《智能信息处理》课程考试

数据挖掘系统对本体的应用研究

侯雅静

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

2020 年 11 月 23 日

数据挖掘系统对本体的应用研究

侯雅静

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘 要 数据挖掘的重点在于从大量数据中把未知的有价值的新知识或规律给挖掘出来，然而专业技术人员对应用领域背景知识不熟悉常常导致此过程不能顺利实施。针对该问题，可以将本体这一知识表示方法引入到数据挖掘中，利用本体的特点，设计一个基于本体的智能教据挖掘系统。本文详细介绍了数据挖掘该系统各部分的构成及工作原理。结合本体讨论了数据挖掘系统对本体的应用以及是如何构造实现的，为今后的研究发展做出了贡献。

关键词 数据挖掘系统 本体 应用

中图法分类号 TP311.20 **DOI 号** 10.3970/j.issn.1001-3695.2019.11.031

Application research of data mining system to ontology

Hou Yajing

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract The key point of data mining is to dig out unknown and valuable new knowledge or rules from a large amount of data. However, professional and technical personnel are not familiar with the background knowledge of the application field, which often leads to the failure in the implementation of this process. To solve this problem, the knowledge representation method of ontology can be introduced into data mining, and an intelligent data mining system based on ontology can be designed by utilizing the characteristics of ontology. This paper introduces the structure and working principle of the data mining system in detail. Combined with ontology, this paper discusses the application of data mining system to ontology and how to construct and realize it, which makes a contribution to the future research and development.

Key words Data mining system; Ontology; Application;

1 引言

数据挖掘就是从大量的数据中提取或“挖掘”知识。近年来，数据挖掘引起了信息产业界的极大关注，其丰要原因是各个领域所涌现出的大量数据，这些数据可以广泛使用，我们迫切需要将这些数据转换成有用的信息和知识。如今，数据挖掘已经应用到了各个领域中，例如市场营销、零售、电子政务、银行、证券、电信、体育、生物医学、DNA、气象预报等等。但是数据挖掘专业的技术人员并不具备相应的应用领域的背景知识，这给数据挖掘的进行造成了障碍。数据挖掘作为一种技术，它的生命周期正处于蓬勃发展阶段，需要时间和精力去研究、开发和逐步成熟，并最终为人们所接受。针对这种现状，将本体思想与技术引入到数据挖掘过程中，用领域本体表示领域背景知识，从而来辅助技术人员顺利实施数据挖掘。

本体概念最初起源于哲学领域，后来随着人工智能的发展，被人工智能界赋予了新的定义。1993年，Gruber 给出了本体最为流行的定义：“本体是概念模型的明确的规范说明”。后来，在此基础上，1997 年 Borst 给出了本体的另一种定义：“本体是共享概念模型的形式化规范说明”。1998 年 Studer 对上述两个定义进行了深入研究，认为本体是:“共享概念模型的明确的形式化规范说明”。从知识共享的角度来说。本体是一种概念化的显式说明，是对客观存在的概念和关系的描述，它将隐含在分析者头脑中的或在实现者的程序中的概念模型表达出来，大大减少了对问题域中概念和逻辑关系可能造成的误解。本体提供在特定域内对数据信息的一致性理解，考虑相同信息的不同含义、不同信息相同含义的差异，从而达到目标系统正确处理信息的目的。概括地来讲，本体是对共享概念形式化的明确表示，从而使计算机能够解释处理信息的语义。共享和明确是本体的两个基本特征，共享是指本体表达了公认的知识，被一组人所接受；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。

本体形式化后是概念、属性、关系的一组定义，是提供表示领域知识的一个符号集合，表示了一个特定领域中各知识库间保持不变的领域知识。如图 1 所示，节点表示的是本体中的概念，结点问的连线表示概念间的关系。

