《智能信息处理》课程考试

基于本体在数据挖掘中的应用

洪川宇

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 11 月 30 日

基于本体在数据挖掘中的应用

洪川宇

[1]（大连海事大学 信息科学技术学院，大连 116026）

摘　要　本文主要讨论了本体的定义，本体的构建准则，以及构建本体的五种方法，然后介绍了数据挖掘的相关定义，步骤和方法。最后提出本体在数据挖掘中的应用。将本体引入到数据挖掘中，与传统的数据挖掘方法相比，使专业技术人员能够了解应用领域的背景知识，设计出更好的数据挖掘算法，提高了数据挖掘的效率和结果。

关键词　本体；数据挖掘；人工智能；

Application of Ontology in Data Mining

HONG Chuanyu1)

[1])(School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract　This paper mainly discusses the definition of ontology, ontology construction criteria, and five methods of ontology construction, and then introduces the relevant definitions, steps and methods of data mining. Finally, the application of ontology in data mining is proposed. By introducing ontology into data mining, compared with traditional data mining methods, professional technicians can understand the background knowledge of the application field, design better data mining algorithms, and improve the efficiency and results of data mining.

Key words　Ontology; Data mining; Artificial intelligence;

1　引言

随着现代信息科学技术的不断进步，现代社会步入了一个高度信息化的时代。巨大的信息量使得现代公司的决策变得困难，对企业的反应速度产生了巨大的影响，如何从巨大的知识库中挖掘出有意义的信息，根据信息预测未来的发展趋势，现在变得十分重要，成为当今研究的一个热点问题。数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程，人们越来越重视数据挖掘的作用和意义，

如今， 数据挖掘在各个领域已经发挥额重大的作用，如：电子商务、市场营销、零售、银行、证券等等。但是数据挖掘的专业技术人员并不具备相应的应用领域的背景知识，因而很难设计出最佳的数据挖掘算法，影响了数据挖掘的效率与效果，将本体概念和技术加入到数据挖掘中，建立背景领域的领域本体辅助技术人员来进行数据挖掘。本文主要介绍了本体的定义、构建规则、构建方法，并研究如何利用将本体技术应用到数据挖掘中，开发出更加有效的挖掘算法，提高挖掘效率和结果。

2　本体

2.1　本体定义

对于本体的定义，在西方哲学方面，本体是关于存在及其本质和规律的学硕，在中国哲学方面，本体是关于万物产生、存在、发展、变化的根本原因的学说。而在人工智能领域，本体有着更重大的意义。我们可以认为，本体即是与任务独立的知识库，即给出构成相关领域的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的外延的规则的定义。在本体发展的过程中，不同的人给出了不同的定义，其中的定义主要有三种。分别是 Gruber 定义、Borst 定义、Studer 定义。其中

Gruber 定义为本体是概念模型的明确的规范说明，Borst 定义为本体是共享概念模型的形式化规范说明，Studer 定义为本体是共享概念模型的明确的形式化规范说明，本文认为，Studer 定义是目前对本体定义概括的最全面的定义。

虽然本体有着不同的定义，但是，本体的本质含义大概是相同的，可以分为以下四种：概念化、形式化、明确、共享。通俗的讲本体可以描述为：某个领域的概念的集合、概念和概念关系的集合、以上两中集合的集合。这样的本体在共享范围内有着明确的唯一的定义，达成一种共识，便于人和及其进行交流。本体的形式化表示：

$$O = (V, C, P, H, R) \quad (1)$$

其中，V表示词汇集合，C表示各个词汇之间的关系与约束，P表示各个属性之间的关系和性质，H表示对同意词集和单词的实例声明，R表示对实例的具体描述。

2.2 构造本体的准则

本体的构造准则要遵循以下的五种准则，即清晰性，完全性，一致性，可扩展性，最小承诺（最小编码偏好）。

A）清晰性：本体必须有效地说明所定义术语的含义。定义应该是客观的并且与背景独立。当定义可以用逻辑公理表达时，它应该是形式化的，应该尽力用逻辑公理表达。定义应该尽可能的完整。所有定义应该用自然语言加以说明。

B）完全性：完全性也称完备性，当一个本体具有完全性，即不论任何时刻这个本体不需要添加其他元素，即这个本体包含的要素是全面的，这个本体也可称为完备的。

C）一致性：本体应该是前后一致的，也就是说，它应该支持与其定义相一致的推理。它所定义的公理以及用自然语言进行说明的文档都应该具有一致性。如果从一组公理中推导出来的一个句子与一个非形式化的定义或者实例矛盾，则这个本体是不一致的。

