a,A)	谓词:a不属于A
同义(A,B)	谓词:A与B的含义相同
近义(A,B)	谓词:A与B的含义相近
反义(A,B)	谓词:A与B的含义相反

我们使用"所有X:〈类〉"表示X是一个类变量,X的取值是民族类中的任意一个民族实例;用"存在X:〈类〉"表示X是一个类变量,X的取值是民族类中的一个民族实例。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。

4.1 公理库对错误和不一致的检验

首先,要保证知识库的知识之间及知识与本体之间是没有矛盾的,即保证一致性。在建立知识库的时候可能由于人为的疏忽或文本错误等原因出现一些矛盾或错误(例:一个民族的城市居民比例大于1),这些问题若没有相应的公理进行检查,单靠知识库本身和已有的本体是无法发现的,因此要建立起一些公理,用来对属性和关系的取值作出一些限制,从而保证知识的一致性,例如表4中的公理1就是对民族的城市居民比例这个属性的值域作出了限制,从而将其取值规定在合理的范围内,这实际上是所有百分比数的一个常识,因此,可以将这条公理进行扩展,使之可以应用于所有的属性值为百分比数的属性。

表4 公理举例(1)(公理的表示方式与表5、6统一)

公理1	所有X:民族,上取值(X·城市居民比例)=1,下取值(X·城市居民比例)=0
公理2	所有X:属性,是一个(X,百分比)→X·值域=(0,1)

公理2说明了若属性X是一个百分比,则X的取值范围在0到1之间。实际上,本体中的每一条属性和关系都需要很多条的公理来支持,才能保证属性和关系的正确使用和不出现矛盾,从这种意义上来看,公理库比属性关系集更加重要,它为属性和关系的使用提供了基础。我们知道民族一般都包含一些支系,下面表5中的20条公理都是与民族分支有关的,它们从各个方面对民族分支方面的属性作出了限制。

表5中一共列举了20条公理。下面对其中几条公理进行解释。公理3表示如果民族有民族分支数,民族本支数,民族旁支数三个属性,则,民族分支数等于民族本支数加上民族旁支数;公理5表示如果一个民族的属性"民族分支"的值包含"等"的字样,则说明属性"民族分支数"的值大于属性"民族分支"的基数;公理6表示如果一个民族的属性"民族分支"的值不包含"等"的字样,则"民族分支数"的值应等于"民族分支"的基数;公理15表示一个民族的"民族分支数"的值应大于这个民族的"主要民族分支数"的取值;公理19表示一个民族的"民族分支"的值应包含"主要民族分支"的取值(由于公理很多,不能一一解释,但基本含义差不多)。

表5 公理举例(2)

公理3	所有X:民族,等于(X·民族分支数,加(X·民族本支数,X·民族旁支数))
公理4	所有X:民族,等于(X·民族分支数,加(X·主要民族分支数,X·其他民族分支数))
公理5	所有X:民族,属于(等,X·民族分支)→大于(X·民族分支数,基数(X·民族分支))
公理6	所有X:民族,!属于(等,X·民族分支)→等于(X·民族分支数,基数(X·民族分支))
公理7	所有X:民族,属于(等,X·主要民族分支)→大于(X·主要民族分支数,基数(X·主要民族分支))
公理8	所有X:民族,!属于(等,X·主要民族分支)→等于(X·主要民族分支数,基数(X·主要民族分支))
公理9	所有X:民族,属于(等,X·其他民族分支)→(X·其他民族分支数,基数(X·其他民族分支))
公理10	所有X:民族,!属于(等,X·其他民族分支)→等于(X·其他民族分支数,基数(X·其他民族分支))
公理11	所有X:民族,属于(等,X·民族本支)→(X·民族本支数,基数(X·民族本支))
公理12	所有X:民族,!属于(等,X·民族本支)→等于(X·民族本支数,基数(X·民族本支))
公理13	所有X:民族,属于(等,X·民族旁支)→(X·民族旁支数,基数(X·民族旁支))
公理14	所有X:民族,!属于(等,X·民族旁支)→等于(X·民族旁支数,基数(X·民族旁支))
公理15	所有X:民族,大于(X·民族分支数,X·主要民族分支数)
公理16	所有X:民族,大于(X·民族分支数,X·其他民族分支数)
公理17	所有X:民族,大于(X·民族分支数,X·民族本支数)
公理18	所有X:民族,大于(X·民族分支数,X·民族旁支数)
公理19	所有X:民族,真包含(X·民族分支,X·主要民族分支)
公理20	所有X:民族,真包含(X·民族分支,X·其他民族分支)
公理21	所有X:民族,真包含(X·民族分支,X·民族本支)
公理22	所有X:民族,真包含(X·民族分支,X·民族旁支)