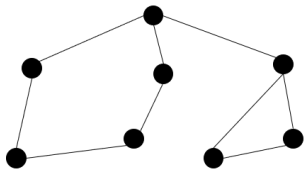

图 1　本体中概念关系图

本文利用本体的特性，研究如何将数据仓库、数据挖掘集成在一起，形成一个具有多种功能的信息系统，从而能有效实现数据挖掘功能、提高系统性能、实现集成的信息处理环境。

2 数据挖掘系统

2.1 数据挖掘功能及过程

数据挖掘相关技术的出现缓解了信息提取压力，同时也可智能化地将大量数据转化为的知识信息。数据挖掘主要基于人工智能、数据库理论、机器学习及统计学技术、模式识别等技术，是一种决策支持过程。该技术在实际应用中能对原有的数据进行高度自动化分析并做出归纳性的推理，挖掘数据中的潜在模式，预测客户行为，进而帮助决策者调整策略，避免风险及做出正确决策。数据挖掘功能在于从数据挖掘任务中需要寻找的模式类型，人们可以通过此方式预测可能发生的行为及预测未来趋势。数据挖掘的任务分为描述功能和检验功能两类，主要用来推断当前数据和预测新的数据。根据数据挖掘的任务不同，数据挖掘的任务主要有偏差检测，因为许多潜在的有意义的信息都包含在这些偏差信息中。概念描述，指浓缩当前数据信息并给出某一类对象内涵的紧凑表示信息。

2.2 从数据挖掘到知识发现

数据挖掘常常与数据库中的知识发现被认为同一概念。有相关研究指出，知识发现中的关键步骤为数据挖掘，主要有以下步骤：1. 数据选择；可从数据库中调取选择与业务相关的目标数据，然而在大型数据库中浏览所有数据是不可能的一件事，也不是明知的行为；2. 数据预处理；根据时间需要等等因素选择数据、根据需要去除噪声并通过建模的方式收集和噪声有关联的信息；3. 数据转化；为使结果更理想化可在转换数据选取数据挖掘工具所需的格式。即选择数据属性时根据目标任务并降

维处理高维数据；4. 数据挖掘；在选择决定参数和相适应的模型、数据挖掘算法时可建立在任务目的的基础上，经这阶段后可从数据中挖掘出相关模式。其中挖掘提供的模式包括演变分析。

3 本体

本体的目标是捕获其他相关领域的共有的知识，确定在这一领域之内可以共同接受的概念并在不同层次的形式化模型上明确化定义这些概念和概念之间上的相互关系，从而实现对这一领域知识的推理。本体的功能主要包括：可以复用专业领域知识；分析专业领域的知识体系结构；通过断言的形式明确专业领域内的学术观点和假设；在用户和软件代理之间达成对于信息组织结构的共同理解和认识及将专业领域的知识从运筹学、知识管理的环境中剥离出来。总的来说，本体作为知识组织的一种形式，其定义有概念化、明确、形式化、共享 4 层含义。本体能明确这些概念以及概念使用中的限制，客观描述概念和已存在概念之间的关系，使计算机能对信息语义能准确理解和处理。共享即被人认可，具体为指本体表示的知识是公认的知识。本体主要活跃在生物、信息科学等领域。

4 数据挖掘系统对本体的应用

数据挖掘会常受人为因素干扰，主要因为此项工作是不断重复的过程。在挖掘准确阶段据挖掘专家会在业务人员与数据库技术人员共同确定目标数据后选择合适的算法、模型。需要数据挖掘专家在解释阶段合理解释所选择的模型和算法。总之要有数据库管理人员、数据挖掘专家、领域专家及业务人员四种类型参与到数据挖掘过程中。尤其将专家引入数据挖掘过程中能补充和促进相关知识和知识背景，有利于挖掘，避免产生无意义的结果。传统数据挖掘有较易受数据挖掘者个人的挖掘偏好影响、脱离情境、规则过载 3 个方面的缺陷。总得来说，本体对后期研究数据挖掘系统起着积极的作用，有利于数据挖掘推算成可理解的形式。

