

44-47

计算机科学 1999Vol 26No. 2

数据库 知识发现

数据库过程

数据挖掘

数据库中知识发现的处理过程模型的研究

Perspective of Knowledge Discovery in Database Process Model

朱廷劭¹ 高文¹ Charlex X. Ling¹

(1 中国科学院计算技术研究所 北京 100080)

(1 Department of Computer Science, University of Western Ontario London, Ontario N6A 5B7, Canada)

Abstract Knowledge Discovery in Database is a process involving several steps to extract useful knowledge from database by using of learning algorithm, data mining is one of these steps. Most of the research works on KDD pay more attention to data mining, but little to the whole process. In this paper, three kinds of process models are introduced, and we hope it is helpful for the research about KDD.

Keywords Knowledge Discovery in Database, Process Model

1 前言

数据库中的知识发现 KDD (Knowledge Discovery in Database) 是近年来随着数据库和人工智能技术的发展而出现的, 它是从大量数据中提取出可信的、新颖的、有效的并能被人理解的模式的高级处理过程^[8]。它主要采用机器学习算法或统计方法进行知识学习, 一般将 KDD 中进行知识学习的阶段称为数据挖掘 (Data Mining)。数据挖掘是 KDD 中的一个非常重要的处理步骤。人们往往不加区分地使用两者。一般来说, 在工程应用领域多称数据挖掘, 而在研究领域人们则多称为数据库中的知识发现。

人们进行的关于 KDD 的研究是为了将知识发现的研究成果应用于实际数据处理中, 为科学的决策提供支持。正是因为这样, 目前所进行的关于 KDD 的研究, 大多只着眼于对数据挖掘的研究, 而忽视了其他方面, 而事实上, KDD 首先是一个处理过程, 如果我们仅仅着重于数据挖掘, 可能就看不到实际工作中, 数据处理过程中的数据提取、组织和显示方式的难度^[7]。数据挖掘所完成的工作对于整个知识发现过程来说, 并不是占据绝大部分, 而只是较小的一部分, 因此对 KDD 中其他阶段及处理过程的研究也非常重要。目前人们对整个处理过程并没有给出非常清楚的划分, 而建立合适的处理过程模型能将各个处理阶段有机地结合在一起, 以便于人们开发及使用 KDD 应用系统。为了使得 KDD 能够

更好地应用于实践, 我们进行了关于 KDD 处理过程模型的研究, 下面分别介绍两类较具代表性的处理过程模型。

2 多处理阶段过程模型 (1)

多处理阶段模型将数据库中的知识发现看做是一个多阶段的处理过程, 在整个知识发现的过程中包括很多处理阶段。这里我们主要介绍两种面向多阶段处理过程的 KDD 处理过程模型。图 1 是 Usama M Fayyad^[1]等人给出的处理模型。在图 1 处理模型中, KDD 处理过程共分为以下九个处理阶段:

1) 数据准备。了解 KDD 相关领域的有关情况, 熟悉有关的背景知识, 并弄清楚用户的要求。

2) 数据选择。根据用户的要求从数据库中提取与 KDD 相关的数据, KDD 将主要从这些数据中进行知识提取, 在此过程中, 会利用一些数据库操作对数据进行处理。

