

●张玉峰 王翠波

概念知识库的设计与学习方法^{*}

摘要 概念知识库包含两类知识元素:与专业领域相关的概念,概念之间的语义关系。概念库中,属分等级关系是概念之间诸关系中最本质的关系,是组织概念库的核心元素。图 2。表 3。参考文献 4。

关键词 概念知识库 设计 概念知识学习 机器学习

分类号 G253

ABSTRACT Conceptual knowledge base includes two categories of knowledge elements, i. e. subject-related concepts and semantic relations among concepts. In the concept bank, hierarchical relations are the most essential relations, and are the basic elements in the organization of concept banks. 2 figs. 3 tabs. 4 refs.

KEY WORDS Conceptual knowledge base. Design. Conceptual knowledge learning. Machine learning.

CLASS NUMBER G253

知识经济时代,人们不满足于获得表层的信息,而期望获得高质量的知识。如何综合应用信息科学、人工智能等多学科的理论与技术,实现网络信息资源的知识组织与知识检索,是人们关注的新课题。各专业领域中概念知识的组织是网络资源知识组织的基础。专业概念知识,例如主题词及其相互关联知识,是最基本的知识元素,用来表达与组织文献内容和用户提问。设计与建立概念知识库是实现知识组织和知识检索任务的基础工作。

本文所说概念知识库的设计采用通用模式,并侧重于概念结构和逻辑结构的设计。该知识库不仅可以浏览显示,而且利用了机器学习技术,能够获取新概念,不断优化和扩展概念知识库。

1 概念知识的获取与表示

1.1 概念知识的获取方法

概念反映了自然界、人类社会以及人类思维活动各领域的事务。它描述了事物的各种特征和问题,通常以词或词组的形式表达。由于标引和检索的需要,人们按照一定的原则(如通用性、使用频率等),构造了各种表达概念知识的人工语言。

概念知识的获取是建立概念知识库的第一步。它要求识别和选取能表达问题主题内容的重要词汇

或词组。概念可能是明确的,也可能是隐含的,它们的类型和形式各不相同,有些是基础的,有些是辅助性的。概念知识可利用不同的技术通过人工方法、半自动化方法和全自动化方法获取。CODER 智能检索系统利用人机结合方法和自然语言处理技术,从专业词汇高度密集的人工智能手册中获取人工智能领域的概念知识。自然语言处理是基本方法,但难度较大。获取概念知识的较易实现的、较有效的自动化方法是,采用机器学习技术,从各专业文献中、从用户和专家的交互信息或检索实例中获取。还可以通过概念相交、概念组配等概念逻辑方法获得新概念。

概念知识的关键部分是概念之间的关系知识。在获取基本概念之后,需要识别概念之间的各种语义关系。首先是评估、识别概念之间的相关性,专家们已经开发了各种自动分类或聚类方法用来测量概念间的相似值。经过聚类形成概念类以后,重要的工作是获取类中概念之间的语义关系,尤其是等级关系。此项工作比较困难,通常采用统计分析、机器归纳学习、利用专家启发式知识的逻辑推理来推导概念之间的等级关系。目前,数据挖掘技术将机器学习技术与数据库技术相结合,从数据库中提取概念知识和概念之间的关联模式。

1.2 概念知识的表示方法

* 本文为教育部人文社会科学研究重大项目(信息可视化与知识检索)研究成果。

知识常常是一种很复杂的结构化的信息集合。在人工智能领域,常用的知识表示方法有一阶谓词逻辑表示法、产生式规则表示法、框架式表示法、语义网络表示法等。这些方法是根据计算机程序运算方式的不同来区分的。从一般逻辑方法看,知识表示方法可分为定义法、概述法、指示法和逻辑标示法等。产生式规则和谓词逻辑方法适合表示简单结构的知识,而语义网络侧重于表示语义关系,相关知识可以从相连的结点推导出来。由于概念之间存在各种纵向和横向的关系,构成一个网络形式,因此,选择语义网络作为概念的知识表示形式较方便。当然,概念知识的表示应该在方便用户使用系统的前提下灵活设计。准确、快速是知识表示遵循的原则。

