计算机科学2004Vol.31№.3

利用面向对象技术表示知识

刘启和　杨国纬
（电子科技大学计算机学院　成都610054）

摘　要　在本文中，我们提出了利用面向对象的技术表示语义和常识的方法。将知网[3]中的义原（即最小的语义单位）表示为类，同时将概念之间、属性之间的语义关系转换为类之间关系以及类的属性之间的关系；将知网概念词典中的概念表示为所在范畴属性类的子类，并将DEF项中其它义原设置为该类中的变量。通过这种转换，知网中的语义和常识就被表示为类和对象。在实践中表明该方法是实际可行的，更重要的是为我们今后进行从文本中提取知识的研究打下基础。

关键词　知网，面向对象技术，关系，知识，自然语言处理

Represent Knowledge by Object-Oriented Technology

LIU Qi-He　YANG Guo-Wei
(College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054)

Abstract　In this paper, we suggest a method of representing semantic and common knowledge by oriented-object technology. The smemes in HowNet, which are the smallest meaning units, are represented by classes, then transform inter-concept semantic relations and the inter-attribute semantic relations into inter-classes relations and inter-attribute of classes relations; concept meaning in the HowNet's dictionary represents a subclass of categorical attribute class, the other sememes in the DEF item of the concept keep in the variables of the class. After processing, the semantic and common knowledge are represented by classes and objects. These show our method is practically feasible, and these results are critical bases of further researches of representing knowledge acquired from text.

Keywords　HowNet, OOT, Relation, Knowledge, NLP

1　引言

当前，从自然语言中提取各种语义知识已经成为计算语言学的研究热点之一，其难点之一在于构建一个能够表示知识的方法。在自然语言的早期研究中，人们在利用人工智能中的框架结构、语义网络方法基础上提出了一些语义理解模型，如语义网络、概念依存理论等。近年来，也出现了一些计算机语义知识库，如WordNet，FrameNet，HowNet等知识库。这些语义知识库为自然语言的分析和理解提供可计算的知识。其中，HowNet（知网）是董振东教授在因特网上公布的汉语和英语语义知识库，为汉语的自然语言的语义分析和理解提供了一个重要资源。目前，对知网的研究主要集中在语义的自动标注，单词的语义消歧以及利用知网研究汉语动词句法。

知网含有丰富的语义知识和常识知识，它使用知识描述语言（KDML）来表示这些知识。知网以上下位关系为核心来组织义原，并用这些义原来描述概念，形成概念词典。因此在知网中，上下位关系是组织和表示知识的核心。

我们的目标是利用面向对象的技术来表示从文本中提取的知识。既然知网含有丰富的语义知识，本文首先就用面向对象的技术来表示知网中的知识。要实现这一目标，就必须将知网中所表示的知识和各种关系（见第2节）用面向对象技术表示出来，我们的基本思想是把义原表示为类，利用继承机制来表示上下位关系，其它关系用类的变量表示出来，这些变量表示了类（即义原）之间的关联以及是何种关联。我们使用Java语言创建义原类库，在类库的基础上，建立了概念与概念之间以及概念的属性之间的内在联系，从而形成一个网状的信息表示结构。同时，将概念词典中的概念表示为其所在范畴属性的一个子类，使用这些类，我们可以方便地找到与之有关的各种知识和关系，为进一步应用打下基础，也说明了利用面向对象技术表示知识是实际可行的。

2　知网简介

知网是一个以汉英双语来表示概念与概念之间以及概念的属性之间关系的知识库。知网是通过义原来标注概念，所谓义原，是不可再分的语义单位。在知网中，包含了大约1500个义原。知网中的概念是通过义原来描述概念和概念的属性，因此，概念之间以及概念属性之间的关系表现为义原之间的关系。这些义原以上下位关系为主干，组成一个层次树结构。知网所表示的关系主要包括：(1)上下位关系；(2)同义关系；(3)反义关系；(4)对义关系；(5)属性-宿主关系；(6)部件-整体关系；(7)材料-成品关系；(8)事件-角色关系。以上关系包含在概念词典和各种特征文件中。在概念词典中，这些关系主要体现在每个记录的概念定义项（即DEF项）中（图1），所有这些，都是通过知识描述语言（KDML）来实现的。

