基于本体的数据挖掘方法的研究

胡森博

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘 要 数据挖掘是个交叉领域，与人工智能、信息科学、统计分析等领域有着紧密的联系。而本体作为一个新兴的研究领域，与数据挖掘在应用的学科领域范围上有着较大的重合，比如在生物科学和化学领域，这两者的结合研究也非常活跃。在数据挖掘中引入本体能够极大地解决数据挖掘面临的问题。本文中,我们将讨论使用本体的方法来协助数据挖掘工作者在实施数据挖掘过程中对众 多可供选择的算法和方法进行选择.

关键词 知识发现，数据挖掘，本体

中图法分类号 G250.73

Research on Ontology of Data Mining Methods

Hu Senbo

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract Data mining is a cross-domain, and artificial intelligence, information science, statistical analysis and other fields are closely linked. Ontology as a new research field, and data mining applications in the field of disciplines have a greater overlap, such as in the field of biological science and chemistry, the combination of the two studies are also very active. The introduction of ontology in data mining can greatly solve the problems faced by data mining. In this paper, we will discuss the use of ontology methods to assist data mining workers in the implementation of data mining process to select a variety of algorithms and methods to choose.

Keywords KDD,Data minging,Ontology

0 引言

“本体论”最早是哲学中的基本概念,它是研究“是”之所以为“是”的理论,可以说是哲学中的哲学,甚至可以认为西方哲学自身的发展就是一个“本体论”的产生、发展、怀疑和批判的过程。近年来，本体论的方法在知识工程领域得到了越来越广泛的应用，在很多有名的知识系统中,如美国 D.Lenat 教授领导研制的大型常识知识库系统 Cyc,Pricnteno 大学 Berkeley 分校研制的语言知识库 WrodNet 等,本体论都有一定的应用。一方面,本体论研究深层次上的指示,把知识工程研究中的知识向更深更本质的方向上推进,另一方面,本体论的研究独立于任何语言,因此本体论将会为不同系统之间知识的共享和互操作提供手段。

早在 1998 年,Gruber 就已经给出了本体的一

收稿日期：2016-11-5
作者简介：胡森博（1994-）男，硕士生在读。

个流行定义,即“本体是领域概念化对象的明确表示和描述”。Guarino 把概念化对象 C 定义为:C<D,W,R>,其中 D 是一个领域，W 是该领域中相关的事务状态集合，R 是领域空间<D,W>概念关系的集合。因此，从概念化对象的定义来看，本体把现实世界中的某个领域抽象成一组概念(如实体、属性、进程等)及概念间的关系。某个领域的本体不仅提供了关于该领域的一个公认的概念集，同时也表达了各概念间所具有的各种语义联系。

随着数据挖掘技术在商业领域中得到越来越广泛的应用，对数据挖掘算法以及方法的研究也日新月异，有关数据挖掘过程各阶段的新思想、新算法、新技术层出不穷。一个简单，但典型的数据挖掘过程可能包括数据预处理阶段，数据挖掘算法的应用阶段，以及对挖掘结果可视化处理阶段。由于数据挖掘是包含多个阶段的知识发现过程，而在每个阶段都会有多个算法或方法供数据挖掘工作者选择，但仅有一些算法和方法组合是有效的。因此，即使是数据挖掘领域的专家，在一个具体的挖掘任务进行到某一个阶段时，也难免会产生困惑: 该阶段可用的技术有哪些？这些现成的技术是否合适？若不合适，采用的新技术以后能否被其他研究者使用？产生的结果是不是用户最需要的?

基于上述原因，在本文中我们将本体的概念引入到数据挖掘方法中，不同于其他基于本体论的数据挖掘方法使用本体来表示领域知识，我们是为已经存在的、被证明可以有效使用的数据挖掘技术建立本体。通过数据挖掘方法本体 ，协助不论是数据挖掘领域的新手还是专家在实施数据挖掘过程中对众多可供选择的算法和方法进行选择。

1 理论背景

1.1 数据挖掘的定义和 KDD 过程

数据挖掘是“从资料中提取出隐含的过去未知的有价值的潜在信息”(1992 年提出)，也被认为是“从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程”(2001 年提出),后者是被广泛引用的数据挖掘定义。 数据挖掘定义的变化伴随着数据挖掘方法的研究深入而变化，在商务智能领域，数据挖掘被定义为是对商业数据库中的大量业务数据进行抽取、转换、分析和其它模型化处理，从中提取辅助商业决策的关键性数据。 尽管数据挖掘的定义在变化，但其总体目标仍然是从现有的数据中挖掘未知的信息，转化和提取为可理解的信息和知，识以便进一步使用。

