数学领域集体词结构形式化处理研究

党 建,亿珍珍,赵 克,殷 鸿
(西安电子科技大学 机电科学与技术研究所,陕西 西安 710071)

摘 要:根据数学领域自然语言理解的特点,结合集合论的思想深入分析了集体词结构。集体词结构是表示一个可数的集体概念,其外延是一个事物类。集体词结构较好地解决了数学领域中的数量词结构的形式化处理问题。首先给出了集体词结构的语义认知基础,并采用基于知识的方法,应用本体论思想,构造了系统的集体词结构模型。然后对集体词结构的群体关系进行分类和介绍。这种集体词结构的处理方法在数学领域智能辅导领域中得到了较好的应用。

关键词:自然语言处理;集合论;本体论;集体词结构
中图分类号:TP182 **文献标识码**:A **文章编号**:1673-629X(2007)05-0121-04

Research of Formalization Processing for Collective Structures in Mathematics Domain

DANG Jian, YI Zhen-zhen, ZHAO Ke, YIN Hong
(Research Institute of Mechatronics, Xidian University, Xi'an 710071, China)

Abstract: According to the characteristic of natural language understanding in mathematics domain and the idea of set theory, made a thorough analysis on collective structures. Collective structures is a numerable collectivity. It preferably solved the problems of formalization processing of quantitative structures in mathematics propositions. Firstly, the semantic cognizing base of the collective structures was presented. Secondly, applied the knowledge-based method and the theories of ontology to constructing this system's collective structures model. Finally, a suitable class of collective structures and its introduction was given. This kind of method of handling collective structures has been applied to intelligent tutoring systems.

Key words: natural language processing; set theory; ontology; collective structures

0 引 言

文中所涉及的自然语言处理系统为智能辅导系统(Intelligent Tutoring Systems, ITS)在中学数学领域内的实现提供了必要的人机交互接口。ITS是把人工智能应用于计算机辅助教育的一种专家系统,是一种更适用于教与学需求的面向辅导的教育系统(Educational Systems)。在系统中,将用户给出以自然语言表述的需求形式转化为计算机可以接受的已知条件信息,从而实现计算机的自动解题,这是系统智能化的关键技术之一。

这里所述的集体词结构及其处理过程就是应用在智能辅导系统中自然语言处理部分中的。关于集体词,国外的研究主要集中在集合名词,主要是从语法的角度来对名词的分类,没有多少应用。国内的在名词分类的著作中,受国外的影响,也有这种分类,例如朱德熙先生(1982)根据名词和量词的关系将名词分为可数名词、不可数名词、集合名词、抽象名词和专有名词[1]。其从认知的角度来说,具有一定的实际意义。但是对于汉语自然语言理解来说,在应用时都有一定的不足。文中从数学领域应用的角度,基于领域自然语言理解,应用本体论的知识对集体词做了一些研究。

1 集体词结构的理论分析

按照集合论观点,集合是一种含义更为广泛的概念——量,将概念的研究转化为集合的讨论,旨在通过用数学方法研究各种集合间关系和各种集合之间的运算来更好地从量的角度把握概念的逻辑内容[2]。

在数学领域,很大部分涉及到一个群体与其他群体的代数运算。所以从应用的角度,把传统的集合名词、数量词+名词结构和其他群体概念综合分析抽象为集体词结构的处理方法在处理数学问题时更具有应

收稿日期:2006-08-17
基金项目:科技部科技型中小企业创新基金(01c26226111002)
作者简介:党 建(1982-),男,陕西人,硕士研究生,研究方向为人工智能和知识工程;赵 克,博士,教授,研究方向为人工智能、创新设计和知识工程。

用价值。例如通过集体词的记录可以方便地分析整体与部分的关系,甚至可以用交、差、并、补等集合关系运算概念来进行集体词结构之间的运算,甚至实现数学关系的自动生成,尤其对于和群体相关关系的计算如“平均值,方差”等。