很容易看出,表5中的公理所涉及的属性都是关于分支和分支数的(其中,民族本支和民族旁支也都是民族分支),并且看起来都比较简单,可以用来检验比较明显的错误和不一致。但是公理的用处不只如此,我们还希望将其用于推理,下面介绍公理在知识推理中的应用。

4.2 公理库在知识推理中的应用

这里,依然以民族分支为例,来说明一些比较复杂的公理。显然,民族与它的分支之间会有许多的共性,但是这些共性在框架中不一定会有描述,如果我们可以针对这些提供公理,就可以由已知的民族的知识推出未知的民族分支的知识。

表6 公理举例(3)

公理23	所有X:民族,所有Y:民族,属于(Y,X·民族分支)→小于(Y·人口数,X·人口数)
公理24	所有X:民族,所有Y:民族,属于(Y,X·民族分支)∧X·人种=(a)→Y·人种=(a)
公理25	所有X:民族,所有Y:民族,属于(Y,X·民族分支)∧X·分布地=(a)→包含于(Y·分布地,(a))
公理26	所有X:民族,所有Y:民族,属于(Y,X·民族分支)∧X·语言=(a)→包含于(Y·语言,(a))
公理27	所有X:民族,所有Y:民族,所有Z:民族,属于(Z,Y·民族分支)∧属于(Y,X·民族分支)→属于(Z,X·民族分支)

表6中列举了5条公理，都可以用来推理，首先要说明，在民族知识库中，依据现在比较通用的看法，将民族的分支也作为一个民族。公理23说明如果民族Y是民族X的分支，则Y的人口数小于X的人口数，这样就算知识库中没有Y的人口数这条知识，也可以根据X的人口数得到Y的人口数的取值的大致范围，甚至如果知道X的其他分支的人口数，还可以用减法计算出Y的人口数(这也可以作为一条公理)。公理24说明若民族X的人种是(a)，则X的分支Y的人种也是(a)，这样，就可以根据X的人种推出Y的人种。公理25说明如果民族Y是X的分支，则Y的分布地一定包含在X的分布地内部。公理26与25类似，但说明的是两个民族语言方面的关系，即若Y是X的分支民族，则Y所使用的语言一定包含在X使用的语言里。公理27则说明了民族分支这个属性的传递性。我们看到，从公理23～公理26都是已知民族的有关知识，推出民族分支的一些知识，实际上，若已知一个民族的分支的某些知识，也可以推出这个民族的一些相关知识。这样，就可以应用这种推理的能力，对知识库中的知识进行补充，并且可以从民族这个学科本身出发，对学科知识进行分析，例如研究民族之间的关系等。

4.3 公理库在知识联通中的应用

上面已经说明了公理库如何保证知识的一致性和进行推理，下面看一下如何利用公理库来实现知识之间的联通。

首先我们来看图12所示的一个框架的例子。

```
defframe 回族:民族
{
        全称:回回民族
        人口数:8602978
              :时间2000年
        主要聚居地:宁夏回族自治区
}
```

图12 回族框架(部分)

我们知道，实际上回族与回回民族是两个等价的概念，也就是说回回民族的人口数应该等于回族的人口数，但是，这种对于人来说是显然的结论，计算机是不可能得出的，因此 需下面的公理28。

表7 公理举例(4)

公理28	所有X:民族，所有Y:字符串，全称(X,Y)→同义(Y,X)
公理29	所有X:民族，所有Y:语言，X·语言＝Y→X·语言·系属＝Y·系属
公理30	所有X:民族，宗教信仰(X,伊斯兰教)→信奉(X,安拉)
公理31	所有X:民族，属于(忏悔节,X·节日)〈--〉属于(基督教,X·宗教信仰)
公理32	所有X:民族，所有Y:岛屿，所有Z:国家，属于(Y,X·分布地)∧所属国家(Y,Z)→属于(Z,X·生活国家)

公理28中说明若Y是民族X的全称，则Y与X的含义相同，这样X除全称这条的所有知识都可以应用到Y上，例如回回民族的主要聚居地＝回族的主要聚居地＝宁夏回族自治区。这样就用公理将两个等价的概念连通起来。