数据挖掘常常与数据库中的知识发现(Knowledge Discovery in Database，简称 KDD)一起出现，两者被认为是同一概念。 而文献[2]中,数据挖掘被认为是 KDD 的关键步骤。Frayyad 将 KDD 的过程分为以下几个步骤：

(1)数据选择。从数据库中选择与业务相关的目标数据。在大型数据库中，遍历所有数据是不现实且不明智的。

(2)数据预处理。根据需要去除噪声。收集必要的信息用以建模和对噪声进行说明，根据决策需

要决定需要丢弃的数据，根据时间需要等等因素选择数据。

（3）数据转化。转换数据为数据挖掘工具所需的格式。这一步可以使得结果更加理想化。具体工作根据目标任务选择数据属性，对高维数据进行降维处理等等。

（4）数据挖掘。根据任务目的选择数据挖掘算法，包括决定参数与选择合适的模型。经过这一阶段，将从数据中挖掘出模式。数据挖掘能提供的模式包括特征描述、关联分析、分类、聚类、离散值分析、演变分析等。

（5）解释模型并评价。对模式进行解释，并评价挖掘效果，根据评价决定是否进行迭代挖掘。实际研究中，数据挖掘的步骤也与上述相差不远。步骤（1）—（4）在数据挖掘中也被认为是数据挖掘的准备工作，因此在本文中将数据挖掘与KDD 视为同一概念。

1.2 本体

本体是从哲学领域引入人工智能的一个概念，1991 年由 Neches 等人最早给出 Ontology 定义，他们将 Ontology 定义为“给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的定义”。Neches 认为:“本体定义了组成主题领域的词汇表的基本术语及其关系，以及结合这些术语和关系来定义词汇表外延的规则。”随着研究的深入，很多学者对本体给出了不同的定义。 其中最著名并被引用得最为广泛的定义是由 Gruber 提出的“本体是概念化的明确的规范说明”，本体定义的发展最重要的是对本体中概念（con-ceptualization）义的提出，“概念化作为知识形式化表达的基础，是所关心领域中的对象、概念和其它实体，以及它们之间的关系”。 本体的定义有概念化、明确、形式化、共享 4 层含义。 作为知识组织的一种形式，本体是一种对知识的概念化的组织说明，是对已存在的概念和概念之间的关系的客观描述。明确意味着这些概念以及概念使用中的限制具有明确的定义，通过概念模型的表达，减少对概念和逻辑关系的误解。 本体使得计算机能够理解和处理信息的语义。 共享是指本体表示的知识是公认的知识，能被人认可。 本体可被表示为概念、属性、关系三元组。与数据挖掘类似的是，本体的活跃领域也包括信息科学、生物、化学等。 特别是在生物科学领域，基因本体的建立使得该方面的研究十分活跃，取得了丰富的研究成果。

2 基于本体的数据挖掘过程

2.1 本体的构造方法

在参考文献 [6] 中,Uschold&Grun-inger 提出一个本体构造的方法学框架, 该框架包括以下部分:(1)确定本体的目的和使用范围, (2)构造本体, 包括,本体捕获:即确定关键的概念和关系,给出精确定义, 并确定其它相关的术语, 本体编码:选择合适的表示语言表达概念和术语;(3)已有本体的集成:对已有本体的重用和修改, (4)评估:根据需求描述、能力问题等对本体以及软件环境、相关文档进行评价。文件记录每一阶段的指导准则在这个框架内, 详细描述本体捕获和形式化的本体设计和评估方法。