1.1 理论基础

1.1.1 本体论(Ontology)

本体是一个关于特定领域共享的理解,被认为是一些类(概念)、关系、函数、公理和实例的集合[3]。

概念和概念之间的关系是本体的两个非常重要的组成元素。

(1)概念(concepts)。

概念是物体或事件的模型知识。例如,线段和延长分别是物体和事件的模型知识,它们都是知识。

(2)关系(relations)。

在领域中概念之间的交互作用,形式上定义为 n 维笛卡儿积的子集:$R:C_1 \times C_2 \times \cdots \times C_n$。

从语义上讲,基本的关系有 4 种(见表 1)。

表 1 基于本体的概念之间的关系

关系名	关系描述
Part－of	概念之间部分与整体关系
Kind－of	概念之间继承关系,类似于面向对象中父类与子类的关系
Instance－of	概念的实例与概念之间的关系,类似于面向对象中对象与类的关系
Attribute－of	某个概念是另一个概念的属性,如“价格”是桌子的属性

在实际建模过程中,概念之间的关系不限于上面 4 种关系,可根据领域具体情况定义相应的关系[4]。

1.1.2 名词

智能辅导系统中的自然语言处理部分把名词按照处理的需求分为:个体名词、物质名词、可分集合名词、不可分集合名词、抽象名词、专有名词、过程名词和无量名词八种。

1.1.3 数量词

数词中可以与量词结合的有:表示确定数目的数词、序数词和概数词。智能辅导系统中的自然语言处理部分把量词按照处理的需求分为:个体量词、集体量词、度量词、容器量词、种类量词、成形量词、不定量词、动量词和时量词九种。

1.1.4 集体标识词

在现代汉语描述集体的句群中,“其中、任意、另、其余、另外、其他、各、每”在句子中的出现往标志着一个相关集体的出现,因此这些词被定义为集体标识词。例如:8 名学生,其中男生 5 名,女生 3 名。

1.2 本体知识库

按照本体模型的定义,将领域内相关的每个知识,构建其概念本体,概念与概念之间由关系相互联系,把知识库构建为一整体的概念网。图 1 是在本体知识库中对概念“三角形”及其属性概念“边”和“角”的局部静态语义关系的描述。

