基于本体的数据挖掘技术的研究

张冰

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘 要 本文阐述了将本体应用到数据挖掘中的科学性和可行性。根据本体的特点，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

中图法分类号 G250.73 **文献标识码** A

The Data Mining Technology Based on Ontology Research

Zhang Bing

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

Abstract This paper expounds the ontology is applied to the scientific nature and feasibility in data mining.According to the characteristics of the ontology, the ontology is introduced into the data mining, compared with traditional data mining methods, make the professional and technical personnel to understand the background knowledge in the field of application, design better data mining algorithm, and improves the efficiency and results of data mining.

Keywords Ontology, Data mining, Artificial intelligence

1 引言

随着信息技术的迅速发展，人们正在步入信息化的社会。浩瀚的信息量使得企业决策越来越复杂，严重影响了企业对市场的反应速度，如何从这浩瀚的知识中挖掘出未知的、有价值的的新知识和规律，根据现有数据预测未来的发展趋势，越来越引起人们的重视和关注，成为当今研究的一个热点问题。数据挖掘就是从大量的数据中提取或挖掘知识，近年来引起了产业界的广泛关注，如今， 数据挖掘已经应用到各个领域中[1] ，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。本文利用本体的特征，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

本体(Ontology) 最早是一个哲学的范畴 ，后来随着人工智能的发展 ，被人工智能界给予了新的定义[2] 。本体是共享概念模型的形式化规范说明 ，本体的最大好处可能是明确了概念与概念之间的关系 ，有比较健全的约束 ，数据的集成以及软件的重用在本体的思想下将变得容易实现

收稿日期：2016-11-06
作者简介：张冰（1993-）女，黑龙江人，硕士生在读。

[3] 。因此在网络知识管理中引入本体 ,使知识对象化 ,必定会给知识的集成和重用也带来益处 ,而且通过将与之匹配的知识也对象化 ,可以使与之匹配的知识对象的关系和属性得到完整和清晰的描述 ;通过这些关系和属性 ,用户可以获取更适合自己需求的知识 ,从而避免在知识获取时大量无关信息的混入[4] 。

建立某个领域的本体 , 从这个本体为出发点去引导数据挖掘过程 ,从而完善数据挖掘的进程 ,提高获取知识的效率和质量。

3. 数据挖掘

3.1 数据挖掘过程

数据挖掘的任务就是在如此海量的数据中发现有用的数据。但是仅仅发现数据那是不够的。我们必须在实施数据挖掘之前，先制定采取什么样的步骤，每一步都做什么，达到什么样的目标，有了好的计划才能保证数据挖掘有条不紊的实施并取得成功。很多软件供应商和数据挖掘顾问公司投提供了一些数据挖掘过程模型，来指导他们的用户一步步的进行数据挖掘工作。对这种模型做出一定的反应，并采取行动，最后将有用的数据转换成信息，信息变成行动，行动转换成价值。这个就是数据挖掘在应用上的一个完整的流程。本文对数据挖掘过程分步骤进行了如下研究。

1.确定业务问题。清晰地定义出业务问题,认清数据挖掘的目的是数据挖掘的重要一步.挖掘的最后结构是不可预测的,但要探索的问题应是有预见的,为了数据挖掘而数据挖掘则带有盲目性,是不会成功的.在开始知识发现之前最先的同时也是最重要的要求就是了解数据和业务问题。如果缺少了这些背景知识，就没办法明确定义要解决的问题，不能为挖掘准备数据，也很难正确的解释得到的结果。要想充分发挥数据挖掘的价值，必须要对目标有一个清晰明确的定义，即决定到底想干什么。有效的问题定义还应该包含一个对知识发现项目得到结果进行衡量的标准。

2. 数据准备。（1）数据选择。搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适合于数据挖掘应用的数据。（2）数据预处理。研究数据的质量，为进一步的分析做好准备，并确定将要进行的挖掘操作的类型。（3）数据转换。将数据转换成一个分析模型，这个分析模型是针对挖掘算法建立的，建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

3. 数据挖掘。根据数据功能的类型和和数据的特点选择相应的算法，在净化和转换过的数据集上进行数据挖掘，建立数据挖掘模型。建立模型是一个反复的过程。需要仔细考察不同的模型以判断哪个模型最适合。一旦决定了预测的类型之后，就需要为这个预测选择模型的类型。模型建立好之后，必须评价它的结果、解释它的价值。从测试集中得到的准确率只对用于建立模型的数据有意义。而在实际应用中，随着应用数据的不同，模型的准确率肯定会变化。更重要的是，准确度自身并不一定是选择最好模型的正确评价方法。