2 概念知识库的设计

概念知识库(简称概念库)是知识库中的一种。它提供各专业领域的概念模型,不仅给出了各领域中的主要概念,而且提供了概念之间的各种语义关系。它不仅是语言规范的指南,也是全面地表达文献知识和实现高效率检索的重要工具。这里采用语义网络方法表示概念库的知识模式,用面向对象的方法设计它的逻辑结构。

2.1 概念库的概念结构

概念知识库包含两类知识元素:一类是与专业领域相关的概念,另一类是概念之间的语义关系。从标引和检索科技文献的主要工具《汉语主题词表》的款目结构中可以知道,概念之间主要有等同关系(用、代关系)、等级关系(属、分关系)、相关关系。现以“航天通信”为例,用语义网络方法表达概念库的概念结构,如图1所示。图中D表示代关系,S表示属关系,F表示分关系,C表示相关关系,Z表示族首词。图中“空间通信”与“航天通信”之间也存在用关系。

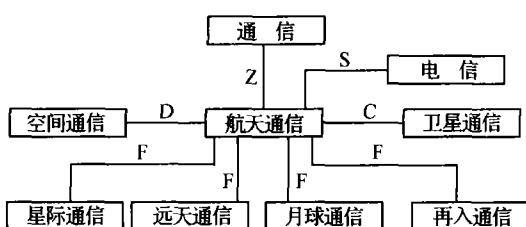


图1 概念库的概念结构

2.2 概念库的逻辑结构及其实现

概念库中,属分等级关系是概念之间诸关系中最本质的关系,是组织概念库的核心元素。面向对

象的方法很适合表达这种具有等级关系的知识。依据各类概念知识及其关系,采用面向对象方法表达的概念对象图如图2所示。

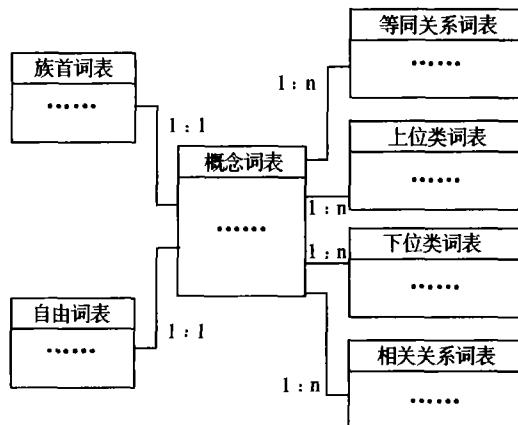


图2 概念对象图

概念对象图中含有7个概念对象:概念词表、族首词表、自由词表、等同关系词表、上位类词表、下位类词表和相关关系词表。其中概念词表是主表,其他为子表。概念的属分关系、等同关系不是唯一的,也不是固定的。一个概念可能有一个或多个属概念,可能有一个或多个等同概念。当然也存在没有属概念和等同概念的情况。

为了解决这种不确定关系,设计过程中利用Oracle9i提供的对象类型扩展。将用、代、属、族、分、参考这些关系设计成对象类型。然后,利用这些简单的对象类型来创建嵌套表类型。嵌套表类型可以用来作为概念实体的基本属性。它还可用来将子表的数据嵌入其父表,动态实现前一段中讨论的不确定关系。概念对象图中的等同关系词表、上位类词表、下位类词表、相关关系词表均定义为嵌套表类型,由系统自动生成。

下面主要讨论概念词表、族首词表、自由词表的实现方法。

(1)概念词表。概念词表综合了正式主题词、非正式主题词和自由词,结构如表1所示。该表是主表,通过词序号(SNO)与其他子表连接。

表1 概念词表

列名	数据类型	可否为空	说明
SNO	Number	yes	概念词的序号

续表

列名	数据类型	可否为空	说明
phrase	Varchar2(20)	no	概念词或词组
flag	Varchar2(1)	yes	(非)正式主题词、自由词
classification	Varchar2(10)	yes	概念词的分类号
Yongs	SYETEM. YongListType	yes	用
Dais	SYSTEM. DaiListType	yes	代
Fens	SYSTEM. FenListType	yes	分
Shus	SYSTEM. ShuListType	yes	属
Zus	SYSTEM. ZuListType	yes	族
Cankaos	SYSTEM. CankaoListType	yes	参考词

