第23卷 第8期 计算机工程与设计 2002年8月
Vol. 23 No. 8 COMPUTER ENGINEERING AND DESIGN Aug. 2002

文章编号：1000-7024(2002)08-0013-04 中图分类号：TP18 文献标识码：A

数据挖掘的概念、系统结构和方法

毛国君

(北京工业大学计算机学院，北京 100022)

摘 要：首先对数据挖掘的概念及相关流派加以归纳，然后给出一个数据挖掘系统的体系结构，并通过它介绍数据挖掘系统的主要功能部件，最后对数据挖掘的主要方法进行分析。

关键词：数据挖掘；知识发现

On concepts, architectures and methods of data mining

MAO Guo-jun

(Beijing Polytechnic University, Beijing 100022, China)

Abstract: In this paper, the typically descriptive concepts of data mining are first given. Then, the architecture of a data mining system is designed and its components are analyzed. Finally the useful methods of data mining are outlined.

Key words: data mining; knowledge discovery

关于数据挖掘和数据库中的知识发现的研究最近几年已经取得重要进展。和其它领域一样，数据挖掘研究经过不断的探索和发展，积累了一批成果，当然也有许多不同的观点产生。随着这种技术的日渐成熟，所涉及的概念越来越清晰，一批重要的理论和方法已经得到公认和开始应用。鉴于此，我们有必要对数据挖掘的概念和方法加以整理。本文将从数据挖掘的基本概念出发，弄清数据挖掘的各种流派的产生和发展，分析它的设计过程和主要方法，其目的是使我们对这一研究领域有个较全面和深入的认识。

1 数据挖掘的概念

1.1 数据挖掘的动机

需要是技术创新的源泉。近年来数据挖掘之所以在信息工业中受到广泛关注，主要在于大型数据系统的广泛使用和把数据转换成有用的信息及知识的迫切需要。60年代，为了适应信息的电子化要求，信息技术一直从简单的文件处理系统向有效的数据库系统变革。70年代，数据库系统的3个主要模式——层次、网络和关系型数据库的研究和开发取得重要进展。80年代，关系型数据库及其相关的数据模型工具、索引及数据组织技术被广泛采用。80年代中期开始，关系技术和新型技术的结合成为数据库研究和开发的重要标志。从数据模型上看，诸如扩展关系、面向对象、对象-关系以及演绎模型等被应用到数据库系统中。从应用的数据类型上看，包括空间、时态、多媒体以及WEB等新型数据成为数据库应用的重要数据源，相关的事务数据库、主动数据库、知识库、办公信息库等技术也得到蓬勃发展。从数据的分布角度看，分布式数据库及其透明性、并发控制、并行处理等成为必须面对的课题。进入90年代，分布式数据库理论上趋于成熟，分布式数据库技术得到广泛应用。数据仓库作为一种新型的数据存储和处理手段，成为异构数据源的集成和管理决策的制订的一种有效的技术支撑环境。

在过去40多年中，硬件技术获得稳定而惊人的进步，这导致高效计算、大容量存储、联机分析、辅助决策成为可能。虽然人们探索了许多组织和应用数据的方法，但是面对日益膨胀的数据，人们往往处于一种尴尬的境地。由于缺少有效的工具，被收集的大量数据已经远远超过我们人类理解的能力。结果是被收集在大型数据库中的数据好多已经成为“数据坟墓”，很少被访问，或者重要的决策时常是在无法利用如此丰富的数据而不得不依靠决策制订者的直觉来

基金项目：国家自然科学基金（项目编号：69883001）、北京市教委基金（项目编号：KPO7O1200102）和北京工业大学青年基金资助项目。
作者简介：毛国君（1966-），男，内蒙古人，副教授，主要研究领域为数据挖掘和分布式系统。 **收稿日期**：2001-11-20