4. 结果分析。解释并评估结果，其使用的分析方法一般应作数据挖掘操作而定，通常会用到可视化技术。

5. 知识的同化。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 数据挖掘方法

数据库技术只是将数据有效地组织和存储在数据库中，并对这些数据作一些简单的分析，大量的隐藏在数据内部的有用的信息我们无法得到。而机器学习、模式识别、统计学等领域却有大量的提取知识的方法，但没有和实际应用中的

海量数据结合起来，很大程度上只是对实验数据或学术研究发挥作用，数据挖掘从一个新的角度将数据库技术、机器学习、统计学等领域结合起来，从更深层次中发掘出存在于数据内部的有效的、新颖的、具有潜在效用的乃至最终可理解的模式。由此，我们可以根据数据挖掘方法所属领域的不同将其分为数学统计方法、机器学习方法、面向数据库的方法、混合方法还有可视化技术、知识表示技术等等，具体来说，主要有以下几种数据挖掘方法:

⑴关联规则挖掘。1993 年，R.Agrawal 等人首先提出了关联规则挖掘问题，他描述的是数据库中一组数据项之间某种潜在关联关系的规则，一个典型的例子是：在超市中，90%的顾客在购买面包和黄油的同时，也会购买牛奶，直观的意义是：顾客在购买某种商品时有多大的倾向会购买另外一些商品，找出所有类似的关联规则，对于企业确定生产销售、产品分类设计、市场分析等多方面是有价值的关联规则是数据挖掘研究的主要模式之一，侧重于确定数据中不同领域之间的关系，找出满足给定条件下的多个域间的依赖关系.关联规则挖掘对象一般是大型数据库。(Transactional Database)，该规则一般表示式为：$A1 \wedge A2 \wedge \cdots Am \Rightarrow B1 \wedge B2 \wedge \cdots Bm$，其中 $Ak\ (k=1, 2, \cdots, m)$，$Bj\ (j=1, 2, \cdots, n)$是数据库中的数据项有 $Support\ (A \Rightarrow B) = P\ (A \cup B)$，$Confidence\ (A \Rightarrow B) = P\ (A\ B)$.数据项之间的关联，即根据一个事务中某些数据项的出现可以导出另一些数据项在同一事务中的出现在关联规则挖掘法的研究中，算法的效率是核心问题，如何提高算法的效率是所要解决的关键最有影响的是 Apriori 算法，它探查逐级挖掘 Apriori 的性质是频繁项集的所有非空子集都必须是频繁的。

⑵决策树方法。决策树(decision tree)根据不同的特征，以树型结构表示分类或决策集合，产生规则和发现规律利用信息论中的互信息(信息增益)寻找数据库中具有最大信息量的字段，建立决策树的一个结点，再根据字段的不同取值建立树的分枝在每个分枝子集中，重复建立树的下层结点和分枝的过程，即可建立决策树，决策树起源于概念学习系统 CLS(Concept Learning System)，其思路是找出最有分辨能力的属性,把数据库划分为多个子集(对应树的一个分枝)，构成一个分枝过程，然后对每一个子集递归调用分枝过程，直到所有子集包含同一类型的数据最后得到的决策树能对新的例子进行分类 CLS 的不足是它处理的学习问题不能太大。为此，Quinlan 提出了著名的 ID3 学习算法通过选择窗口来形成决策树.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。网络能够模拟人类大脑的结构和功能，采用某种学习算法从训练样本中学习，并将获取的知识存储于网络各单元之间的连接权中，神经网络和基于符号的传统 AI 技术相比，具有直观性、并行性和抗噪声性。目前，已出现了许多网络模型和学习算法，主要用于分类、优化、模式识别、预测和控制等领域在数据挖掘领域，主要采用前向神经网络提取分类规则。神经网络模拟人的形象直觉思维。其中，最大的缺点是“黑箱” 性，人们难以理解网络的学习和决策过程。

⑷遗传算法。遗传算法(GA :genetic

algorithms)是模拟生物进化过程，利用复制(选择)、交叉(重组)和变异(突变)3 个基本算子优化求解的技术，遗传算法类似统计学，模型的形式必须预先确定，在算法实施的过程中，首先对求解的问题进行编码，产生初始群体，然后计算个体的适应度，再进行染色体的复制、交换、突变等操作，优胜劣汰，适者生存，直到最佳方案出现为止。遗传算法在执行过程中，每一代都有许多不同的种群个体同时存在，这些染色体中个体的保留与否取决于它们对环境的适应能力，适应性强的有更多的机会保留下来，适应性强弱是由计算适应性函数 f(x)的值决定的，这个值称为适应值(fitness)适应函数 f (x)的构成与目标函数有密切的关系，这个函数基本上是目标函数的变种。