3) 数据预处理。主要是对阶段 2 产生的数据进行再加工, 检查数据的完整性及数据的一致性, 对其中的噪音数据进行处理, 对丢失的数据可以利用统计方法进行填补。

4) 数据缩减。对经过预处理的数据, 根据知识发现的任务对数据进行再处理, 主要通过投影或数据库中的其他操作减少数据量。

5) 确定 KDD 的目标。根据用户的要求, 确定 KDD 是发现何种类型的知识, 因为对 KDD 的不同

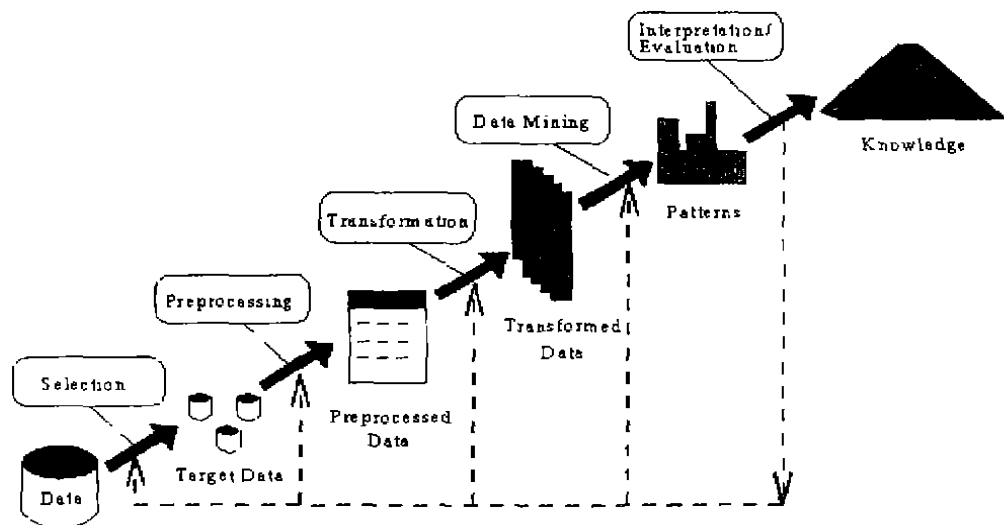


图1 多阶段的数据库中知识发现的处理过程模型(1)

要求会在具体的知识发现过程中采用不同的知识发现算法。

6)确定知识发现算法。根据阶段5所确定的任务,选择合适的知识发现算法,这包括选取合适的模型和参数,并使得知识发现算法与整个KDD的评判标准相一致。

7)数据挖掘。运用选定的知识发现算法,从数据中提取出用户所需要的知识,这些知识可以用一种特定的方式表示或使用一些常用的表示方式,如产生式规则等等。

8)模式解释。对发现的模式进行解释,在此过程中,为了取得更为有效的知识,可能会返回到前面处理步骤中的某些步以反复提取,从而提取出更有效的知识。

9)知识评价。将发现的知识以用户能了解的方式呈现给用户。这期间也包含对知识的一致性检查,以确信本次发现的知识不与以前发现的知识相抵触。

在上述的每个处理阶段KDD系统会提供处理工具完成相应的工作。在对挖掘的知识进行评测后,根据结果可以决定是否重新进行某些处理过程,在处理的任意阶段都可以返回以前的阶段进行再处理。

3 多处理阶段过程模型(2)

下面我们介绍的数据挖掘处理过程模型是George H John在文献[3]中给出的,虽然在某些地

方与上面给出的处理模型有一些区别,但这种区别主要表现在对整个处理过程的组织和表达方式上,在内容上两者并没有本质的区别。

这种模型强调由数据挖掘人员和领域专家共同参与KDD的全过程。领域专家对该领域内需要解决的问题非常清楚,在问题的定义阶段由领域专家向数据挖掘人员解释,数据挖掘人员将数据挖掘采用的技术及能解决问题的种类介绍给领域专家,双方经过互相了解,对要解决的问题有一致的处理意见,包括问题的定义及数据的处理方式。

在数据挖掘人员得到准确的问题定义和分析后,开始收集需要使用的数据,进行再加工以使得数据更适合后面的挖掘算法使用。根据解决问题的需要选择合适的挖掘算法。提取出来的知识需要向领域专家进行解释,以对知识及整个过程进行评价。图2给出了这种处理过程模型。下面对该模型中的各个处理阶段进行介绍。

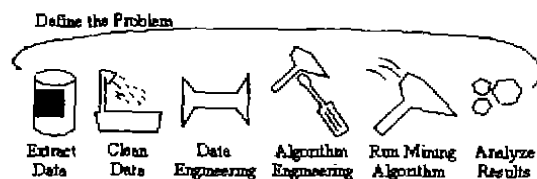


图2 多阶段知识发现的处理过程模型(2)