作出。另外，考察以前知识工程的代表——专家系统，它依靠用户或领域专家来主观输入知识到知识库中。这种机制不可避免带有偏见和错误，并且它的局限性和效率成为进一步应用的障碍。要想改变这种局面，必须根本上改变人们利用数据的方式和能力。采用新的技术把数据转换成知识是大型信息系统面对的新的挑战。数据挖掘这种致力于数据分析和理解、揭示数据内部蕴藏知识的技术，自然成为成为信息技术革命的下一个目标。

1.2 数据挖掘的概念

简单地说，数据挖掘就是从大型数据集中抽取知识。从这个意义上说，叫“知识挖掘”可能更合适。但是数据挖掘这个词已经使用这么长时间了，所以我们只要把它理解成“从数据中挖掘知识”就可以了。数据挖掘的概念产生和发展可以从以下两个视点分析：

(1)作为数据分析手段

数据仓库的发展出现了OLAP(联机分析处理)应用，数据挖掘可以被看作是OLAP的高级形式，因此名词OLAM(联机分析挖掘)在一些论文中出现[1]。按此观点，大多数数据挖掘工具应该在一个集成的和已经清洗过的数据集上工作，因此必须经过数据清洗、数据转换和数据集成等预处理步骤生成可以用于挖掘的数据集。通过这些预处理所建造的数据集的最理想方式就是数据仓库。这也是为什么早期的数据挖掘总是伴随数据仓库一起出现的原因之一。然而，现在的数据挖掘概念已经远远超过为数据理解而使用数据仓库系统进行分析处理这一狭隘观点。

(2)作为知识发现手段

从这个观点看，数据挖掘是一种新型的用于发现知识或模式的技术。难怪提到数据挖掘总是让人联想起一个比它出现还早的名词——KDD(数据库中的知识发现)。关于数据挖掘和KDD的关系也有不同的表述。典型的表述有两种：(a) KDD是数据挖掘的特例[2,3]，即把用于挖掘的数据集限制在数据库这种数据组织形式上，因此数据挖掘可以看作是KKD在挖掘对象的延伸和扩展。这种观点是目前比较流行的。(b)近年来也有人把数据挖掘看作KDD过程中一个步骤[4]。既然数据仓库也是由源数据库集成而来的，即使是像WEB这样的数据源恐怕也离不开数据库技术来组织抽取的信息，因此KDD是一个更广义的范畴，它包括数据清洗、数据集成、数据选择、数据转换、数据挖掘、模式生成及知识表达等一系列步骤在内的知识发现过程。数据挖掘作为KDD的重要步骤，通过和用户及知识库的互动达到从大数据集中获得知识的目的。这种观点有它的合理性，它把知识发现的整个过程看作是基本构件的系统化协同工作，而把数据挖掘看作是整个过程的一部分。一些相关的概念，如知识抽取、模式评估、数据捕捞、异构集成等的出现，表明人们对知识发现的过程研究向着更深入的方向发展。当然，不论数据挖掘的广义或狭义概念，其实质都是从数据中挖掘有用的知识。

值得注意的是，近几年来数据挖掘概念的延伸更加广泛，所涉及的数据种类日益多样化，不仅有结构化的数据库表单，而且半结构化及无结构的多媒体、文本及WEB等也得到广泛研究。

1.3 用于数据挖掘的数据类型

从原理上说，数据挖掘应该可以应用到任何信息存储方式。近年来的研究表明，可用于数据挖掘的数据对象有关系数据库、数据仓库、事务数据库、普通文件和WEB数据以及面向对象数据库、对象-关系型数据库、空间数据库、时态数据库、文本数据库和多媒体数据库等。由于挖掘的挑战性和技术会因为存储系统的不同而不同，因此弄清这些数据对象的特点是研究相关挖掘系统结构和挖掘算法的前提。

(1)关系型数据库

关系型数据库是一系列数据表的收集。它本身的发展是相当成熟的，它有成熟的语义模型(像实体-关系模型)，有成熟的DBMS(像Oracle)，有成熟的查询语言(像SQL语言)，而且有一批可视化的工具可以使用或借鉴。在一个或多个关系型数据库中直接挖掘出所需要的知识，是数据挖掘研究的最早课题，而且已经在金融等众多领域得到应用。

