

一种智能辅导系统中建模的实现方法

罗瑶, 赵克, 李亚涛, 丁蛟腾

(西安电子科技大学 机电工程学院, 陕西 西安 710071)

摘要:针对自然语言中数学领域的特点,提出了一种智能辅导系统中建立模型的方法。该方法通过对语义理解结果的名词聚类分析,调用静态知识库中相应的内涵模型,根据静态知识库中对内涵模型的描述在语义理解结果中搜集相应的信息使抽象概念具体化。在此基础上依据领域知识对题目进行信息挖掘,并根据题目类型对信息进行约简,从而建立模型。这种建模方法在智能辅导系统中得到了较好的应用。

关键词:自然语言理解;静态知识库;概念模型;内涵模型

中图分类号:TP182

文献标识码:A

文章编号:1673-629X(2008)09-0046-04

An Approach of Building Model for an Intelligence Tutoring System

LUO Yao, ZHAO Ke, LI Ya-tao, DING Jiao-teng

(College of Electromechanical Engineering, Xidian University, Xi'an 710071, China)

Abstract: According to the characteristic of natural language understanding in mathematics domain, an approach of building model for an intelligence tutoring system is introduced in this paper. By using the concept model theory and analyzing the result of semantic analysis, the right connotation model can be transferred, corresponding to the connotation model, collecting the information from the result of semantic analysis to embody abstract concept, based on which dig over the information then reducing it according to the type of the subject, thereby to build the model. This kind of method has been applied to intelligent tutoring systems well.

Key words: natural language understanding; static knowledge base; conceptual model; connotation model

0 引言

文中研究的智能辅导系统是将计算机网络和人工智能技术相结合而开发的。系统采用基于知识的推理技术的开发平台,和产生式规则的表示方法,用 Java 专家系统开发工具 Daok 来实现知识库。自然语言理解^[1]建模的任务是建立一种计算机模型,这种计算机模型能够给出像人那样理解自然语言(即人们日常使用的汉语、英语等)的结果^[2]。模型建立的成功与否对自然语言理解起着至关重要的作用。只有自然语言中的信息被正确地组织与表示,机器才能对自然语言做出正确反应。自然语言理解中模型的建立作为研究热点近几年一直被广泛关注着。文中论述了一种领域内的建模方法。以概念模型为理论基础,根据概念间的操作关系结合静态知识库中的概念从属树与内涵模型,使语句中的核心概念由抽象到具体,从而建立正确

的问题模型。

1 概念模型的理论知识

在逻辑学、语言学、心理学、知识工程和数据挖掘中,概念被看作知识最基本的单元^[3]。概念是知识表示和操作理论的中心。概念模型包括概念和概念关系模型。在概念模型中,概念是构成自然语言最基本的单元,概念模型被用来表示概念的内涵与外延以及之间的关系。在此基础上建立的概念关系模型在概念间进行具体的操作。在建模中根据概念模型建模是人们依据的主要理论依据。

1.1 概念

设 Q, V 为概念 R 的内涵与外延,其中 $q_1, q_2, \dots, q_i, \dots, q_n$ 为概念的基本永真命题, $q_{i1}, q_{i2}, \dots, q_{im}$ 为 q_i 的等价永真命题,则内涵可表示为:

$$Q \leftrightarrow q_1 \wedge q_2 \wedge \dots (q_i \wedge q_{i1} \wedge \dots q_{im} \wedge \dots \wedge q_n) \wedge \dots \wedge q_n \quad (1)$$

外延的属性为 v_1, v_2, \dots, v_k , 相应的值域为 D_1, D_2, \dots, D_k , 这些值域组成的笛卡儿空间可表示为 $D = D_1 \times D_2 \times \dots \times D_k$, D' 为满足 Q 约束的子集, D' 为概

收稿日期:2007-12-09

基金项目:科技部科技型中小企业创新基金(01c26226111002)