1)问题的理解和定义。数据挖掘人员与领域专家合作,对问题进行深入的分析,以确定可能的解决途径和对学习结果的评测方法。

2) 相关数据收集和提取。根据问题的定义收集有关的数据。在数据提取过程中,可以利用数据库的查询功能以加快数据的提取速度。

3) 数据清理。了解数据库中字段的含义及其与其他字段的关系,对提取出的数据进行合法性检查并清理含有错误的数据。

4) 数据工程。对数据进行再加工,主要是冗余属性剔除、从大量数据中选择具有代表性的数据以减少学习量以及对数据的表述方式进行转换以适于学习算法等。

5) 算法选择。根据数据和所要解决的问题选择合适的知识发现算法,并决定如何在这些数据上使用该算法。

6) 运行数据挖掘算法。根据选定的知识发现算法对经过处理后的数据进行模式提取,即数据挖掘。

7) 结果的评价。对学习结果的评价依赖于需要解决的问题,由领域专家对发现的模式的新颖性和有效性进行评价。

8) 优化。根据对结果的评价可能需要对处理过程的某些阶段进行优化。在此过程中,领域专家的参与非常重要,可以根据专业知识给出很好的改进意见。

优化包括对问题的再定义及相关数据的进一步处理。

9) 结果的使用。在实际工作中应用这些知识,为决策提供支持。如根据学习得到的知识可以设置某些触发器,当满足条件时进行特殊处理。

上面给出的处理过程模型主要强调 KDD 需要领域专家的参与,由领域的专业知识指导数据库中的知识发现的各个阶段,并对发现知识进行评价。这种模型主要是根据实际应用的需要而提出的。

4 以用户为中心的处理模型

Brachman & Anand^[1]从用户的角度对 KDD 处理过程进行了分析。他们认为数据库中的知识发现应该更着重于对用户进行知识发现的整个过程的支持,而不是仅仅限于在数据挖掘的一个阶段上。通过对很多 KDD 用户在实际工作中遇到的问题了解,他们发现用户的很大一部分工作量是在与数据库的交互上。所以他们在开发数据挖掘系统 IMACS (Interactive Marketing Analysis and Classification System) 时特别强调对用户与数据库交互的支持。图 3 给出了该模型的框图。

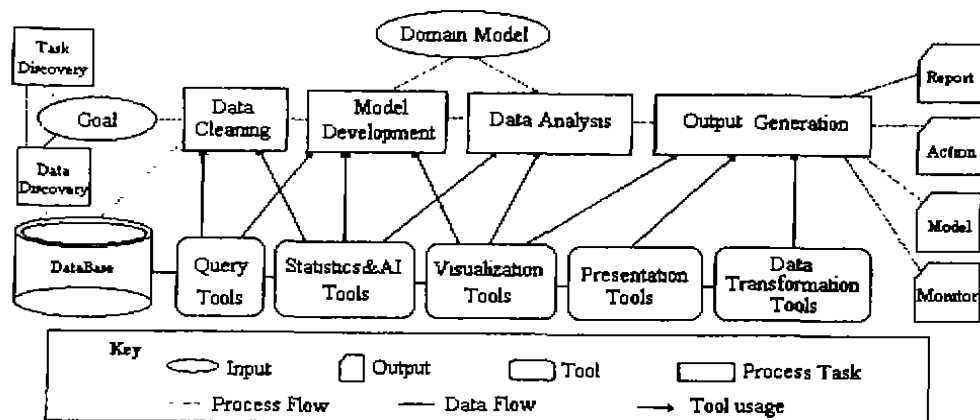


图3 以用户为中心的处理过程模型

该模型特别注重对用户与数据库交互的支持,用户根据数据库中的数据,提出一种假设模型,然后选择有关数据进行知识的挖掘,并不断对模型的数据进行调整优化。整个处理过程分为下面一些步骤:

1) 任务定义。通过与用户或用户集体的多次交流,确切了解需要处理的任务。任务定义是为了明确需要发现的知识的类别及相关数据。

2) 数据发现。了解任务所涉及的原始数据的数据结构及数据所代表的意义,并从数据库中提取相

关数据。

3) 数据清理。对用户的数据进行清理以使其适于后续的数据处理。这需要用户的背景知识,同时也应该根据实际的任务确定清理规则。

4) 模型的确定。通过对数据的分析选择一个初始的模型。模型定义一般分为三个步骤:数据分隔、模型选择和参数选择。

5) 数据分析。包括四个处理阶段:对选中的模型进行详细定义,确定模型的类型及有关属性;通过对

相关数据的计算,计算模型的有关参数,得到模型的各项属性值;通过测试数据对得到的模型进行测试和评价,根据评价结果对模型进行优化。

6. 输出结果生成。数据分析的结果一般都比较复杂,很难被人理解,将结果以文档或图表形式表现出来则易于被人接受。

该处理过程模型以用户为中心,通过对用户在进行数据挖掘过程时的工作方式的分析,在设计 KDD 系统时更注重于对用户的整个数据挖掘的全过程提供支持。

结束语 数据库中的知识发现是一个多阶段的处理过程,作为进行知识提取的数据挖掘在整个过程中虽然起着很大的作用,但其他处理阶段对于知识发现的任务来说也是非常重要的。

处理过程模型描述了 KDD 中各个处理阶段之间的关系,它对于设计实现以及使用 KDD 系统都非常重要。目前关于数据库中的知识发现的研究大多局限于数据挖掘即学习算法的研究,这对于它的研究和发展是不利的。我们应该看到数据库中的知识发现是面向实际应用并最终服务于用户,这就需要不仅在学习算法方面进行大量的研究,而且应从总体上对其进行深入的研究,设计出适合实现并易于使用的系统,它们对于 KDD 的发展是同等重要的。

参考文献

- 1 王军 数据库知识发现的研究 [中科院软件所博士学位论文] 1997
- 2 Ram A, Leake D B. A Framework for Goal-Driven Learning. In: Proc. of the 1994 AAAI Spring Symposium on Goal-Driven Learning. 1994. 1~11
- 3 John G H. Enhancements to the Data Mining Process. [Ph D thesis at Stanford University] 1997
- 4 Yoon J P, Keuschberg L. A Framework for Knowledge Discovery and Evolution in Databases. George Mason U, ISSE, 1994-July-03
- 5 Kero B, et al. An Overview of Data Mining Technologies. In: The KDD Workshop in the 4th Intl Conf. on Deductive and Object-Oriented Databases. Singapore, 1995
- 6 Chen M -S, et al. Data Mining: An Overview from Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 1997
- 7 Brachman R J, Anand T. The Process of Knowledge Discovery in Databases: A Human-centered Approach. In: Advance In Knowledge Discovery And Data Mining. AAAI/MIT Press, 1996
- 8 Fayyad U M, et al. From Data Mining To Knowledge Discovery. Same to [7]
- 9 Fayyad U M, et al. Knowledge Discovery and Data Mining, Towards an Unifying Framework. In: Proc. of 2nd Intl Conf. on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 1996

(上接第 51 页)

Rao, Georgeff 等人 1996 年提出的关于 BDI Agents 系统方法和建模技术^[5], 不难预测今后的 BDI 发展将和面向对象 (Object-Orient) 技术结合起来, 在 Agent 结构中引入类、对象和实例的概念, 并在完善 BDI Agents 技术的过程中不断扩展面向对象技术。

参考文献

- 1 Cohen P R, Levesque H J. Intention is choice with commitment. *Artif. Intell.*, 1990, 24: 213~261
- 2 Georgeff M P, Rao, A S. The Semantics of Intention Maintenance for Rational Agents. In: Proc. Int. Jt. Conf. Artif. Intell., 14th, 1995, 704~710
- 3 Jennings N R. On being responsible. In: Werner E, Demazeau Y, eds. Decentralized Artificial Intelligence. Elsevier/Holland, Amsterdam, 1992. 93~102
- 4 Jennings N R. Specification and implementation of a belief-desire-joint-intention architecture for collaborative problem solving. *Int. J. Intell. Coop. Inf. Syst.*, 1993, 2(3): 289~318
- 5 Kinny D, et al. A Methodology and Modelling Technique for Systems of BDI Agents. In: Proc. Eur. Workshop Model. Auton. Agents Multi-Agent World (MAAMAW-96), 7th, 1996. 56~71
- 6 Rao A S, Georgeff M P. Modeling rational agents within a BDI architecture. In: Allen J, Fikes R, Sandwell E, eds. Proc. of the 2nd Intl. Conf. on Knowledge Representation and Reasoning. Morgan Kaufmann, San Mateo, CA, 1991. 473~484
- 7 Rao A S, Georgeff M P. An abstract architecture for rational agents. In: Nebel B, Rich C, Swartout W, eds. Proc. of the 2nd Intl. Conf. on Knowledge Representation and Reasoning. Morgan Kaufmann, San Mateo, CA, 1992. 439~449
- 8 Rao A S. AgentSpeak(L). BDI Agents Speak Out in a Logical Computable Language. In: Proc. Eur. Workshop Model. Auton. Agents Multi-Agent World (MAAMAW-96), 7th, 1996. 42~55