(2)数据仓库

数据仓库中的数据是按主题来组织的。存储的数据可以从历史的观点提供信息。面对多数据源，经过清洗和转换后的数据仓库可以为数据挖掘提供理想的发现知识的环境。假如一个数据仓库模型化成一个多维数据模型或多维数据立方体来存储，那么基于多维数据立方体的操作算子可以达到高效率的计算和快速存取。虽然目前的一些数据仓库辅助工具可以帮助完成数据分析，但是发现蕴藏在数据内部的知识模式及其按知识工程方法来完成高层次的工作仍需要新技术。因此，研究数据仓库中的数据挖掘技术是必要的。数据挖掘不仅伴随数据仓库而产生，而且随着应用深入产生了许多新的课题。

(3)事务数据库

一个事务数据库包含一系列的事务。事务数据库可以直接应用到诸如采购、市场调查等这些商业活动中。同时，事务对于任何数据存储形式的管理都是重要的。因此，从事务数据库中发现知识成为数据挖

掘中的较活跃部分[5,6]。对事务数据库的挖掘，人们可以获得动态行为所蕴藏的模式，进而指导以后的商业行为。

(4)在关系模型基础上发展的新型数据库

面向对象、对象-关系以及演绎等新型数据库也成为数据挖掘的新的研究对象。随着数据库技术的发展，这些数据库系统诞生并发展以满足新的应用需求。在这些新型数据库系统上的数据挖掘成为不可回避的挑战性课题。

(5)新的数据库应用带来数据的多样性

新的数据库应用包含处理空间数据、时态数据、工程设计数据和多媒体数据等。这些应用需要高效的数据结构和可用的处理复杂结构、长变量记录、半结构或无结构数据的方法。在这些数据集或数据库上的知识发现工作为数据挖掘提供了丰富的研究及开发土壤。

(6)Web 数据

Web数据是复杂的，有些是无结构的(如Web页)，通常都是用长的句子或短语来表达文档类信息。有些可能是半结构的(如 Email,HTML 页)。当然有些具有很好的结构（如电子表格），揭开这些复合对象蕴涵的一般性描述特征成为数据挖掘的不可推卸的责任。近年的研究已经在基于内容/关键词、关联/链接、对象聚集等方面取得进展[2,3]。

2 数据挖掘系统的体系结构

一个数据挖掘系统是从被挖掘的数据集中形成特定知识表示过程的实现机制，因此它与被挖掘的数据组织形式和所采用的知识表示及推理方式有关。数据挖掘系统体系结构的研究可以依据数据挖掘的过程来探讨系统的主要功能部件及其相互联系，为具体应用提供指导。这种从普遍到一般的方法取得了一批成果，对数据挖掘系统的发展和应用起到了推动作用。同时也推动了数据挖掘辅助工具的开发和使用。另一方面，针对特定的领域、采用特定的方法而研制的数据挖掘系统可以快速而准确地解决特定的问题，这些系统在银行业、天空测量及生产和销售业等领域得到应用。本文将在已有的数据挖掘系统的体系结构及其相关功能部件的基础上[1-5]，设计一个具有综合功能的数据挖掘系统模型。在此基础上介绍数据挖掘系统的可选功能部件，从中可以更清楚地了解数据挖掘系统应具有的基本功能和扩展功能以及数据挖掘的较完整实现过程。

图1给出一个具有综合功能的数据挖掘系统的体系结构，下面将分别介绍相应功能构件或挖掘工具。

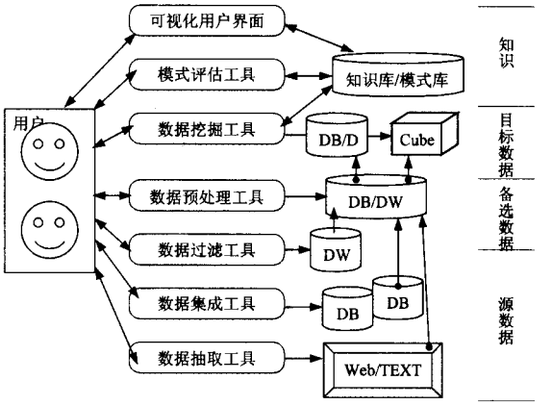

图 1 数据挖掘系统的体系结构示意图

(1)源数据存储体

源数据可以是一个或多个数据库、数据仓库及像Web等这样的其它信息存储源。它们必须经过集成/过滤/抽取来形成易于进一步处理的备选数据。这个过程用户可通过指定约束条件等达到快速形成格式化数据的目的，其结果可用数据库或数据仓库形式存储。