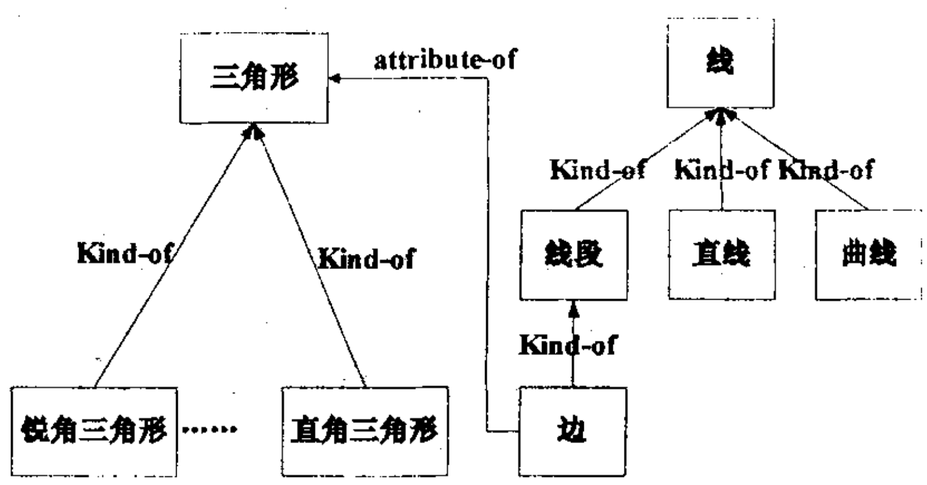

图 1 概念的局部静态语义关系

1.3 集体词结构的分类

通过对数学领域内大量的阐述各种群体概念的句群分析,把集体词的结构分为以下 4 类。

1.3.1 数量词 + 名词结构

①量词为个体量词、集体量词、容器量词、种类量词和成形量词,数词为数目大于一的数词。例如:八个学生。

这类是显式集体词结构,可以从数量词结构直接导出。

②量词为集体量词,数词表示的数目为一。例如:一队巡逻的士兵。

这类通过在知识库中查集体量词的类型,由定量集体量词与不定量集体量词分别生成定量集体词结构与不定量集体词结构。

③多个序数词 + 量词组合构成集体词结构。例如:猎人去打猎,第一次打了 3 只兔子,第二次打了 5 只兔子,第三次……。

这类通过识别序数词的个数,生成定量的集体词结构。

1.3.2 由枚举的元素生成

例如:样本:4 6 9 3 1。

这类通过枚举类型识别程序判断生成集体词结构。

1.3.3 可分集合名词

例如:“船队”是一个“船”的群体。

这类通过查概念本体的“Member－of”关系生成不定量集体词结构(见图 2)。

图 2 本体知识库中“船队”和“船”的关系

1.3.4 名词中隐含集体词结构

例如:“三角形”中隐含有“三条边”、“三个顶点”、“三个角”等集体词结构。

这类通过概念本体的“Attribute－of”关系,查找其定量属性概念(见图 3)。

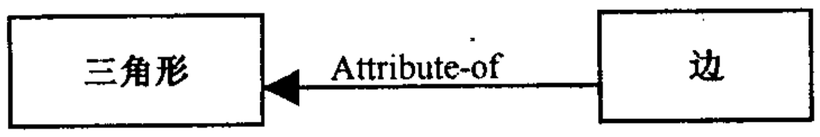

图 3　本体知识库中“三角形”和“边”的关系

1.4　集体词结构间关系(群体关系)

集体词结构间关系是指任意两个(或两个以上)集体词结构间的关系。它们是包含关系、相等关系,交叉关系和全异关系。

1.4.1　包含关系

给定集体词结构 A 与 B,如果 A 中的任一元素都是 B 的元素,则称 B 包含 A。记作:A⊆B。

例 1:有 9 名游客,其中有 5 个走路的(游客),4 个骑车的(游客)。

令 A=｛5 名走路的游客｝;B=｛4 名骑车的游客｝;C=｛9 名游客｝,则有 A⊆C,B⊆C。

1.4.2　相等关系

给定集体词结构 A 与 B,若 A⊆B,且 B⊆A,则称 A,B 两集体词结构相等,记作:A=B。

例 2:有两组数据:

第一组:3 5 8 9 0;第二组:0 8 9 3 5

令 A=｛第一组数据｝;B=｛第二组数据｝,则有 A=B。

1.4.3　交叉关系

给定集体词结构 A 和 B,如果有 A 的元素是 B 的元素,也有 A 的元素不是 B 的元素,并且有 B 的元素不是 A 的元素,则称 A 与 B 有交叉关系。

例 3:学校购买了两批图书,第一批有 4 种书:英语、数学、政治、历史;第二批 6 种书:语文、数学、政治、地理、生物、物理。

令 A=｛英语 数学 政治 历史｝;B=｛语文 数学 政治 地理 生物 物理｝,则 A,B 是交叉关系。

1.4.4　全异关系

给定非空集体词结构 A 和 B,若 A 的每一元素都不是 B 的元素,并且 B 的每一元素都不是 A 的元素,则 A 与 B 是全异关系。

例 4:在例 1 中集体词结构 A 与 B 就为全异关系。

2　集体词结构的理解过程

2.1　领域自然语言理解模型简介

在自然语言中,经常有语句的歧义存在,因此全面自然语言理解的实现存在很大困难,而基于领域则可以大大减少系统的复杂性[5]。同时,将静态知识添入事实库进行合理的抽象,可以大大降低处理的难度。因此,我们的模型采用基于领域和基于本体知识的处理方式,句子进入系统后将依次进行词法分析、句法分析、语义分析和篇章分析。