另外，由于民族学与很多其他学科关系密切，因此民族知识库也需要与其他学科的知识库联通起来，这样才能使我们的系统更加完整，比如当有人想知道某个民族所使用的语言是属于哪个语系时，就需要调用语言知识库中的知识，利用公理29我们就可以实现这个调用，即若民族X所使用的语言是Y，则X的语言的系属等于Y的系属，类似的关于语言还有许多的公理，这里不再详述。另外，宗教也是一个与民族密切相关的学科，大家都知道伊斯兰教的教徒信奉的唯一对象是真主安拉，因此，我们有了公理30，若一个民族的宗教信仰是伊斯兰教，则这个民族信奉安拉，类似的还有，若一个民族信仰基督教，则这个民族信奉上帝等。另外，还可以将民族与节日知识，宗教知识联系在一起，例如公理31，若一个民族的节日中有忏悔节，则这个民族的宗教信仰会包含基督教，反之亦然。公理32将民族与地区联系起来，若一个民族X的分布地包括某个岛屿Y，而这个岛屿Y又属于某个国家Z，则X的生活国家包含Z。上面的公理都是有利于知识的连通，通过这些公理，可以将民族内部的概念联系起来，还可以将民族知识与其他学科的知识联系起来，从而有利于知识的共享。

上面，分三部分介绍了我们在本体中建立的公理库，实际上，所有的公理都是用于知识的一致性分析的，由于篇幅限制，不能给出全部的公理库，只能给出一个大概的描述。下面来看一个知识分析的实例，在图11的形式化例子中，德昂族的属性"人口数"没有给出侧面"时间"，这一条知识是不完整的；另外，巴塔克族形式化框架中有这样一段如图13所示。

```
defframe 巴塔克人:民族
{
    ……
    民族分支:代里人,和帕克帕克人,和托巴人,和卡罗人,和曼代林人,和昂科拉人
        :根据方言,和文化
    民族分支数:5
    ……
}
```

图13 巴塔克人框架(部分)

图13中，属性"民族分支数"的值为5。但是，如果细心一点就会发现，属性"民族分支"的值中含有6个元素，这是一个很明显的错误，可是仅从知识库本身来说，对这种错误是无能为力的，但由前面给出的公理6，就可以让计算机自动地检验出这样的错误，以保证知识的一致性。

总结 前面提到了许多一致性的问题，在建立知识库的时候，一致性是首先需要保证的，实际上，一致性分为两个方面，一是公理之间的一致性，这一点基本上不会有问题，因为在建公理时，首先就会保证公理之间没有矛盾；二是公理库与知识库之间是一致的，相对来讲，这一点更为重要，因为建立公理库的主要任务就是保证知识的一致性。但是，有一个很重要的问题就是我们所建立的一致性公理是否是完备的，完备性是所有的知识库都希望满足的性质，但由于知识源等方面的限制，我们不可能保证将所有的知识都收入知识库中，这基本上是一种理想的情况，很难达到。在很多知识库，例如Cyc[5]是不考虑完备的，将这个问题尽可能地回避，我们一直在考虑完备的问题，但也只能说希望我们建立的公理库是尽可能完备的。

这篇文章主要介绍了一种应用本体论进行民族知识获取和分析的方法，从民族学本身出发，建立了能够充分对民族进行描述的民族本体，这个本体主要包括本体体系、类和公理集，其中类中提供了属性关系集，主要是用来对民族知识进行获取，形成框架形式的民族知识库，而公理库则用来对已经获取的知识进行分析，从而确保知识的一致性。实际上，我们在

(上接第81页)

明笔划轮廓模板中包含的笔划轮廓数目随着字形样本的增加趋于稳定;三组字形都获得了较大的压缩比。在图6中给出了三个字形的还原结果,通过与原字形的对比可以看出字形还原的质量比较高。