在基于本体的数据挖掘中，关键是建立实用，有效，能精确表达的本体。本文根据参考文献[6]提出一种自学习型的本体构造方法流程，如图 1。

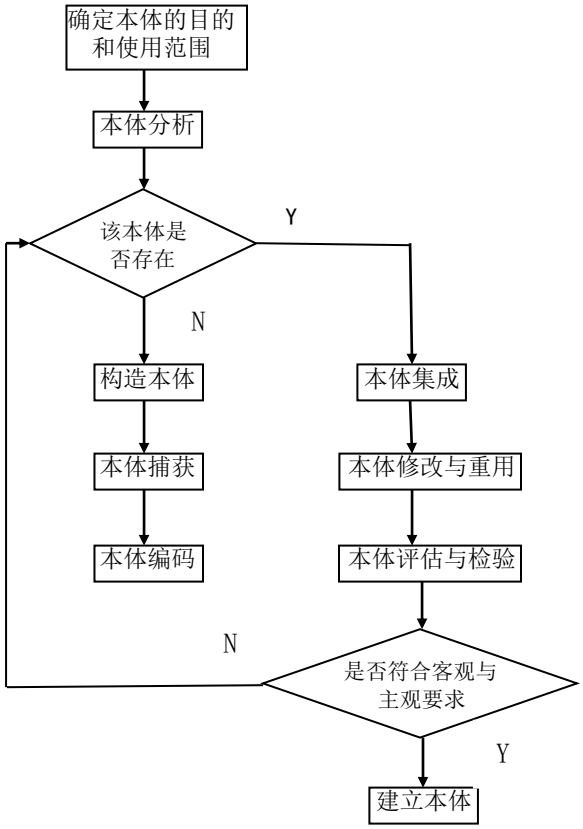

图 1 本体构造方法流程图

2.2 基于本体的数据挖掘系统模型

基于本体的数据挖掘方法，首先，利用领域知识或背景知识，将本体划分为三个层次:基础本体、核心本体、精确本体，其中基础本体是独立于领域应用逻辑的基本类型和概念结构，核心本体是符合上层定义的领域核心概念，精确本体是精确的领域概念。在精确本体层上进行数据挖掘，产生语义规则，些规则由精确本体层的精确领域概念组成；其次，该系统能够自动进行数据挖掘，利用本体进行数据预处理及后处理。

数据挖掘的过程中根据数据库中的数据和通过对现有知识库的对比提供给智能代理进行自适应性学习，一方面把精确本体存入本体库，另一方面把推理的知识存入知识库，如图 2。

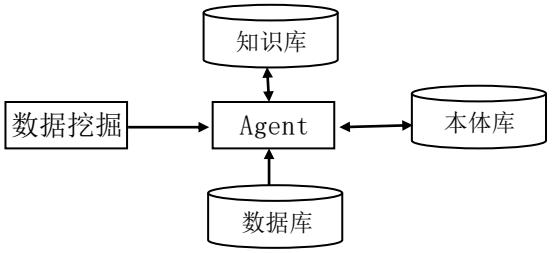

图 2 基于本体的数据挖掘系统模型

3 数据挖掘方法本体的建立

在挖掘过程中，本体是用来协助用户构成有效的 DM 过程(可执行方案)集合。数据挖掘方法本体需要定义已经存在的数据挖掘技术以及其特征属性。首先，我们对每个操作(算法或过程)建立本体，本体中包含如下信息[7] :(1)每个操作的可读信息;(2)对于每个操作，说明其执行环境，包括前提条件以及该操作和前驱操作的兼性;(3)操作执行的结果的详细说明;(4)说明阈值况;(5)对影响操作属性如速度、精度、模型复杂性的估计。APRIOR 算法、小波变换聚类方法的本体描述如图 3 所示。将所有方法本体综合在一起就可以构成数据挖掘方法的本体。本体对所有操作按照逻辑的形式进行分类，形成不同的逻辑组。在数据挖掘工作进行到某个阶段时，这些逻辑组可以用来减少构成有效 DM 过程的操作的数量。 一个有效的DM过程并不会违背本体中任何叶节点的基本约束。例如，如果输入的数据集合包含数值属性的数据，则将数据简单地应用于小波变换聚类

方法是不合理的，因为小波变换聚类方法仅仅能处理分类属性。此时可以使用离散化规则预处理数据使数据能够完成从数值型到分类属性的转换，再使用小波变换聚类方法。 数据挖掘方法本体建立后，还可以为彼此工作相互独立的挖掘工作组或工作者提供共享新成果的平台。 例如，当工作组 A 和工作组 B 彼此独立工作时，工作组 A 若发现现有的技术都不能较好地实现用户需求时，可以设计出新的算法或过程，只需将它加入到我们的本体系统中，则不论是工作组 B 还是其他的工作者以后都可以使用该新技术。 这也是我们提出为数据挖掘方法建立本体的初衷之一，即如何实现数据挖掘工作者之间的信息共享，使彼此的工作不再处于独立状态。