作者简介:罗瑶(1982-),女,陕西西安人,硕士研究生,研究方向为人工智能、知识工程、创新技术;赵克,教授,硕士生导师,研究方向为人工智能、知识工程、创新技术。

念的变化范围, $d^k = (d_1, d_2, \dots, d_k)$ 是概念中的一个点。 (Q, d^k) 称为概念 R 的实例概念。概念 R 可表示如下:

$$R = \{Q, [d^k \mid d^k \in D' \subseteq D]\} \quad (2)$$

概念模型被用来表示概念的内涵、外延和概念内涵外延之间的关系,然后建立概念间的操作关系。概念的内涵与外延从不同侧面对概念进行了描述,二者的有机结合才是对概念的完整描述,概念的外延实际上是该概念内涵约束下的概念的可行空间。概念内涵中所涉及到的概念称为该概念的具体概念。

在知识工程领域,作为基于知识的自然语言处理系统中的知识库,本体是由概念集合以及概念之间的联系所构成的知识实体^[4]。从具有较高抽象意义的知识下划分出具体知识的过程称为特例化知识,它是人们从抽象到具体认识的结果。这种特例化知识具有抽象知识的性质,并且具有自身细化于抽象概念的属性。下面介绍概念间的两种操作——例化操作与实例化操作。

例化操作:设概念 R 的外延 V 由变量 v_1, v_2, \dots, v_k 组成,变量 v_i 的值为 D_i , v_i 在 D_i 中的任意一次取值记为 $\gamma(v_i)$,称作概念 R 的一次例化,记为 $\gamma_i(R)$ 。例化操作的定义可表示为:

$$\gamma_i(R) = \{Q, [d' \mid d' = \langle d_1, d_2, \dots, d_{i-1}, d_i, d_{i+1}, \dots, d_k \rangle \wedge Q, d'] \in R \wedge d_i = \gamma_i(v_i)\} = \{Q \wedge \gamma_i(v_i), [d \mid d = \langle d_1, d_2, \dots, d_{i-1}, d_{i+1}, \dots, d_k \rangle \wedge (Q \wedge \gamma_i(v_i), d) \in R]\}$$

实例化操作:假设概念 R 的外延由属性 v_1, v_2, \dots, v_k 组成,当外延的每一个属性具体化后,叫做概念 R 的实例化。

1.2 概念间的抽象具体关系

概念间的抽象具体化关系是由概念的实例化和细化形成的一种关系。实际上,每一次实例化和细化操作是一个内涵增加外延减少的过程。概念外延减少的原因是外延约束增加,概念内涵增加的原因是新的基于存在命题增加,新命题对新外延的范围起决定性的约束。所以,它们相互补充,不可分割。

例如在一个一元方程中, Q 为方程的内涵, 1 为方程中未知量的个数,外延的可变范围由在一元方程中两个独立的属性 x 轴和 y 轴决定,则一元方程的可变范围 V 包括第一象限的所有空间。对直角三角形进行细化操作,添加新的性质——未知量的最高次幂为 2,过程如下:

$$\delta_{\text{最高次幂}=2}(\text{一元方程}) = \{Q, [d' \mid d' = \langle \text{未知量}, \text{未知量最高次幂} \rangle \wedge (Q, d') \in \text{一元方程} \wedge \text{最高}$$

次幂为 2]\}

对一元方程进行操作的命题公式为:最高次幂 = 2,即对概念的变化范围增加最高次幂为 2 的约束。在操作过程中一元二次方程的本体与变体发生以下变化:

(1)本体事实增加。当新事实最高次幂 = 2 加入本体,推理可得相关新事实,同时一元方程的事实经推理后也得到简化。

(2)随着本体约束的增加,变体变化范围缩小。变体的变化范围第一象限内所有的点,变为第一象限内的一条射线,即从原点以 45° 发射出的射线。

这就说明,在从抽象概念到相对抽象概念的演化过程中,概念的内涵逐渐扩大,外延逐渐缩小。

2 智能辅导系统中问题模型的分析

2.1 问题模型的分析

在智能辅导系统中所指的问题模型主要是指用来解决数学领域问题的模型,分为几何与代数两个方面。文中是基于代数领域来进行论述的。模型就是求解一类问题时所有必要信息的集合。集合内各条信息都是对问题的描述,通过集合内各条信息的约束使模型的内涵扩大外延减少,因此模型建立的越准确,就能越好地反映问题。

2.2 关于建模的分析

在自然语言中,理解一个问题要经过两个部分即拆分与合并过程。拆分就是在进行建模之前,通过对语句的词语切分、句法分析、语义分析,生成相应的动态名词、动态关系、动态事件等句子的初步理解结果。合并是将拆分后的结果按照概念间的操作关系进行合并的过程^[5]。建模主要是在这一部分完成的。问题模型的建立可分为以下四个步骤。

2.2.1 中心名词汇聚

数学领域中的每一句话都是围绕一个核心概念进行描述的,这个核心概念就是所说的中心词,其它的概念都是围绕它的属性、组成、过程等来描述的。在静态知识库中各名词概念都是按照名词间的继承与包含的关系存储在概念从属树中。概念从属树是一棵与或树,在不同的分类标准下,内涵相近的概念聚成一个集合。本系统所采用的中心名词汇聚的方法,就是结合静态知识库中的名词概念从属树通过爬树找出包含句中名词概念数最多的概念,则该概念就是中心词。以该核心名词为基点,在概念从属树中查找其他名词与核心词的关系从而实现中心名词的汇聚。

例如:设 x_1, x_2 是方程 $2 * x^2 - 6 * x + m = 0$ 的两个根,且 $x_1^2 + x_2^2 = 6$,则求 $x_1^2 - x_2^2$ 的值。根据名词间

的属性联系则可得如图 1 所示的关系。

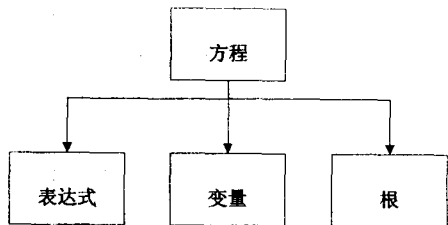


图 1 例子中名词间的联系

根据查找中心词的方法可知方程包含的概念数最多,则该句的中心词就是方程,其他的名词是围绕它来描述的。中心词汇聚的结果就是(方程(表达式)(根)(变量))还不能反映题目信息只有经过抽象具体化才可使其转变为概念的一个实体。

2.2.2 抽象具体化

概念的抽象具体化顾名思义就是使概念由抽象变具体的过程,它包含抽象概念具体化与概念属性的具体化两部分。中心名词汇聚的结果只是一个抽象的概念,还不能作为一个描述问题的模型。要实现从抽象概念到具体问题的转变,首先要通过查找名词概念从属树使抽象概念落实到具体概念缩小其概念范围,然后利用语义理解结果对具体概念中的各个属性进行例化操作实现概念的实例化。

抽象概念的具体化也就是领域的确定。通过领域的确定可为信息挖掘提供正确的领域知识。抽象概念的具体化首先要确定该抽象概念在整个概念从属树中的位置确定其领域,然后查找以该抽象概念为根从属树,通过概念模型的内涵找出与句中概念特征相符的子概念使其领域进一步细化。例如:中心名词“方程”的概念从属树如图 2 所示。