自然语言理解的知识库分为静态知识库和推理规则两部分。静态知识在执行推理之前首先加载,静态知识库的组织与存储采用基于本体论的方法,将数学领域的概念、术语以及它们之间的层次关系合理地组织在一起。

事实(fact)是产生式规则中推理的基础,事实靠模板来记录。模板(template)是产生式规则系统中记录事实所必需的东西,通过它的单值槽(slot)和多值槽(multislot)可以方便地记录一个概念或者术语的相关信息。如下是集体词结构的模板例子:

(deftemplate 集体词 (multislot 关系)(slot 标识)(slot 对象)(slot 内容)(slot 相关标识)(multislot 元素)(multislot 位置)(multislot 顺序)(multislot 参数)(multislot 其余))。

2.2　集体词结构的理解过程

系统处理集体词结构的过程分为在语义分析部分和在篇章分析部分两部分进行,如图 4、图 5 所示。

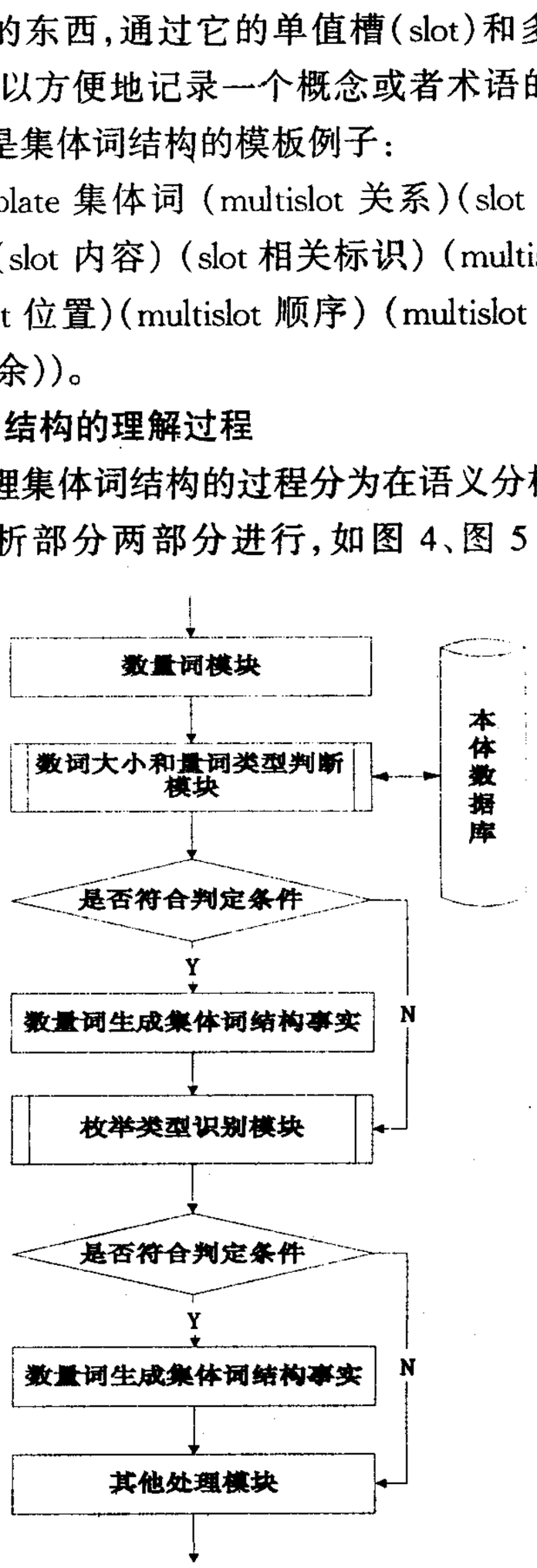

图 4　语义分析部分集体词结构处理

语义分析部分是根据数量词结构和相关事实直接生成集体词结构。这部分生成上文所述集体词结构的第一类和第二类。

篇章分析部分是首先根据已有的集体词结构通过判断模块发掘出隐含的集体词结构,然后分析所有集体词结构事实,确定结构之间的关系。这部分生成上文所述集体词结构的第三类和第四类。

