第 33 卷 第 7 期 计 算 机 工 程 2007 年 4 月
Vol.33 No.7 Computer Engineering April 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)07—0190—02 文献标识码：A 中图分类号：TP18

基于领域本体网络模型的知识获取技术

张德政，庄洪波
（北京科技大学信息工程学院，北京 100083）

摘 要：知识获取一直是人工智能的瓶颈，如何有效地从文本中提取知识是知识工程所关注的问题。该文提出并构建了领域本体网络模型，将其用于中医领域文本的知识获取，分析了领域本体的数据结构、本体概念的实例化以及语义场的结构与组织方法。基于中文信息处理技术，提出了获取文本知识的框架，实现了原型系统，并用于中医医案知识的获取，取得了较好的效果。

关键词：本体；语义场；分词；知识获取

Technique of Knowledge Acquisition Based on Domain Ontology Network Model

ZHANG Dezheng, ZHUANG Hongbo
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083)

【Abstract】 The knowledge acquisition is a bottleneck of the artificial intelligence. How to extract knowledge effectively from the text is what knowledge engineering is concerned. This thesis has advanced network mode on the domain of ontology and applies this to get text knowledge in the herbalist doctor field. It discusses the domain ontology some attributes such as the describer of computer's data structures, ontology concept's instantiation and the organize method in semantic-field. On the basis of chinese information processing technology, it puts forward the framework about how to get text knowledge, and realizes prototype system. It applies this to get the knowledge about herbalist doctor case further, and gets a good effect.

【Key words】 Ontology; Semantic-field; Word segmentation; Acquisition knowledge

当代名老中医的学术思想和临床经验的获取是中医理论传承与发展的关键，也是中医信息处理亟待解决的问题，并归结为如何从名老中医的学术著作、大量医案中有效地获取知识。对大量医案的深层次挖掘是获取部分知识的方法。对全国多位名老中医治疗肝病的医案分析可以看出，相似或相近疾病或病证的数量较少，通过挖掘来获取小样本医案的难度较大。由于名老中医经过多年的理论与临床实践，其知识结构稳定、中医理论与知识完备，最能反映中医学的思辨规律与特点，单份医案包含着丰富的知识，因此从单份医案文本获取知识的应用与理论研究价值十分重大。

基于应用与理论研究需求驱动，本文提出了基于中医领域常识的文档知识获取技术，并研发了相应的软件系统——KATCM(Knowledge Acquisition of TCM)，该系统以领域本体知识库为基础，建成网状数据结构模型。文本进行词法分析后，抽取出特征词，在对特征词进行术语概念标准化基础上，对领域专家的知识进行实例化，获取单文档知识，组织形成语义场。

1 中医知识及知识获取

中医理论是一个完整的知识体系，包含中医辨证论治各个方面知识，同时其知识结果复杂。独特的学术思想是医者长期读书、临证、思考的经验结晶，值得挖掘整理，使之升华为理性的东西。例如，汉代著有《伤寒杂病论》的张仲景，开辨证论治之先河，树外感热病治疗之圭臬。清代创立温病论的叶天士，以伤寒主六经，温病主卫气营血，形成两门学问。于是，引起伤寒和温病 2 个学派长期论争。再如，邓铁涛教授提出“五脏相关说”，主张其可取代五行学说。认为人体是以五脏为中心的内外相联系的统一体。在生理情况下五脏系统之间互相促进和制约，协调机体的正常活动；在病理情况下，五脏系统互相影响。用五脏相关说取代五行学说，既可以继承五行说的精华，又可赋予现代系统论的内容，体现中医整体观的理论。这些学术思想反映了当代中医的学术水平，促进了中医学术的进步和发展。

知识获取过程都是以原有的知识为基础来获取新的知识。用一种清晰的方法把领域知识分解为一组知识元以及它们之间的相互关系，这些知识元和相互关系组织在一起就构成了该领域的本体。以中医领域本体知识为基础，构成了中医领域问题的理论框架。根据这个框架去获取知识，即可生成一个具体的知识库模型。文献[1]在中医领域通过对领域概念和概念关系组织成本体，即生成了知识获取需要的原有知识库。在文档分析和处理时，需要组织成相应的数据结构，这里引入图论的方法。一个有向图是一个三元组(V,E,f)，其中 V 是一个非空的集合，它的元素称为有向图的节点，E 是一个集合，它的元素称为有向图的弧(边)，f 是一个从 E 到 $V\times V$ 上的映射(函数)。为了把本体知识库存储起来，本文组织了邻接链表的形式。为图的建立、遍历和查找建立相应的数据

基金项目：国家“十五”计划基金资助项目“基于信息挖掘技术的名老中医临床诊疗经验及传承方法研究(老中医学术思想、经验传承研究)”(2004BA721A01H07)；北京市自然科学基金资助项目“基于认知计算的经验知识获取与知识构建理论技术研究”(4062022)
作者简介：张德政(1964－)，男，副教授，主研方向：认知计算与创新，知识发现，中文信息处理；庄洪波，硕士生
收稿日期：2006-07-05 **E-mail：**ustb_zhb@163.com

结构形式。在这些资源的基础上，就可以建立知识获取的系统框架。对文档进行分析处理，生成了新的知识，达到了知识获取的目的。

2 系统结构与功能

KATCM 系统对单份医案文本进行处理。进行分词和特征词抽取后，形成特征词集，对本体知识库的信息进行概念实例化和形成语义场。此系统包括 3 个模块：文本预处理模块，本体知识库模块和知识获取模块。系统结构如图 1 所示。文本预处理模块负责对文本进行词法分析，提取出文本的特征词。本体知识库模块负责中医领域本体的组织、数据结构表示。知识获取模块负责文本特征词的实例化，获取文本特征词的网状知识库，组织形成语义场。