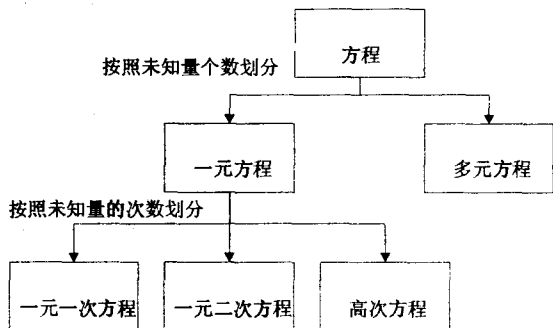


图 2 方程概念从属树

比较后可知一元二次方程的内涵与题目中的描述相符。就可使相对抽象的核心概念“方程”上升为相对具体的概念“一元二次方程”。概念的具体化降低了领域间的耦合度从而使建模的难度降低。

概念属性的具体化也就是概念的实例化操作。通过对概念属性的例化操作实现概念的实例化。例如,根

据例子的语义理解结果中的信息对一元二次方程的表达式这一属性进行例化操作可得:

$$\delta_{\text{表达式}} = 2 * x^2 - 6 * x + m = 0 (\text{一元二次方程}) = \{Q, [d' | d' = \langle \text{表达式}, \text{变量}, \text{根}, \text{条件} \rangle \wedge (Q, d') \in \text{一元二次方程} \wedge \text{表达式} = 2 * x^2 - 6 * x + m = 0]\}$$

同理,例化概念的其他属性,就可实现该概念的实例化。经过这一步骤得到的句子信息中将不存在抽象形式,再对句子进行重新的语义分析,便可得到新的语义理解结果。但若句子中本不存在抽象信息则不需要重新的语义分析。

2.2.3 信息挖掘

核心概念经过具体化操作后变成一个描述具体个体的概念实体。然而解决问题时不仅需要概念的表面信息,更重要的是掌握许多与此相关的深层知识,也就是根据题目的语义理解结果对题目的已知信息与未知信息进行挖掘。

已知信息指的是题目中直接给出,但不是概念外延中的那部分信息。已知信息可直接通过对语义理解结果挖掘来获得。例如,题目中的: $x_1^2 + x_2^2 = 6$ 这就属于已知信息。

对于隐含信息,计算机必须结合特定领域内的知识将题目中的深层信息挖掘出来这也就是所说的语用推理。语用推理就是根据语用准则,结合一定语境,从语句的字面意义推出其深层意义的推理。例如,在求解一元二次方程的过程中经过前面的抽象具体化,得出判别式这一属性的具体对象,结合静态知识库关于一元二次方程的领域知识,由根的个数为 2 便可推出判别式大于等于零。有时同样的信息在不同的语境中所推理出的结果截然不同,因此,运用语用推理知识挖掘隐含信息,推理过程必须要和具体的语境结合,这样才能准确地描述出语言中的所有信息。

经过信息挖掘后题目信息的具体形式如下:(1) 等式;(2) 不等式;(3) 题目求解信息。

其中题目的已知信息分别以等式与不等式的形式表现出来,需求信息存放在题目求解信息中。

2.2.4 信息约简并建立模型

经过以上几个步骤题目中的信息已经全部具体化,但是在确定的领域内并不需要将概念所有的属性例化后才能实现概念的实例化。为了降低冗余、提高系统运行速度,就需要对信息进行约简。约简是通过题目信息与所需静态条件之间的匹配去掉那些与问题解决无关的信息来实现的。进行约简首先要根据语义理解结果确定题目的类型,然后再根据静态知识库中与此类问题有关的必需静态条件对现有信息进行筛选。例中题目是求 $x_1^2 - x_2^2$ 的值,条件中的未知量最高

