数据库 知识获取 人工智能 (113)

计算机科学 1994 Vol. 21№. 5

48-50

大规模数据库中的知识获取

TP392

丁德恒

(中山大学计算机科学系，广州 510275)

摘 要 Knowledge discovery in large databases is a developing area which would have a lot of applications in many real fields. In this paper, the relation between knowledge dicovery and machine learning are analyzed. While we introduced its present situation and some problems to exist, the tactics we should take are proposed. In last, a system frame is given.

关键词 Database, Knowledge base, Knowledge discovery, Machine learning.

一、前言

数据是知识的源泉，拥有大量的数据与拥有许多有用的知识完全是两回事。为了有效地利用大量的公共数据，必须更好地理解这些数据，并从其中快速、准确地发现知识。这里所说的知识是指大量数据中存在的规律性(regularity)或不同属性值之间所存在的[IF THEN]规则。将所获取的知识附加于仅由事实数据(fact data)构成的传统数据库上，既可强化数据库的查询能力，又可给数据库提供推理能力，并由此可构造基于规则的大规模知识库(VLKB)。

近几年来，为了更有效地利用 VLDB 中的数据，人们越来越重视数据库的知识获取研究。机器学习领域的创始人 D. Michie 认为，如何将机器学习领域已研究开发的工具有效地用于大规模数据的分析，是一个重要而又有意义的研究课题[1]。实际上，在数据库研究领域这一问题也已被逐步重视。例如，在一个由美国国家科学基金会(NSF)发起的旨在探讨九十年代数据库的研究课题的研讨会上，从数据中挖掘知识(knowledge mining)被列为数据库研究的最重要的研究课题之一[2]。最近的 VLDB 以及其他一些与数据库相关的国际学术会议也设立了相应的专题。人工智能领域也分别召开了两次以数据库中的知识获取为主题的研讨会[3,4]，本文拟就数据库中的知识获取问题、分析其与机器学习的关系，并介绍目前的研究动态和存在的问题。最后给出解决问题的策略和展望。

实际上，从大量数据中获取知识有两层意思：一是与科学发现相关。从观测客观世界的大量实验数据(往往是数值)中发现数据的整体结构特性和数据间的函数关系，并根据统计特征推断客观世界中数据间存在的规律性。例如，从关于气体的大量实验数据中抽取与压力、体积及温度有关的数据，并由此导出 Boyle-Charle 定律。P. Langley 等人开发的 BACON 系统就是这类系统中的典型[5]。二是指，研究如何从商业数据等事实数据所构成的 VLDB 中，发现其中隐含的规律性或规则。这是一类将人工智能技术与数据库理论相融合的应用性研究课题。本文主要讨论后者。

二、作为人工智能研究的应用技术

数据库中的知识获取研究，其对象主要是数据库，将数据库中各种各样的数据进行正确的抽象或泛化，以达到从看来杂乱无章的数据中发现某种规律性或规则的目的。随着数据量或复杂程度的增大，人的能力已难以应付，要把人从这种繁杂的数据分析工作中解脱出来，重要的解决途径就是有效地利用机器学习领域现有的研究成果和技术，研制能完成这种功能的计算机系统。当然，将人的能力完全移植给计算机是难以做到的，但可实用的知识获取系统目前已得到某种程度的实现。例如，在医疗、化学、CAD/CAM 和股票交易方面，八十年代初就开始研究如何从数据库中获取有用的知识[6]，并开发出了实用性系统[7]。也提出了多种通过学习从大量数据中获取规律或规则的方法。在这种形势下，我们认为开发和应用实用知识发现系统的时机已经成熟。

数据库作为大规模数据库管理系统的对象和作为机器学习的研究对象有许多不同点，参照文[6]我们归纳成表 1。

丁德恒 副教授，系副主任，主要研究领域包括：模式识别与机器智能、人工神经网络和图象处理。

表1 数据库作为不同对象的差异

数据库管理系统的对象	机器学习的研究对象
数据库是动态更新的	数据库只是静态的数据集合
记录值是不完全的，也可能包含错误信息	实例通常是完全的，不包含噪声
各字段值一般都是数值	各事件序列是二值的
数据库通常包含数百万个记录值	数据集合通常只有数百个实例