在图1中，DEF是用义原来表示词条，第一位置表示其词条的范畴属性，这是最重要的属性，“#”表示“教师”与“occupation|职位”相关，“*”表示“教师”可以是“teach|教”的施事。第四位置表示另外的属性。因此，概念词典中词条的语义是定义在DEF项中，DEF项中表示了概念与属性、属性与属性之间的关系。

刘启和　讲师，博士研究生，主要研究方向：自然语言处理、人工智能；**杨国纬**　教授，博士生导师，主要研究方向：自然语言处理、人工智能、计算机网络。

NO.=043281
W_C=教师
G_C=N
E_C=
W_E=educator
G_E=N
E_E=
DEF=human|人,#occupation|职位,*teach|教,education|教育

图1 概念词典中的一个词条

entity|实体
├thing|万物
├ physical|物质
├ animate|生物
├ AnimalHuman|动物
├ human|人
| └ humanized|拟人
└ animal|兽
├ beast|走兽
├ livestock|牲畜
├ bird|禽

图2 知网义原的层次树

由上可知，知网的语义知识结构和 WordNet 的语义知识结构有着本质的不同，WordNet 通过同义关系将概念组织在层次树中，在知网中，概念是通过义原以及知网描述语言(KDML)来描述的，概念本身并不组成一个层次树结构，而义原被组织成层次树的结构(图2)。知网通过一个有限义原集合，使用知识描述语言(KDML)来描述概念。因此，义原成为知网描述知识的关键要素。

知网共有1500左右义原，这些义原被分成9类：1)Entity|实体；2)Event|事件；3)attribute|属性；4)aValue|属性值；5)quantity|数量；6)qValue|数量值；7)SecondaryFeature|次要特征；8)EventRole|动态角色和 EventFeatures|动态属性；9)syntax|语法。其中，前8类是最基本的，是对概念进行定义的重要义原，第8类用于动词的概念描述，其描述方法类似于格语法中的格关系，第9类是用来描述词的语法特征。

3 基本思想

知网使用一个有限的义原集合和知识描述语言(KDML)来表示概念，而义原本身被组织成一个层次树结构，义原与义原之间有各种复杂的关系。概念之间、概念与属性以及属性之间的关系是通过义原之间的关系体现出来。正因为义原的关系错综复杂，在计算机进行处理时难以进行操作，所以应将知网的各种知识结构转化为一个易于操作的结构。

在面向对象的程序设计中，其继承机制可以被描述为一个层次树结构(Java 语言是单根继承)，可以表示义原的上下位关系，因义原是通过上下位关系进行组织的，所以可以把一个义原用一个类来表示，这样义原的上下位关系就表示为对应类的继承关系，义原之间的其它关系在知网中是通过知识描述语言(KDML)表示的，这些关系我们可以在类中设置变量表示出来(图3)，从而把这些关系转化为类之间的关联。通过这种转换，就可以把知网中的各种关系利用面向对象表示出来。使用类来表示义原，就可以把语义知识表示为一个可以操作的类结构。不仅如此，类还可以提供方法，这些方法为我们操作提供了方便。此外，我们还可以利用面向对象技术中其它方法为系统提供更多的灵活性。

知网符号	类中的表示	类型
#	relation	集合类
!	定义为一个引用对象	
*	canSubject	集合类
$	canObject	集合类
%	isPartOf	集合类
+	hideRole	集合类
……	……	

图3 知网的 KDML 被表示为类的变量

在知网中，概念是用一个或多个义原来描述概念的(如图1中的 DEF 项)，这些义原与该概念的关系是由位置和知识描述语言(如“*”)决定的。其中第一位置(如图1中“human|人”)称为概念的范畴属性，是最重要的一个描述。我们将概念表示为所属范畴属性类的一个子类，另外一些义原可以设置为类的变量或为变量的初始化值，这样，一个概念就表示为一个类。对具体词所表示的概念就对应为概念类的一个对象。例如在图1中，我们创建一个“Teacher”类，它是“Human”的子类，并在类中设置3个变量来表示此类与“occupation|职位”，“teach|教”，“education|教育”的关联。