4. 基于本体的数据挖掘

4.1 基于本体的数据挖掘模型

将本体应用于数据挖掘中的数据预处理、数据准备、数据挖掘步骤中。通过领域本体的指导将各类异构的数据源转化为基于本体的标准模式并建立数据仓库，为决策支持打好基础，通过准确的定义，获取问题的语义，将决策问题准确描述，建立任务本体，使数据挖掘技术工作者在语义层次上理解专业领域知识，从而提供更好的数据挖掘算法，提高数据挖掘效率。下面图 1 提供了一个基本的模型。

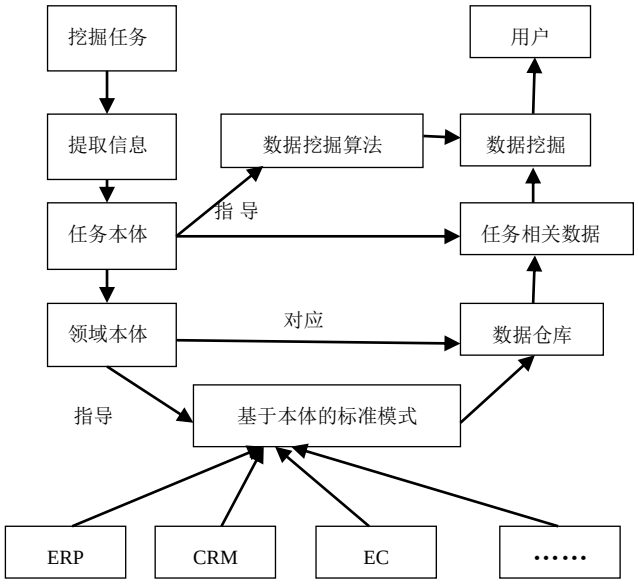

图 1 基于本体的数据挖掘模型

4.2 基于本体的数据挖掘工作步骤

（1）建立领域本体及任务本体

建立各个挖掘领域的领域本体，然后通过分词技术，从用自然语言描述的问题中提取相关的描述知识，再用过滤器过滤掉虚词等各种无关信息得到问题的信息特征，最后将领域本体中的概念或属性与特征信息进行匹配并根据一些规则建立任务本体。

（2）数据仓库的建立

在现实中，不同的数据库系统有不同的数据格式如：ERP、CRM、EC 等等。XML 定义的结构是无歧义的，且是自我描述并独立于平台的，所以可用 XML 作为各种数据库之间进行交流的媒介语言。将各种异构数据在本体的指导下转换为 XML 标准形式，然后通过领域本体之间概念关系的定义辅助确定数据仓库维之间的关系。最后将数据载入数据仓库中，为数据挖掘作准备。

（3）确定数据挖掘的范围

在数据挖掘中，数据量往往非常庞大，不加区分的对整个数据仓库进行挖掘是很不明智的，特别是由于产生的模式会随数据量的增大成指数级增长，造成挖掘过程的效率低下，并且在发现的模式当中有很多用户并不感兴趣。本体提供了用来描述某个领域内事实的词汇集，所有的领域本体构组成了该领域的知识。本文利用本体是领域知识的表示这个特点，根据步骤建立的任务本体，通过规则知道，确定挖掘的范围。

（4）数据挖掘

此步骤对已经进行了预处理了的数据仓库挖掘出用户感兴趣的知识。在此步骤中关键是选择

合适的算法，专业技术人员往往不懂专业领域的知识，而设计不出良好的数据挖掘算法，影响了挖掘效果。通过建立领域本体，使机器能够从语义的层次上理解领域的知识，从而帮助专业技术人员开发出优良的数据挖掘算法。

（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. 结束语

许多数据挖掘方法仅仅在内容上产生规则，然而背景知识的利用可以补充发现过程，从而产生具有语义的规则。本文中将本体引入到数据挖掘中，解决了因专业技术人员不了解背景领域知识而无法开发出高效的数据挖掘算法的困扰，改善了数据挖掘效果且提高了效率。

参考文献

[1] 曹尧，姜超．基于本体与搜索引擎的 web 数据挖掘 [J].广西质量监督导报 .2008,9. 总第 93 期 ,52-54.

[2] 许琳。基于本体的个性化信息服务用户模型构建研究[D].燕山大学.

[3] 李健康, 张春辉. 本体研究及其应用进展[J] .图书馆论坛, 2004(12) : 80-86.

[4] 朱廷劭,高文. KDD: 数据库中的知识发现[J] .计算机科学, 1997(6) : 55 -59.97．Sandinia：ChiaLaguna，1997．

[5] 梁凯强，陆菊康．基于领域本体与概念格的关联规则挖掘 [J]. 计算机工程与设计 .2007,7.13(28),3033-3036

[6] Kantardzic M.数据挖掘----概念、 模型、 方法与算法[M].北京 : 清华大学出版社 ,2003. [9] 卢美律.数据库里的知识发现[J] .人大复印报刊资料,1998(2): 127-130.