(2)数据预处理工具

备选数据可以在用户的指导下，使用数据清洗和数据转换技术，生成目标数据，这些目标数据可以用数据库、数据仓库以及数据立方体等形式存储。

(3)知识库或模式库

它存储挖掘出来的中间或最终结果模式或知识，也存储用于指导搜索或评估结果模式的领域知识。这样的知识可能包含不同的抽象层次、适应于不同粒度的知识、多种模式、多种知识表示形式等。

(4)数据挖掘工具

这是把数据抽象成知识的一个重要构件。它应具有知识推理机的功能，能反复利用已获得的知识和用户互动，达到最终形成知识模式的目的。理想情况应包含诸如描述、关联、分类、簇类分析以及进化和偏差分析等在内的功能模块。

(5)模式评估工具

对于一个多策略系统来说，探索并最终选定知识模式是件重要的工作。可结合现在广泛采用的诸如可信度等的兴趣测度[4]方法，达到和数据挖掘工具相互作用以集中搜索感兴趣的模式。也可以使用兴趣阈值来过滤发现的模式。根据被使用的数据挖掘方法，模式评估功能也可以集成到数据挖掘工具中。为提高挖掘效率，模式评估工作应尽可能深入到挖掘的不同层次中，这样可以保证搜索限制在感兴趣的模式中。

(6)可视化用户界面

数据挖掘系统必须允许用户通过指定数据挖掘查询或任务，对有经验的用户来说，理想情况应允许

他们使用约束条件等形式指导不同阶段的工作。通过用户和知识库/模式库的互动，帮助系统聚焦搜索。在需要时，实现基于中间结果的探索性数据挖掘工作。用户也可以通过它直观而多样化地显示任务的执行结果。另外，可以通过允许用户浏览数据库、数据仓库、知识库等，帮助用户了解系统的状况。

虽然市场上已经有许多所谓的数据挖掘系统或工具，但是不是所有的都能实现真正的数据挖掘工作（全部的或典型的）。一个不能处理大型数据的数据分析系统应该属于机器学习系统、统计数据分析工具或实验系统原型。一个仅能完成包括发现聚类值等的信息检索系统应该更靠近传统的数据库系统或信息检索系统。完成在大数据集中的演绎查询系统，应该归类到演绎数据库系统中去。

为了更直观地了解数据挖掘系统或工具的现状和技术，表1给出了一些数据比较有代表性的数据挖掘原型系统或工具。

表1 数据挖掘原型系统或工具

名称	研究机构或公司	主要特点
DBMiner	Simon Fraser	多规则（特征/序列/关联）；多层次交互
Quest	IBM Almaden	面向大数据集，多目标（关联规则/分类）
Explora	GMD(Klosgen 等)	多模式（规则/聚类/预测/统计）
IBM Intelligent Miner	IBM	包含多种技术的辅助挖掘工具
Darwin	Thinking Machines	基于神经网络的辅助挖掘工具
DataEngine	MIT GmbH	使用模糊逻辑、神经网络及信号处理等技术
ReMind	Cognitive System	基于实例推理和归纳逻辑的辅助挖掘工具