2.1 从技巧发展到工程

在人工智能研究领域，尤其是作为其应用的专家系统研究领域，主要目标在于探索人的知识的结构并逼近之。为了达到这一目的，所采取的研究方法是，限定问题或其应用对象领域，在最大限度利用相关领域的背景知识的基础上，进一步提取深层知识，达到逼近核心知识之目的。这种方法曾获得了某种程度的成功，但缺乏通用性。对于包含大量原始数据的数据库来说，对象领域本身就难以限定，所以这种方法受到了限制。这就要求研究开发适应范围更广的知识获取方法，即提出更工程化的研究方法。

2.2 仅由正实例构成的学习

在归纳学习中，基于实例的学习（概念获取）通过归纳导出规则的逻辑基础是完全性和无矛盾性，其中的无矛盾性就是通过正实例和反实例的观测来导出有效的规则。但是，数据库中保存的数据通常只有正实例，没有反实例，所以无矛盾性条件不能得到保证。为了仅从正实例中获得知识，在构筑数据库时应附加属性间的必要约束、属性值的层次结构与包含关系。

2.3 非单调知识扩展

在从事大规模数据库维护的工程人员之间都流行"9比1法则"，即数据库中占一成的违反规则的数据，需要花费占整个系统九成的管理时间。在数据更新量大的大规模数据库中，某一时刻满足的规则在数据更新后很可能不满足了。为了解决这个问题，必须考虑处理例外事件的知识的非单调扩展（non-monotonic knowledge evolution）。在人工智能等相关领域，对这一问题的处理往往有以下方法[8]：

●例外事件处理 ●规则动态修正 ●再定义（overriding） ●缺省逻辑 ●异常谓词 ●PAC学习（probably approximately correctly learning） ●模糊逻辑

以上述方法为基础的非单调知识扩展研究引起了许多学者的注意。

2.4 多媒体带来的研究课题

人的知识根据时间地点的不同往往采用不同的媒体（如文字，图形和声音等）即多媒体表示。但是，目前的知识库，专家系统中知识获取或表示的对象往往仍是文字媒体。必须进一步开展以多媒体为对象的大规模数据库研究和多媒体信息的知识获取研究。要对多媒体信息知识库化，关键的是确立以多媒体为对象的知识表示方法，和构筑既保持数据间复杂的相关性又能有效存储的数据库框架。面向对象数据库研究给人们展示了希望。人们期待着开发带有演绎功能的面向对象演绎数据库系统。

三、面向元组和属性的算法

现有的数据库系统中以关系数据库最为典型。基于关系数据库的自动知识获取算法与数据库的结构紧密相关，可以归结为两种：

1. 面向元组的算法

着眼于关系数据库元组间满足的依赖性而求得规则的算法，称为面向元组算法，假设给定大学学生数据如表2所示，通过分析其面向元组（表中的行）的依赖性，可以发现规则：[如果是22岁的女性，则是中文系的学生。]

表2 大学学生数据

姓 名	性 别	年 龄	住 址	所属系
李 军	男	19	石家庄	计算机
刘小丽	女	22	番禺市	中 文
王志兵	男	23	保 定	数 学
王 红	女	22	花都市	中 文
戴志强	男	23	花都市	外 语
何振海	男	22	广州市	经 济

G. P. Shapiro 提出了一个面向元组的算法，该算法对所有数据进行并行检索且对各元组只要读一次[9]。其思想来源于传统关系数据库领域流行的函数依赖性研究。

2. 面向属性的算法

利用属性间概念的层次结构来求得规则的算法称为面向属性算法。J. Han 等人所提出的一种面向属性的算法[10]，是把学习过程的知识表示或概念树（concept hierarchy），对每一属性从树叶回塑到根，将各属性值用具有更一般概念的值来置换，从而导出某种规则。例如，对表2的住址属性中的"石家庄"，"保定"置换成"河北省"，"花都市"、"番禺市"和"广州市"置换成"广州市"。对所属系这一属性，可将"计算机"，"数学"置换成具有更一般概念的值"工科系"，"中文"，"外语"和"经济"置换成"文科系"。对属性值进行了一般化的元组，再通过选择，投影得到规则：[河北省的学生属于工科系，广州市的学生属于文科系]。

一般将所获取的规则分为两类，表示数据库中所有数据都满足的特征的规则称为特征规则，陈述某类概念与其他类概念差异的规则称为分类规则。例如，表示某大学学生会成员特征的规则为特征规则，而表示学生会成员与其顾问有哪些不同特征的规则为分类规则。以关系数据库为对象，J. Han 等人给出的算法就是从概念树出发，导出这两类规则[10]。该算法在使用示例学习中的泛化规则、条件削减规则和泛化树上升规则的基础上，同时采用了概念聚类方法。