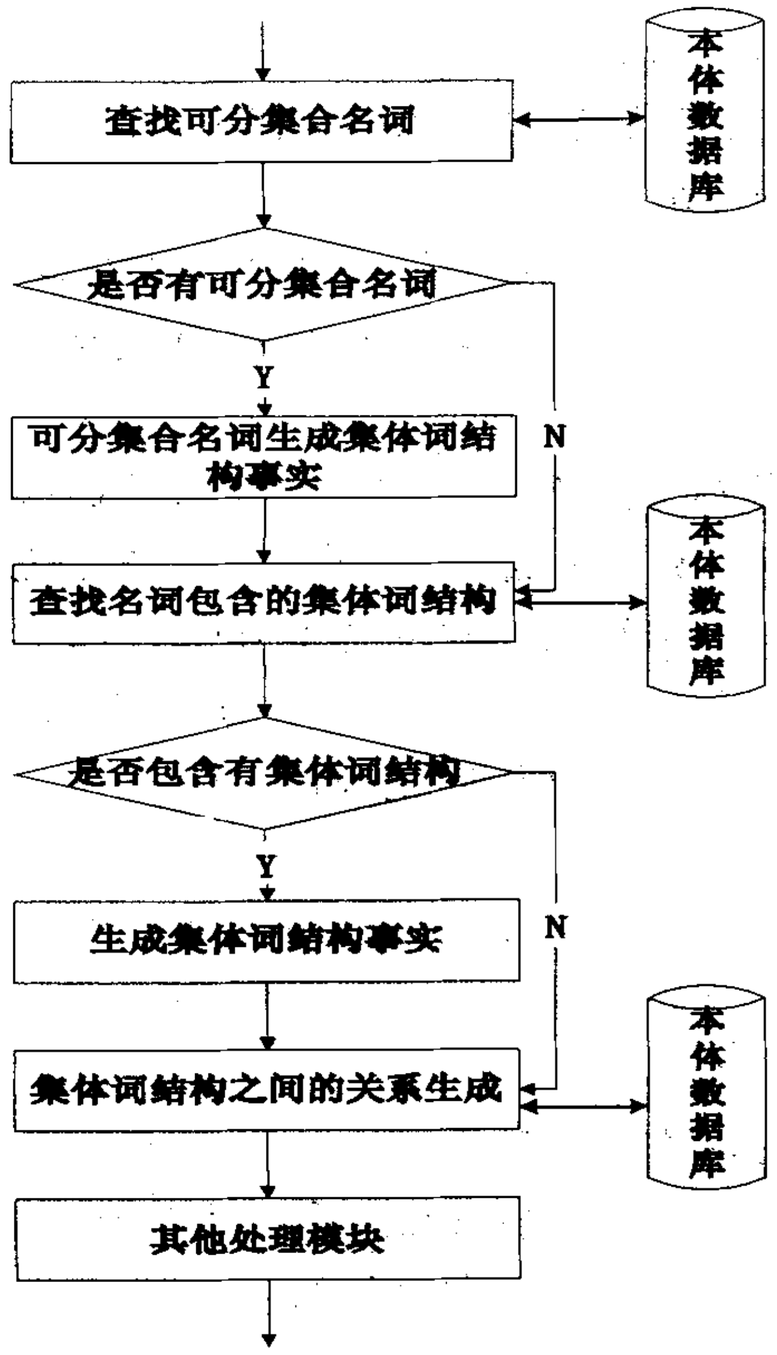

图 5 篇章分析部分集体词结构处理

集体词结构关系确定部分是理解集体词的关键。集体词结构之间的关系是集体词结构实际应用价值的主要体现。系统通过集体词结构差异查找模块完成结构间关系的确定。集体词结构差异查找模块包括结构的中心名词之间的关系比较、结构的元素数量比较、结构包含的所有元素的比较和集体标识词的位置四个方面,用来判断确定集体词结构关系。