4.1 基于本体的数据挖掘系统架构设计

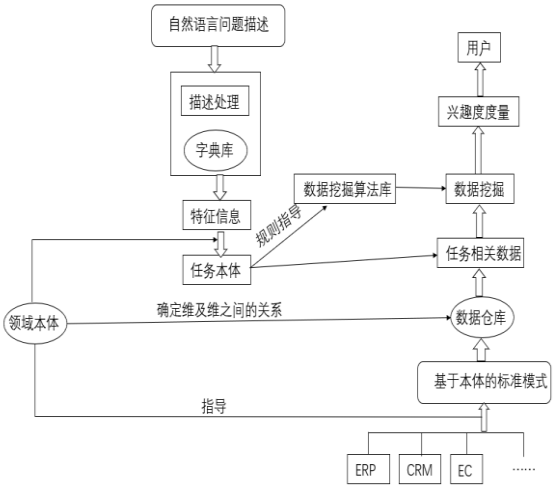

图 2 基于本体的数据挖掘系统

一般成功挖掘数据要经预处理数据、选择挖掘模型及评估发现知识。图 2（基于本体的数据挖掘系统）将各类异构的数据源通过领域本体的指导转化为基于本体的标准模式，之后可建立相应的数据仓库做好决策支持基础。通过准确定义获取问题语义，在语义层面可建立任务本体使机器准确理解决策者的意图，进而选择和所要挖掘数据相适应的模型、方法，提高挖掘效率。

4.2 本体在数据挖掘系统中的应用原理

（1）理解领域本体问题，建立挖掘任务；自然语言是描述问题的主要载体，通过过滤器将种无关信息得到问题的特征信息给过滤掉，并通过分词技术从问题描述中提取相关的描述知识。通过同领域本体中概念或者属性的匹配将特征信息组成任务本体。例如最直观且最简单的规则即在本体中将特征信息对应的概念向外关联一层或两层。

（2）构建基于异构数据抽取及数据仓库；数据库系统不同，那么数据格式也相对比较复杂，然而这些复杂的数据可通过不同的数据库系统都有各自不同的复杂格式，但这些数据 XML 标准语言展开交互。除此之外，关于 XML 定义结构没有较大的争议，即自我描述并不限制于某个体系中，归属于独立平台。XML 在一定程度上类似于 HTML，前者也会使用属性和标记。但唯一不同的是，HTML 会对每个标记和属性表示的含义给予指定，而 XML 为表达数据的逻辑结构和含义只简单运用了附加 tag，并没有定义数据文件中数据出现的具体规范。使用标记在于从表述中分离数据，由此一来，读取

数据的应用程序会全权接管数据解释权，此时 XML 会将各种异构数据转换成各种标准形式，主要因为 XML 会自动理解和规范相关程序。最后利用本体之间的概念关系具体内涵确定数据仓库维之间的关系，在数据仓库中载入数据后开始挖掘。

（3）确定挖掘数据的范围；如果在数据挖掘中不对各种类型的数据进行区分，除了会降低挖掘效率，还会由于所产生的模式可能随着数据仓库的大小指数地增长而增大挖掘量。除此之外，用户兴趣和大部分挖掘模式没任何关系，虽然领域专家会筛选可利用的属性，但筛选工作任务十分耗时费力，通常筛选过程中被遗漏的相关属性和不相关数据属性会危害系统，进而导致所选择的数据挖掘方法不能发挥应有的作用，同时还会降低发现模式质量。数据工作量也会因不相关或冗余的属性而增大，会减慢挖掘进程。

（4）依据任务本体对挖掘知识类型进行确定；在领域本体中的概念单独增加一个属性表示挖掘方法。领域本体是形成任务本体的主要来源，所以自然会涵盖相关概念的挖掘方法。但大部分与挖掘任务有关联的概念都包含在任务本体中，他们在领域本体中有不同的挖掘方法吗，因此要定义相关规则后才能确定挖掘方法。（具体方法如图 3 所示）

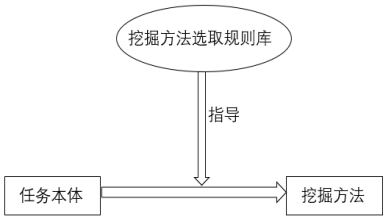

图 3 挖掘方法的选取

（5）评估领域本体对发现模式，知识发现进行指导；关于挖掘任务数据集和挖掘方法上述内容已经选取了，主要继续挖掘即可。通过兴趣度度量挖掘到的结果，有可能并不符合用户需求。此时，根据兴趣度反馈信息对任务本体进行重新构建并挖掘。