四、研究进展与动态

由 G. P. Shapiro 领导的研究小组最近开发了一个知识获取平台[7]，并以数值数据所构成的商业数据库为对象，采用可视化技术和聚类分析、分类和简约化(summarization)等数据分析方法，在工作站上用 Common Lisp 实现了一个原型系统，它能发现数据之间满足的规则。他们正在开发其与各种商用关系数据库的接口，以便建造一个实用系统。

D. B. Lenat 等人在研制 CYC 知识库过程中，有效地探讨了构筑大规模知识库系统的可能性[11]。W. M. Shen 为 CYC 系统，开发了一个发现 $P(x,y) \wedge R(x,z) \rightarrow Q(x,z)$ 型规则的方法[12]。例如，"X 与 Y 非常熟悉"且"Y 是讲 Z 语言的"，那么根据常识我们可认为"X 是讲 Z 语言的"。数据之间满足的许多有用的规则都可以表示成这种形式。CYC 通过不断发现自身知识库中这类常识性知识又追加于其中的手法，使知识库功能越来越完善。

随着数据库规模的增大，字段值产生错误的概率往往会随之增大。为了解决数据库规模增大带来的这些问题，华盛顿州大学的 J. C. Schlimmer 利用学习理论，建立了一个可从数据库中发现各字段之间满足的规则，并能检测错误的模型。在此基础上开发出了实用的系统 CARPER[13]。

前面已指出了以多媒体为对象的研究之重要性，但目前基于多媒体信息的知识获取研究工作少有报导。唯有遗传因子形状抽取研究工作十分活跃，人们企图从遗传因子三维结构数据中获取形状知识。例如，D. Cohen 的研究小组开展过推测 DNA 的水解模式的研究工作，他们采用聚类分析和 AI 中决策树技术来完成 DNA 结晶的分类[14]。这一工作既属于模式识别领域的研究，也可看作与知识获取相关的研究，因为其实质就是从大量关于 DNA 结构的数据中，抽取关于其形状特征的知识。

大规模数据库中的数据量往往有数百万到数十亿字节，因此研究高效实用的知识获取算法是关键之一。为了提高算法效率，目前采用的方法有：

1. 划分可独立计算的部分，实行并行计算。

2. 不进行全数据库检索，由采样的部分数据所包含的信息导出有用的规则。

五、展望

近几年，关于数据库与知识库的异同点的讨论颇多，Y. Freundlich 曾归纳了两者的差异[15]，如表 3 所示。

从表 3 中可知，数据库中的知识获取研究是以非专家式的一般信息收集系统中的大量数据为对象，它是构造多目的大规模高级知识库的核心技术。在当今的信息化社会，数据库中的知识获取研究作为机器学习的一个应用性研究分支，有待于进一步发展。主要的研究课题包括：

- 多种知识获取技术的融合；
- 考虑用户界面的交互式系统的研制；
- 数据内在的不确定性处理技术；
- 从复杂对象数据(面向对象数据库等)中获取知识的技术。

表 3　数据库与知识库的差异

比较项目	数 据 库	知 识 库
信息收集	一般人员	专　　家
使用目的	信息检索	多种目的
信息种类	事　　实	深层信息(知识)
理论基础	计算理论	语义学解释

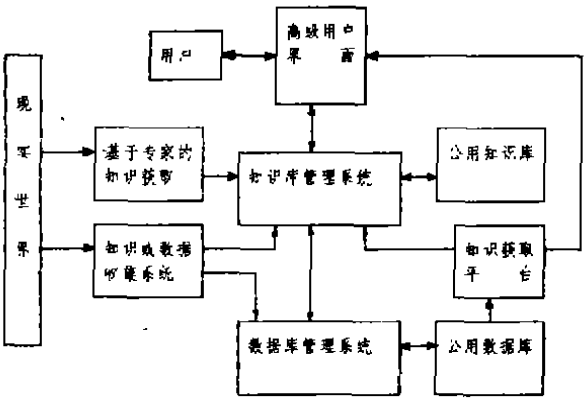

图 1　大规模知识库系统框架

笔者在此给出一个如图 1 所示的大规模知识库系统框架。在该系统框架中，不仅考虑到专家知识的获得，更重要的是从大规模数据库中获取知识，并将其反馈给系统，以形成内容更丰富的公用知识库。当务之急是研制出可从公用数据库中获取知识的软件平台。我国计算机界应尽快在各种科学基金资助范围内设立相应的研究课题，组织国家级研究开发项目，跟踪或赶上国际水平。(参考文献共 15 篇略)