D）可扩展性：本体的可扩展性是指本体提供一个共享的词汇，这个共享的词汇应该为可预料到的任务提供概念基础。它应该可以支持在已有的概念基础上定义新的术语，以满足特殊的需求，而无须修改已有的概念定义。也就是说，人们应该能够在不改变原有定义的前提下，以这组存在的词汇为基础定义新的术语。

E）最小承诺（最小编码偏好）：本体应该处于知识的层次，而与特定的符号级编码无关。本体的表示形式的选择不应该只考虑表示上或者实现上的方便。概念的描述不应该依赖于某一种特殊的符号层的表示方法，不能依赖于某种确定的语言，因为实际的系统可能采用不同的知识表示方法。

2.3 五种构造本体的方法

本体的构建方法主要分为以下几种方法，分别为：

(1) Uschold 和 King 方法。此方法的构建顺序为 a）明确目标 b）构造本体 c）评价本体 d）完善文档四个步骤。

(2) Gruninger 和 Fox 方法。此方法的构造顺序的表示方法如图一所示。

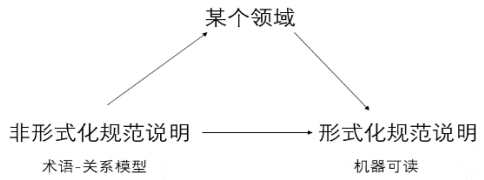

图 1. Gruninger 和 Fox 方法

(3) Berneras 方法。此方法的构建顺序为 a）目标与需求说明 b）基于高层本体的初步设计 c）本体的提炼和构造。

(4) MethOntology 方法。此方法的构建顺序为 a）目明确构造本体时采用的行为 b）规划构造本体的生命周期 c）实施每种行为/每个阶段。

(5) 基于 SENSUS 方法。SENSUS 方法的构造说明如图 2 所示。

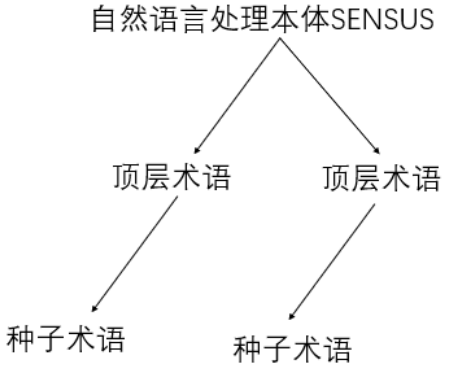

图 2. 基于 SENSUS 方法

3 数据挖掘

3.1 数据挖掘的定义

数据挖掘与人工智能和数据库中的数据是一样的，它是一个小的过程，显示准确的信息，隐藏在大量的数据挖掘领域，以前未知的和可能的主要分析了机器学习、模式识别、统计、数据库、可视化等方面的信息自动分析，帮助决策者制定市场营销策略，科学发现数据清理、信息清理、信息构建、结果表达、数据解释等三个方面都能与挖掘用户和一些学者建立联系。

数据挖掘技术是一种基于数据分析的技术，它分为三个步骤：清理、清理、调用两个系统，选择信任度设置信息，从信息，收集信息，发现规律，学习所提供的信息，不显示任何可能的规则，用户调查作为一个特殊的数据单元来分析和分析异常。

3.2 数据挖掘的过程

数据挖掘的任务就是生成如此大量的可用的数据。但是仅仅找到数据是不够的。在我们实现数据挖掘之前，我们必须一步一步来，每一步做什么，实现什么，只有有了良好的计划，数据挖掘才能有序地进行。许多软件供应商和数据挖掘顾问 STVC 提供了一些数据挖掘过程模型来指导用户对数据逐步进行挖掘。对于这个模型做出一定数量的回应，并采取最终会有用的行动。这是数据挖掘中应用的一个完整过程。

1.识别业务问题。清楚地定义业务问题并识别数字数据挖掘目的是数据挖掘的一个重要环节。挖掘最后结构是不可预测的，但要探索的问题应该是预想的。为了数据挖掘而进行的数据挖掘是盲目的，这是行不通的。

2.数据准备。(1)数据选择。搜索所有业务配对，包括相关的内部和外部数据信息，并从中选择。(2)数据预处理。调查数据的质量，准备进一步的分析，并确认。确定要执行的挖掘操作的类型。(3)数据转换。将数据转换为分析模型，即指针建立一个挖掘算法，建立一个真正适合的挖掘算法，分析模型是数据挖掘成功的关键。

3.数据挖掘。根据数据函数的类型和数据的特点在经过净化和转换的数据集上点选择相应的算法进行数据挖掘，建立数据挖掘模型。建立模型是一个迭代的过程，需要仔细研究不同的模型来确定。决定哪种模型最适合。一旦确定了预测的类型，需要为这个预测选择模型的类型。然后对其结果进行评价，并解释其价值。