4.3 本体在数据挖掘系统中的应用实现

4.3.1 任务本体的形成

在自然语言的问题被抽取语义的过程中理解问题。机器会自动匹配和决策问题、中心词等问题领域本体定义相关的内容。尤其机器会在搜索过程中通过自身具备有用户交互以及人机接口来对搜索进行纠正。使问题中心词所在的本体能顺利被机器设备找到。根据中心词所定义的概念含义在本体中搜索信息数据，最后形成任务本体。

4.3.2 基于语义的客户数据抽取模块

解决异构问题一直是数据抽取难题。一般造成语义异构有以下方面：①由于各信息源有不同分布领域，导致其信息源中的概念存在不同的联系，这种隐含的关系不能体现出。本体信息在于认识特定领域知识，相关领域应用可借此进行相互转化和相互操作。尤其本体可通过逻辑推理获取概念之间蕴涵的关系，重点描述概念的含义和内在关系，从而有较强的表达概念语义和获取知识能力，因此可充分利用本体解决语义异构的问题；②多种术语表示可表达不同的信息源。③在不同的信息源中问一术语表达不同的含义。

4.3.3 数据挖掘模块

对数据仓库的客户数据进行处理时建立在任务本体的基础上，从大量数据中将与任务本体有关联的数据维找出，不挖掘余下数据。除了能在数据挖掘算法库中找出对应的算法，还能根据前面自然语言问题在营销领域本体中解析出的特征信息，以此找出相对应的挖掘算法，并借此挖掘所需知识，大幅度减轻数据挖掘推理时的负担。

结束语

许多数据挖掘方法仅仅在内容上产生规则，然而可以运用背景知识来补充其分析过程。本文从数据挖掘系统在应用过程中存在的问题得知，从该系统应用问题得知，在数据挖掘过程中引入以知识表达方式为主的本体可提高外挖掘系统的智能化，有利于大规模推广数据挖掘，解决了因一些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境，改善了数据挖掘效果并且提高了效率。同时该系统的实现在一定程度上也能突破传统应用时缺乏相关知识的弊端，并逐步完善不同部分的规则库，从而进一步实现基于本体的智能化数据挖掘模型的深入研究。

参 考 文 献

[1]. 朱勇,丁刚.基于开放本体的数据关联分析研究[J].数字技术与应用,2020,38(09):34-36.

[2]. 高晓灵.基于本体论知识源模型的神经网络数据挖掘技术研究[J].通讯世界,2020,27(07):55-56.

[3]. 赵斌,韩晶晶,史覃覃,吉根林,刘信陶,俞肇元.语义轨迹建模与挖掘研究进展[J].地球信息科学学报,2020,22(04):842-856.

[4]. Hiba Belhadi,Karima Akli-Astouati,Youcef Djenouri,Jerry Chun-Wei Lin. Data mining-based approach for ontology matching problem[J]. Applied Intelligence: The International Journal of Research on Intelligent Systems for Real Life Complex Problems,2020,50(6).

[5]. 陈晨.数据挖掘技术在医院信息化中的应用[J].数字技术与应用,2019,37(03):119-120.

[6]. 王志勇,王建宇.数据挖掘在知识管理系统中的研究与应用[J].航天工业管理,2019(03):38-41.

[7]黄飞.基于本体的数据挖掘技术在商务智能中的应用[J].中国商论,2018(36):24-25.

[8]. 李兴春.计算机信息检索中的本体构建研究[J].重庆文理学院学报（社会科学版）,2013,(3).87-91.

[9]. 段妍羽,巩青歌,彭圳生.基于数据挖掘的本体构建与重构技术研究[J]. 计 算 机 测 量 与 控 制 ,2017,25(8):244-247. DOI:10.16526/j.cnki.11-4762/tp.2017.08.063.

[10]. 杨怡. 基于数据挖掘的语义 web 系统设计与实现[D].电子科技大学,2014.

[11]. 闻中慧.数据挖掘中的本体应用研究综述[J].软件导刊.2012,(7).104-106.

[12]. 谢茜茜.基于本体的分布式数据挖掘系统构建[J].企业技术开发,2011,30(20):76-77.