通过以上处理，就实现了知网中的知识使用面向对象的技术表示出来，同时在具体实现时，我们把整个系统设计为一个开放的形式，可以根据需要添加新的义原和概念。为其上层应用提供接口。

4 系统的具体设计

系统的具体实现使用的 Java 语言，并使用 Jbuilder 工具进行开发。

由于类库是对义原的表示，义原之间的上下位关系反映为类之间的继承关系，所有类库的组织按知网中的义原分类和组织方式进行设计，为便于今后的扩展和方便，为所有的类提供一个公共基础类“HowNet”，使用 UML 表示的部分继承图如图4。

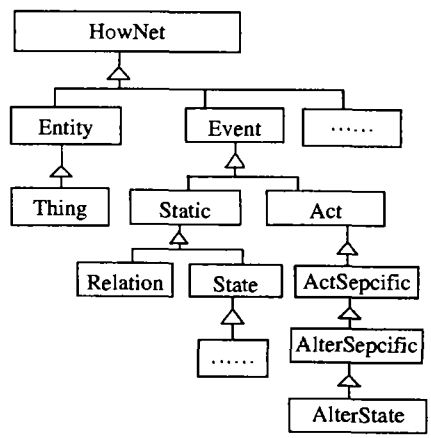

图4 义原类库的继承图

由于知网有1500左右的义原，则对应的类库中有1500左右的类，如果用人工编写类库，其工作量相当大，因此，我们编写了一个转换器，这个转换器完成从知网中提取各种关系，并将这些关系用类表示出来。在知网中，一些关系通过知网描述语言(KDML)已经显示地表示出来，我们根据其语义提取出来即可(图3)。

知网中还有一些关系是通过其位置或特殊书写方式来隐

(下转第111页)

4 实验

本文使用中国证券市场1997～2001年共1125个交易日近500支股票的收盘价时间序列作为测试集，比较了 E-Apriori 和 ES-Apriori 算法的性能。

实验使用 PIII667，512M 内存，操作系统是 Win2000 Server 的计算机。每支股票的涨跌幅度分成两段，频繁项最大长度为6，时间窗口为3，最小支持度为1%，分别使用含有5k～30k 个数据项的数据测试，结果如图4所示。

由图中可知，当项的总数小于20k 时，E-Apriori 和 ES-Apriori 的执行效率都很高。但是随着数据的增加，E-Apriori 的内存使用量将急速增加，导致运算时间骤然变长；而 ES-Apriori 无论在内存上还是在时间上都呈现平稳增加的态势。在试验中，当项的个数大于30k 后，E-Apriori 会耗尽计算机内存而无法继续运行；而 ES-Apriori 却可以顺利运行。实验结论证明，分析较大数据量的多元时间序列的跨事务关联规则时，ES-Apriori 算法在时间/空间性能上要优于 E-Apriori 算法。

结论 本文提出了一种新的多元时间序列的跨事务关联规则分析方法 ES-Apriori。由于 ES-Apriori 在计算中使用了分步策略，使得每步计算的搜索空间急剧下降，内存的开销也较 EH-Apriori 小很多。所以，在增加数据、增大时间窗口、加大涨跌幅度分段数时，算法仍然能够顺利运行。本文使用中国证券市场1997～2001年间1125个交易日，近500支股票的时间序列作为实验数据，证明 ES-Apriori 非常有效。

参 考 文 献

1 Agrawal R, Imielinski T, Swami A. Mining assocation rules between sets of items in large databases. In: Proc. of the ACM SIGMOD Conf. on Management of Data, 1993

2 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. of the 20th Conf. on Very Large Data Bases, 1994

3 Han J, Fu Y. Discovery of multiple-level association rules from large databases. In: Proc. of the 21th Conf. on Very Large Data Bases, 1995

4 Srikant R, Agrawal R. Mining generalized association rules. In: Proc. of the 21th Conf. on Very Large Data Bases, 1995

5 Kamber M, Han J, Chiang Y. Metarule-guided mining of multi-dimensional association rules using data cubes. In: Proc. of the Knowledge Discovery and Data Mining, 1997

6 Lu H, Han J, Feng L. Stock movement and n-dimensional inter-transaction association rules. In: Proc. of the SIGMOD Wrkshop on Research Issues on Data Mining and Knowledge Discovery, 1998