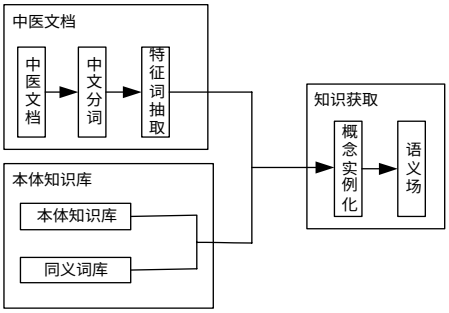

图 1 知识获取系统结构

3 文本预处理

3.1 中文分词

在中文信息处理中，分词是基础的工作。现有的分词算法可分为 3 类：基于词典的机械分词方法，基于统计的分词方法和规则的分词方法[3]。本文采用基于统计模型的最大概率法进行分词。最大概率法相当于一元语法模型，每次选择出现概率最大的路径作为切分结果。运用一元语法模型可以达到 90%以上的切分正确率。利用大规模的中医语料库和成熟的n元语法统计模型，可以很容易将切分正确率提高很多。

3.2 去停用词和抽取特征词

在对文本进行分词处理后，所有的分词项划分为：功能词，高频噪声词和特征词。功能词本身没有实质意义，只起到连接、指代的语法功能的词语，多为虚词。有些词在文本中出现的频率极高，如“ 的 ”、“ 着 ”，但这些词是文本分析的噪声词，对文本分析没有贡献。特征词对文本分析有突出作用，决定整篇文档的信息和知识。

本文采用了预制停用词库的方法，可以达到滤词和抽取特征词的作用。首先对大量病例病案进行切分和标注，训练出词频和词性的信息，把词性为助词、代词、介词、语气词、语素词的词汇作为停用词的候选词汇。从候选停用词中，选择那些可能会在文本中频繁使用，而无宜于语义表达的词语作为停用词。本文最终构造停用词 800 个，包括符号 40 个。在对文本进行滤词后，可以抽取出特征词，结合本体概念词库和本体概念同义词库，抽取出文档对应本体网络的标准概念。为下一步本体概念的实例化奠定了基础。

4 本体知识库

4.1 领域本体的数据结构

本体是一个哲学的概念，被哲学家用来描述事物的本质。文献[4]本体中的关系表示概念之间、概念和个体实例之间的关联。领域本体是用于描述指定领域知识的一种专门本体。它由概念、关系和子领域本体组成。开发一个本体的过程包含：定义本体中的类，定义概念之间的关系。通过添加特定的属性插件赋值信息和限制条件，就可以建立起一个知识库。这里采用 Protégé-3.2 工具构建了中医领域本体知识库，并将部分中医领域本体中药症关系用绘图软件 NetDraw 以形象化的形式表示出来，如图 2 所示。

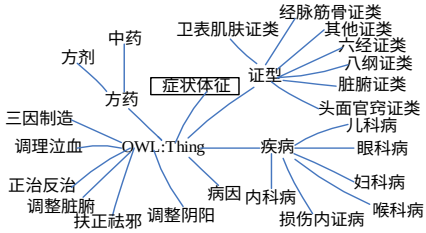

图 2 部分中医本体网状图

在此基础上，本文运用图论的理论构建了中医领域本体的网状数据结构[5]，定义了结点类型和邻接表内条目类型如下:

```
//顶点基本信息
struct Vertex
{ string name;//本体概念结点
    vector<Edge> adj;//邻接结点
    Vertex *prev;//前 1 个结点
    int scratch;//相关信息    };
//邻接表内条目
struct Edge
{
  Vertex *dest;//第 2 个本体概念结点
  string relation;//概念间的结点关系    };
```

对知识库中的概念及关系组织成网状数据结构，以各个概念名称作为头结点，与它相关的概念和概念之间的关系作为邻接链表的结点，其中几味药的邻接链表数据结构见图 3。

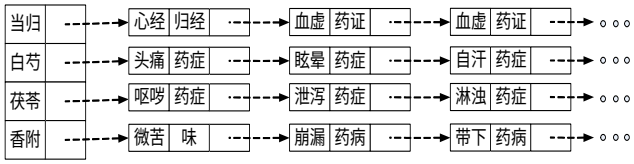

图 3 部分本体邻接链表结构

4.2 概念标准化

在组织好本体数据结构后，就可以进行相关的本体知识获取的研究。对文本进行特征词抽取后，得到的是相关的特征词的集合。为了对应到本体的概念，需要对特征词进行本体术语概念的标准化。这里采用统一词库的方法，把标准术语概念和特征词之间作映射。一个本体概念可以对应多个特征词和不规范词汇。

5 知识获取

5.1 概念实例化

概念标准化后，就可以对文本分析而来的概念进行实例化的映射。把与单片文档相关的概念以及概念间的关系从本体知识库中抽取出来，形成多个概念及概念关系的语对。把这些语对再组织成网状的数据结构形式，在此基础上就可以进行语义场的构建。

5.2 语义场

语义场指意义有关联的词共同构成的一个集或区，场内每个成员的意义取决于与成分之间的相互制约关系。语义场

（下转第 200 页）