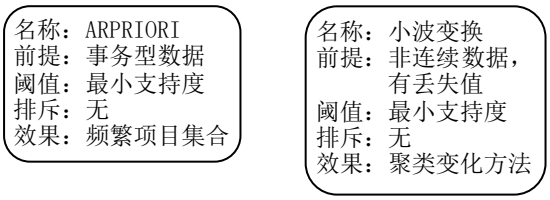

图 3 ARPRIORI 算法与小波变换聚类方法的本体描述

4 算法设计

在本节中，我们将给出使用数据挖掘方法本体生成有效 DM 过程的算法伪码。算法设计思想如下 ：我们通过生成一棵树来生成所有可执行方案。初始时，生成树只有一个根节点，在该节点中保存有初始数据特点。同时根据目标任务描述将叶节点的最小子类按照从左到右的顺序排列形成一个子类的有序集合(目标任务描述的作用是在挖掘算法本体中将彼此独立、不可能连续执行的技术排除，例如：当用户要完成关联分析任务时，则分类、偏差检测等本体将不会被包含在该集合中)，每一个子类中有若干本体。 我们对生成树的所有叶节点按照广度优先的次序完成如下操作：根据当前叶节点的数据特点，在第一个子类中所包含的方法本体中查找前提与之相一致，不存在操作排斥性的技术，有则生成一个新节点，在该节点中记录三方面信息：技术的名称，排斥信息以及执行完该技术后数据的特征，即方法本体中所描述 的效果。 新节点作为当前叶节点的子节点插入，对于同一个当前叶节点，可能会生成有多个子节点 ，同时，还需要插入一个节点，该节点在技术名称位置为空，表示没有采用该本体中的技术，其数据特点是当前节点的数据特征，当处理完生成树同一层次中最后一个叶节点时，表明在该阶段的所有可能的技术组合都已经考虑，可以进行下一阶段操作，既考虑子类集合中的下一个子类，直至将集合中的所有子类都遍历之后，我们也生成相应的生成树了。

输入：只有根节点的树 T，子类的有序集合{C_1，C_2，…，C_n}

输出：有效 DM 过程的生成树

过程：

```
for (i=1;i<=n;i++)
{leaves=getleaf(T);//
  For each r in leaves
  { for each ontology Oj in Ci
   {if (r.数据特征=Oj.前提 and r.排斥<> Oj.
        名称)
    new(t.Oj.名称,Oj.效果,Oj.排斥);
```

//生成新节点 t，名称为 O_j所代表的技术，数据特征为执行该技术后的结果add(t,r);//将t作为r的子节点加入T中

}

new(t,null,r.数据特征,r.排斥);

//生成一个没有名称的新节点，表示没有采用 C_i 中的任何技术。

add(t,r);

}

}

我们对T进行遍历生成所有最长路径，一条最长路径上的所有节点，即为一个可执行计划方案中所有细节。

5 结论

本文将本体引入数据挖掘方法中，对数据挖掘方法本体和其相关算法进行初步设计，提出一种基于本体的数据挖掘方法，目前已经在动手建立部分本体，以实现本文中所提出的算法。将来的工作重点是智能代理模块的完善，将增加智能代理的自适应性的多种算法，具有识别本体和知识的精确性。

参考文献

[1] Alon Y. Levy,Marie-Christine Rousset.Combining Horn rules and description logics in CARIN[J].Artificial Intelligence,1998,(104):165- 209.

[2]易国洪, 章瑾. 基于本体的数据挖掘方法研究[J]. 计算机与数字工程, 2007, 07:42-44.

[3] Gruber TR. A translation approach to portable ontology specifica －tions[J].Knowledge Acquisition，1993，5：199-220.

[4] Brost W N. Construction of engineering ontology for knowledge sharing and reuse[D].University of Twente，Enschede，1997.

[5] Studer R，Benjamins V R，Fensel D.Knowledge engineering principles and methods[J].Data and Knowledge Engineering，1998，25（1-2）.

[6]邹力鹃, 王丽珍, 姚绍文. 数据挖掘方法本体研究[J]. 计算机科学, 2005, 03:197-199.

[7]B．Mobasher and H．Dai．Using Ontologies to Discover Domain—Level Web Usage Profiles[J]．Proceedings of the Second Workshop on Semantic Web Mining，PKDD02，Helsinki，Finland，2002．

[8]Cingil．I.Dogac．A and Azgin．A Broader Approach to Personalization．Communications of the ACM．V01.43.No．8：136-141．

[9]王栋, 向阳, 张波. 本体在数据挖掘系统中的应用研究[J]. 计算机工程与应用, 2009, 05:11-12.