4.结果分析。用分析的方法解释和评估结果，此方法通常用于数据挖掘操作，并且这项技术通常是可视化的。

5.知识同化。将分析所得到的知识集成到业务信息系统的结构组织中去。总之，数据挖掘过程需要多次的循环反复，才能达到预期的效果。

3.2 数据挖掘的方法

⑴关联规则挖掘。

1993 年，R. Agrawal 等人首先提出了用于描述数据的关联规则挖掘问题，他描述的是库中一组数据项之间的潜在关联规则，一个典型的例子是:在超市，90%的顾客都在购买面包和黄油的同时，也会买牛奶，直观的意思是:客户购买产品的可能性有多大，对于其他产品，查找所有类似的关联规则企业产销确定、产品分类设计、市场分析等关联规则是数据挖掘的主要研究课题，具有多方面的主要模式之一，重点识别找出满足给定条件的多个域之间的依赖关系关联规则挖掘对象通常是大型数据库。

⑵决策树方法。

决策树是基于不同的特征，用树形结构分类或决策集表示，规则的生成和发现利用了信息论中的相互信息(信任)，找到数据库中信息最多的字段并构建它，建立决策树的一个节点，然后根据字段的不同值构建，一棵直立的树的枝干集中在每一个枝干上，树的底部被反复地建立起来，层节点和分支的过程可以建立决策树，从示例学习最优化的角度分析，理想的决策树分为 3 种：①叶子数最少；②叶子结点深度最小；③叶结点数最少且叶子结点深度最小，寻优最优决策树已被证明是 NP 困难问题。

⑶神经网络(neural network)。

它是由大量的简单神经元，通过极其丰富和完善的连接而构成的自适应非线性动态系统，并具有分布存储、联想记忆、大规模并行处理、自组织、自学习、自适应等功能。

4 基于本体的数据挖掘

将本体应用到数据挖掘中，主要分为以下部分。

（1）建立领域本体及任务本体。建立挖掘领域的领域本体，并进行细分。从自然语言描述的问题中提取相关信息的词技术，描述知识，然后过滤掉各种无关信息，如虚词，得到了问题的信息特征，总

结了领域本体中的概念，将概念或属性与特征信息进行匹配，并按照一定的规则进行构造，创建任务本体。

（2）数据仓库的建立。不同的数据库格式一般不同，但 XML 格式一般是无歧义的，因此一般用 XML 用作标准语言，其作用是数据库之间的交流。将不同的本体转化成 XML，建立起各数据仓库之间的关系。

（3）确定数据挖掘的范围。在数据挖掘的过程之中，数据量十分巨大，因此要对数据加以区分。本体提供了一个词汇集，用来描述该领域事实，本文是利用本体的这个特点建立任务本体。

（4）数据挖掘。数据挖掘的关键是选择合适的算法。通过本体的建立，可以让机器学习更好的理解语义知识，从而开发出合适的算法。

（5）评估及反馈。对于挖掘的结果，采取一个经过良好的定义的兴趣度量来度量挖掘结果的好坏。并且采用反馈机制来重新构建任务本体，重新进行数据挖掘。

5 总结

目前，基于本体的数据挖掘研究已经取得很多成果，本文仅从相关定义方面对基于本体的数据挖掘进行了系统的阐述和比较，并简单综述了国内代表性的研究。基于当前研究成果，笔者认为，今后基于本体的数据挖掘研究还需从以下几个方向予以深入：

(1)利用本体进行数据挖掘的前提是要将被词语转换成本体中的概念词，因此，准确有效实现词语向本体概念词的映射很重要。

(2)网络自身的分布性使得各个领域，甚至是同一个领域，都必然使用自己的本体来描述数据，这就带来了本体异构问题。相似性的度量可以在同一本体内进行，也可以在不同本体内进行。因此，应加强跨本体，尤其是异构本体的构建相关研究。

(3)除了本体的结构信息和被比较概念词在本体中的位置信息，应加强基于本体实例的混合式算法研究，充分利用本体库的统计特性,更好的挖掘两类意思相近的词语。

(4)任何一个基于本体的数据挖掘算法都不可能解决所有问题，因此，要加数据挖掘技术研究，如:如何根据具体任务选择调用相关算法和确定相关参数。

参 考 文 献

[1] 张冰.基于本体的数据挖掘技术的研究[D].大连海事大学,2016.

[2] 赵玲.Ontology 综述[D].大连海事大学,2016.

[3] 崔韬世,麦范金.词语相似度计算方法分析[J].网络安全技术与应用,2012,05:55-56+72.

[4] 孙海霞,钱庆,成颖.基于本体的语义相似度计算方法研究综述[J].现代图书情报技术,2010,01:51-56.

[5] 马良荔,孙煜飞,柳青.语义 Web 中的本体匹配研究[J].计算机应用研究,2017,05:1-3.