次幂为2、二次项系数不为零对题目的求解没有帮助可以约去。约简后的题目信息进行组织后便可生成进入解题系统的问题模型。

3 数学领域自动建模的实现

在数学领域中,依据以上分析自动建模的具体流程如图3所示。

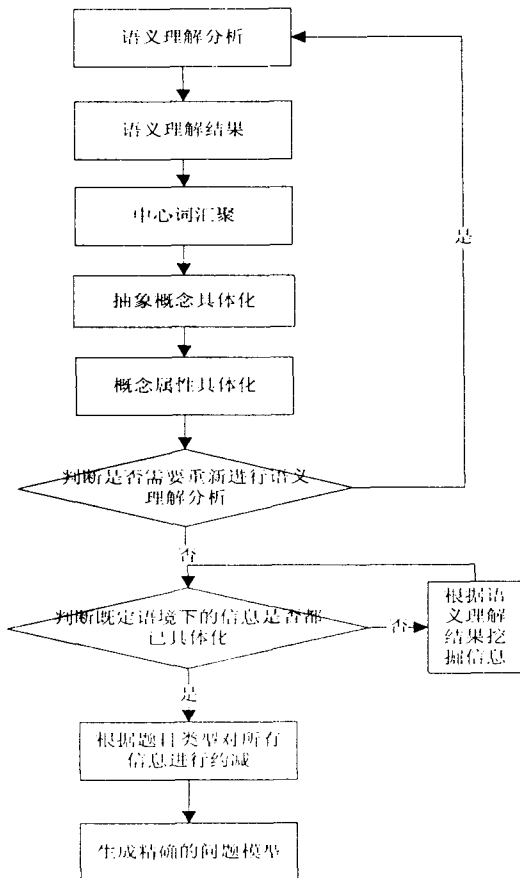


图3 数学领域建模流程图

经过以上处理后自然语言中的信息就可被计算机直接处理。

4 测试结果

本系统在设计时搜集了近300道的一元二次方程的题目,对其进行分类,设计处理流程。一元二次领域中的题目类型涉及面约为95%,其中包括以下几类;

(1)求未知量值(已知一根求另一根,求两根,已知两根关系求与其相关的第三未知量的值);

(2)求代数式的值(已知两根关系与方程系数求未知表达式的值);

(3)求范围求符号(仅由不等式推出,需要通过不等式与方程联合才可求解)。

测试后通过率可达90%以上。已可与自动解题系统相结合得到最终的结果,基本实现了初步的设想。

5 结束语

论述了一种智能辅导系统中模型建立的基本过程,并使用一元二次方程问题为例子,对其具体实现进行了详细的描述。为同类问题的解决和系统的进一步的壮大提供了借鉴的基础。但由于系统所掌握的知识程度有限对于需要深层理解的纯文字的应用题目目前还无法处理。

参考文献:

- [1] Allen J. 自然语言理解[M].第2版.北京:电子工业出版社,2005.
- [2] 俞士汶. 自然语言理解与语法研究[M]. 北京:商务印书馆,1999.
- [3] 周经野. 基于自然语言计算模型的汉语理解系统[J]. 软件学报,1993,4(6):41-46.
- [4] Jurafsky D, Martin J H. 自然语言处理综合[M].北京:电子工业出版社,2005.
- [5] 王继成,武港山,周源远,等. 一种篇章结构指导的中文Web文档自动摘要方法[J]. 计算机研究与发展,2003(3):27-31.

(上接第45页)

- Job Scheduling on Computational Grids Using Multiple Simultaneous Requests[C]//Proceedings of 11th IEEE International Symposium on High Performance Distributed Computing. Washington, DC, USA: IEEE Computer Society, 2002:359-366.
- [5] 罗红,慕德俊,邓智群,等. 网格计算中任务调度研究综述[J]. 计算机应用研究,2005(5):16-19.
 - [6] Cao J, Kerbyson D J, Nudd G R. High performance service discovery in large-scale multi-agent and mobile agent systems[J]. International Journal of Software Engineering and

Knowledge Engineering, 2001,11(5):621-641.

- [7] Nudd G R, Kerbyson D J, Papaefstathiou E, et al. PACE a toolset for the performance prediction of parallel and distributed systems[J]. International Journal of High Performance Computing Applications, 2000,14(3):228-251.
- [8] Buyya R, Murshed M. GridSim: a Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing[J]. Concurrency and Computation: Practice and Experience, 2002,14(13-15):1175-1220.