概念词表中, Yongs、Dais、Fens、Shus、Zus、Cankaos 为嵌套表类型。表中的正式主题词、非正式主题词和自由词,通过 flag 标记加以区分。表中的自由词作为增添的概念检索入口。

(2)族首词表。将概念词表中每个族首词都抽取出来,单独建为一个族首词表(相当于为主题词做索引),以提高知识库的检索效率。该表结构如表 2 所示。其中, subject 是对象引用类型,表达族首词 phrase 与相应的概念词表之间的关联。

表 2 族首词表

名称	数据类型	可否为空	说明
SNO	integer	yes	概念词的序号
phrase	Varchar2(20)	no	概念词或词组
classification	Varchar2(10)	yes	概念词的分类号
subject	REF SubjectType	yes	指向概念词表

(3)自由词表。把概念词表中的自由词提取出来单独做一个索引。单独成表是为了能对自由词的使用频数进行统计,为下一步的学习方法打基础。该表结构如表 3 所示。其中, subject 是对象引用类型,表达自由词 phrase 与相应的概念词表之间的关联。

表 3 自由词表

名称	数据类型	可否为空	说明
SNO	integer	yes	概念词的序号
phrase	Varchar2(20)	no	概念词或词组
count	integer	yes	概念词使用次数计数
flag	Varchar2(1)	yes	自由词
subject	REF SubjectType	yes	指向概念词表

3 概念知识的学习与维护

学习是一种自然的认识处理,是人(或计算机)增长知识和改善其技能的方法。如果一个计算机系统具有学习能力,它就可自动改进自身的执行性能而不需要重新进行程序设计。概念知识库具有学习功能,可以适应系统内外环境的变化,不断学习新的概念知识,以保证知识库内容的正确性、完整性和新颖性。

这里介绍两种机器学习方法:一种是基于权值优先的学习方法精炼概念知识,另一种是基于自由词学习新的概念。

基于权值优先的学习方法,就是依据用户的访问和反馈信息学习。系统对每一个主题概念分配一个权值,表示用户对它的满意程度和访问频率。系统通过观察记忆机制对用户经常检索和访问的主题进行统计,并转换成权值。另一方面,通过反馈学习算法根据用户反馈信息修改相关权值。系统依据权值优先的原则,周期性地检查精炼概念知识。

基于自由词的学习是,在机器获取了自由词的情况下,如何判断自由词是否能被提取出来作为备用的正式主题概念,即正式主题词或非正式主题词。当然这些备用的主题概念还需提交专家进一步审核。学习系统所起作用是记录自由词并统计其使用频率,依据限定的条件将其转化为备用词,为专家决策起辅助作用。

系统基于概念知识库中的自由词表,在用户进行检索时,自动将检索词与自由词表进行比较,如果该检索词在自由词表中,则将其 COUNT 加上 1。可以将 COUNT 值做一下限定,假设当自由词被查询的次数达到一定阀值时,系统则自动改变该自由词标志 flag,将 flag 设置为一个特殊值,用以表示该词可以供专家参考。

参考文献

- 1 张玉峰. 智能情报系统. 武汉:武汉大学出版社,1991
- 2 蒋永福. 论知识组织的语言学基础. 图书情报工作, 2001, 17(5)
- 3 周六炎. 科技文献管理. 武汉:武汉大学出版社,1992
- 4 Iadh Ounis. Organizing Conceptual graphs for Fast Knowledge Retrieval. IEEE, 1998

张玉峰 武汉大学信息资源研究中心教授、博士生导师。通讯地址:武汉市。邮编 430072。

王翠波 武汉大学信息管理学院硕士研究生。通讯地址同上。
(来稿时间:2002-12-10)