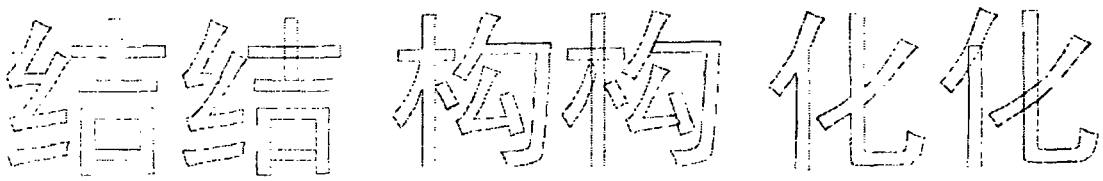

图5 字形还原结果

结论 汉字字形结构式压缩方法利用了汉字字形的特点,能够充分压缩汉字字形中的重复信息。对汉字轮廓字形进行结构式压缩,将字形轮廓信息转换为笔划轮廓模板和字形结构信息,可以在轮廓字形压缩的基础上进一步压缩字形的存储空间,并且得到高质量的还原字形。基于笔划抽取和聚类的结构式压缩方法充分利用了现有轮廓字形的笔划形状信息,简化了组件模板的生成,实现了结构式字形的自动生成,提出了汉字字形结构式压缩技术研究的新方向。我们设计实现的结构式字形计算机辅助设计系统适用于各种的系统平台和字形数据格式,能够实现轮廓字形数据到结构式字形数据的自动转换。它有效地提高了字形设计和制作的工作效率,为汉字结构式字形的应用提供了基础。我们的实验结果证明了汉字字形结构式压缩算法可以获得较高的数据压缩率和高质量的字形还原效果。汉字字形结构式压缩方法同样适用于与汉字字形有相同特点的日文和韩文,因此我们正在将结构式字形压缩的研究成果推广到CJK(中文、日文和韩文)字型技术。

参考文献

1 宋晓丹,罗予频. Outline字体结构式压缩算法及其实现. 中文信息学报,2002,16(3):52~57
2 彭寿全,黄可. 汉字信息处理. 成都:电子科技大学出版社,1994
3 孙星明,胡运发,等. 结构化汉字信息处理. 长沙:国防科技大学出版社,2001
4 赵恒,金通光,王国瑾. 骨架汉字字形存储与显示技术. 中文信息学报,1997,11(1):37~43
5 樊建平. 智能汉字字形设计技术及一个试验性系统ICCDS. 中文信息学报,1990,4(3):1~11
6 Ma Xiaohu, Pan Zhigeng, Zhang Fuyan. Automatic generation of Chinese Outline font based on stroke extraction. Journal of Computer Science and Technology, 1995, 10(1): 42~52
7 Lim Soon-Bum, Kim Myung-Soo. Oriental character font design by a structured composition of stroke elements, Computer Aided Design, 1995, 27(3): 193~207
8 Knuth D E. The METAFONT book. Addison-Welsey, Reading Mass. 1986

(上接第54页)

这方面已经做了很多的工作,建立了民族类和一个包含800个民族左右的民族知识库,已经完成了比较基本的工作。但是,本体还需要进一步的完善,首先属性和关系集还不完全,我们会在今后的工作中不断将其补充完整;其次,也是非常重要的一点,公理库要添加更多的公理,以保证知识的一致性;另外,民族知识库还缺少很多民族的知识,就是已有的民族的知识也需要不断的补充和更新,我们可以通过在网上进行知识挖掘等方式,来补充我们的知识库。

参考文献

1 Harvard大学. ad2000项目. 网址 http://www.ad2000.org
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 R, Schnattinger K, Romacker M. Automatic Knowledge Acquisition from Medical Texts. Text Knowledge Engineering Lab, 1996
4 Lu R Q. New Approaches to Knowledge Acquisition. World Scientific Publishers, 1992
5 Cycorp. Features of CycL. 网址 http://www.cyc.com
6 曹存根. 面向专家的知识获取. 北京:科学出版社,1998
7 Welty C. The Ontological Nature of Subject Taxonomies. In: Proc. of the First Intl. Conf. (FOIS'98), June 6-8, Trento, Italy. 317~327
8 Guarino N. Formal Ontology and Information Systems. In: Proc. of the First Intl. Conf. (FOIS'98), June 6-8, Trento, Italy. 3~15
9 Guarino N, Welty C. Ontological Analysis of Taxonomic Relationships. In: Intl. Conf. on Conceptual Modeling. Springer-Verlag LNCS Vol. 1920, Oct. 2000. 210~224
10 Guarino N, Welty C. A Formal Ontology of Properties. In: Proc. of EKAW-2000: The 12th Intl. Conf. on Knowledge Engineering and Knowledge Management. Spring-Verlag LNCS Vol. 1937, Oct. 2000. 97~112
11 Smith S, Mark D M. Ontology and Geographic Kinds. Proceedings of the International Symposium on Spatial Data Handling (SDH'98), Vancouver, Canada, July, 1998. 12~15
12 《中国大百科全书》之民族卷. 中国大百科全书出版社
13 Chaudhri V K, Farquhar A, et al. The Generic Frame Protocol 2.0: [SRI International Technical Report]. 1997
14 赵锦元,戴佩丽. 世界民族通览. 中央民族大学出版社,2000
15 Cao Cungen. Extracting and Sharing Medical Knowledge. Journal of Computer Science and Technology, 2002, 3
16 任新建. 略论中国民族关系史上的文化交流和整合. 中国传统文化网中华文化研究通讯栏目导航,1999(7)
17 张德海,曹存根,张宇翔. 国家和城市知识获取与本体论分析. 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
18 唐素勤,曹存根. 智能教学系统:综述与改进. 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