7 史忠植. 知识发现. 清华大学出版社，2002

（上接第107页）

式表示的。如上下位关系使用一种缩进的书写方式来表示，对义和反义关系是分别用对义和反义关系表表示的。我们首先提取出这些隐式关系，然后将这些关系表示在类中。

总之，这个转换器就是将义原的各种关系挖掘出来，并用类表示出来。在图3中，将变量"canSubject"定义为集合类是为了它能容纳多个事件。

概念词典中有62000左右词条，我们同样使用一个转换器，以把每个概念表示为范畴属性的子类（如图5）。

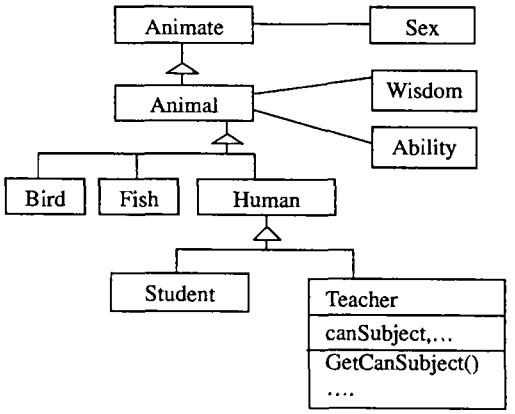

图5 概念类 Teacher 的继承关系以及部分关联

现在，假如判别"教师"是不是"动物"，如果在知网里面，需要找到"教师"的范畴属性，然后看"动物"是否是其范畴属性的上位义原，才能判断，其操作不易。但如果使用类库，其操作就在"教师"和"动物"类中进行即可完成。由此，使用对象技术表示知识后，其各种操作将得到简化并提供更高的灵活性。

另外，为便于系统的扩展，我们将为上层提供的接口从义原类库中分离出来，提供一个公共的接口，实现对各种关系的查询和判别，经过这样的设计，系统变成一个开放式的系统，便于今后的扩展和补充。

结论 本文在充分理解和分析知网的基础上，使用面向对象的技术来表示知网的语义知识，把知网中的上下位关系用继承机制表示出来，其它关系在类中设置变量来表示，进而把义原集合转化为面向对象技术中的类库，将概念词典表示为词条的 DEF 中的范畴属性类的子类。通过以上的处理，就可以将知识转化为利用面向对象技术表示的知识，使得这些知识变得容易操作，为上层的应用提供方便快捷的服务。更重要的是，通过这样的处理，说明利用面向对象的技术表示知识是确实可行的，为我们今后利用面向对象技术表示从文档中获得的知识提供基础。

利用面向对象的技术表示知网中的知识是我们从文本中提取知识的第一步，为我们今后从文本中提取知识提供丰富的资源，利用这些资源从文本中提取知识并将这些知识利用面向对象的技术表示出来，这样既可以用计算机加以处理，又为信息检索、文本过滤、信息检测提供方便使用的工具。

参 考 文 献

1 Wong Ping Wai, Yang Yongsheng. a maximum entropy approach to HowNet-Based Chinses Word Disambiguation. processing of COLING-ACL'02 . 2002

2 Yang XiaoFeng, Li Tangqu . A Study of Semantic Disambiguation Based HowNet. International Journal of computational linguistics and Chinese Language processing, 2002, 7(1): 47～78

3 董振东. 知网. http://www.keenage.com

4 周强，冯松岩. 构建知网关系的网状表示. 中文信息学报，2001，14(6):21～27

5 Wang C Y. Sense Pruning by HowNet - a knowledge-based Word Sense Disambiguation:[MPhil Thesis]. Hong Kong University of Science and Technology

6 Miller, George A. Nouns in WordNet: a lexical inheritance system. Five Papers on WordNet:[CSL Report 43]. Cognitive Science Laboratory, Princeton University, 1993

7 Baker C F, Fillmore C J, Lowe J B. The Berkeley FrameNet Projiect . In: processing of COLING-ACL'98, 1998. 86～90

8 Richardson S D, Dolan W B, Vandervende L. MindNet: acquiring and struct-uring semantic information from text. Processing of COLING-ACL'98, 1998. 1098～1092