3 结 语

群体的处理在数学领域占有很大的比重,文中基于数学领域提出了集体词结构概念及其应用。而笔者已经将这种处理模式应用到了智能辅导系统中。经过反复的、大量的句子测试,推理过程符合人的思维习惯,达到了预期的目的。但是汉语表达形式千变万化,这些分类并不能完全涵盖整个语料库,还需要进一步细划。文中的下一步工作就是继续查找语料库,完善知识库并重点完善与集体词结构相关的处理。

参考文献:

[1] 朱德熙. 语法讲义[M]. 北京:商务印书馆,1984.

[2] 周建设. 中国逻辑语义论[M]. 长沙:岳麓出版社,1996:231-240.

[3] 余以胜,张玉峰. 基于本体论的知识库系统研究[J/OL]. 情报杂志,2003(7). http://e37.cnki.net/kns50/detail.aspx?QueryID=7&CurRec=1.

[4] Zhu Cheng, Wang Zhenjie. Ontology Mapping For Interaction in Agent Society[C]//Services Computing, 2004(SCC2004) Proceedings. 2004 IEEE International Conference. China: Shanghai Jiao Tong University, 2004:619-622.

[5] Allen J. 自然语言理解[M]. 第 2 版. 刘 群,张华平,骆卫华等译. 北京:电子工业出版社,2005.

(上接第 17 页)

恶意代码的二次攻击,RSS 阅读器通过数据库和网页保存用户订阅的信息,如果用户下载恶意的 Feed,那么以后每次用户查看该信息时,恶意代码就会被执行。恶意脚本攻击,为了丰富页面元素,允许方便地用 HTML 的元素来规定字体、颜色、图片显示等元素。很多 RSS 阅读器中自带的 Web 浏览功能,并没有完善地过滤 JavaScript 等脚本语言的功能,因此诸如页面跳转、弹出窗口、运行 ActiveX 控件、修改用户注册表等功能都很容易实现。

6 结 论

RSS 是近年来迅速普及的技术,文中从 RSS 的发展、工作原理、应用、优点、缺点等方面作了全面的阐述。与传统的阅读方式相比,RSS 有成本低、效率高、针对性强的优点,因此能广泛应用于新闻阅读、博客、电子商务网站等领域。但另一方面,RSS 存在一定的不足,如应用过于单一、版权问题、安全性问题等。

参考文献:

[1] Hammersley B. Content Syndication with RSS[M]. [s.l.]: O'Reilly, 2003.

[2] Bates C. XML in Theory and Practice[M]. [s.l.]: John Wiley & Sons, 2003.

[3] Hammersley B. Developing Feeds with RSS and Atom[M]. [s.l.]: O'Reilly, 2005.

[4] 曾宇昆,王清明,杨卫冬,等. XML 模式到关系范式的映射[J]. 计算机工程,2005,31(8):37-39.

[5] 江泽文. RSS:即将到来的互联网新革命[J]. 传媒观察,2005(9):46-47.

[6] 张德杰,高厚礼. RSS 技术及其电子商务应用分析[J]. 计算机与现代化,2005,19(11):82-84.

[7] 李子臣,王晓丽. 引擎竞争的两大焦点:RSS 技术和桌面搜索模式[J]. 中国信息导报,2004(10):54-56.

[8] 伍玉伟. RSS:网络信息"聚合"利器[J]. 图书情报论坛,2006(1):72-73.

[9] 魏 英. Internet 环境下自动新闻发布系统[J]. 计算机应用,2004,24(12):294-296.