3 数据挖掘方法

3.1 挖掘知识或模式的方法

数据挖掘的目的是发现知识，知识要通过一定的模式给出。可用于数据挖掘系统的知识表示模式是丰富的。最新的研究向着多种模式综合于一个系统的方向发展。在一些情况下，用户可能并没有对感兴趣的模式有确定的想法，因此可能喜欢同时搜索几个不同的模式类型。因此一个理想的数据挖掘系统应该能挖掘不同的模式以满足不同的用户期望和应用特点。进一步，数据挖掘系统应该能发现在不同粒度中的模式。同时，既然一些模式可能并不是对一个目标数据集中的所有数据是有效的，那么模式的兴趣测度一般有总是和模式一起存在。下面对近年来出现的数据挖掘的主要方法以及知识模式类型给出简述。

(1) 概念描述

数据可以和概念联系在一起。概念描述类的方法总是用特定的模式（如概念树/偏差分析等）把数据蕴藏的特征化知识或对照类间的比较信息组织起来。可以采用基于属性归纳、基于数据立方体分析、神经网络以及机器学习等技术实现[2,4,7,8]。基于概念层次的数据挖掘可以为不同粒度的应用提供理想的知识表达和推理手段。

(2)关联规则

形式化为X->Y的关联规则被解释成在数据集中满足X条件的也很可能满足Y。它通过数据的分析和归纳挖掘出数据集中蕴藏的关联信息。关联规则的研究和应用是数据挖掘中最活跃和比较深入的分支，许多关联规则挖掘算法已经被提出。Apriori算法[5]是提出最早而且最成功的发现关联规则的算法，由Agrawa等人建立的基本原理：频繁项目集的子集是频繁项目集或非频繁项目集的超集是非频繁项目集，一直是优化关联规则挖掘算法的重要理论基础。其它比较有代表性的是Partition，Sampling和DIC。近年来作者也在这方面做了一些工作，由于其成果尚未正式发表，不宜在参考文献中列出。

(3)分类及聚类分析

分类能被使用来预测数据对象的类标识。聚类可以被用来产生类标识。可以通过分类规则（IF-THEN）、决策树、数学公式和神经网络等表示知识。所获得的知识可以用于决策、预测及识别等进一步应用中。另外，有关机器学习和模糊及粗糙集的一些研究成果可以被应用。

(4)进化分析

数据进化分析主要描述随时间改变的数据蕴藏的规律和趋势。显然它可以用诸如规则、分类、聚类等手段完成基本功能。但是这样的分析有它鲜明的特色，因此近年来如时间序列数据分析、模式匹配的序列、基于类似度的数据分析及其遗传算法等技术和方法得到研究和发展。

(5)其它方法

数据挖掘是一个多学科交叉研究领域。近年来的研究向着多方法集成化的方向发展。除了上面提到的技术外，还有神经网络、模糊或粗糙集、归纳逻辑程序设计、空间数据分析、模式匹配、图象分析、符号处理、Web技术、以及经济学、商业科学、生物信息学及心理学等相关方法。

3.2 数据挖掘系统的分类

由于数据挖掘涉及的范围很广，因此使得提供一个清楚的数据挖掘系统的分类是困难的。但是分类可以帮助潜在的用户区分数据挖掘系统以适合不同的需要。因此，本文将从不同角度刻画数据挖掘系统的分类。

(1)根据数据挖掘类型分类：如上所述，可以分成关系型数据库、事务数据库、面向对象数据库、对象-关系数据库和数据仓库、空间数据库、时态数据库、文本数据库、多媒体数据库及Web挖掘系统等。由于它们所采用的技术各有特点，因此这种分类可以帮助我们了解和研究对应数据类型的数据挖掘技术与算法。

(2)根据挖掘知识模型分类：如上所述，可以分成概念描述、关联规则、分类/聚类及数据进化模型等。进一步，由于数据挖掘系统挖掘的知识粒度和水平的不同，其知识表示形式也有所差异。

(3) 根据实现的技术类型分类：这些技术可以根据用户的互动程度(如自治系统、互动探索系统、查询驱动系统)或采用的数据分析方法(如面向数据库、面向数据仓库、机器学习、统计学、可视化、模式识别、神经网络等)来描述。一个复杂的数据挖掘系统时常采用多种数据挖掘技术。

(4) 根据应用的特点分类：数据挖掘可以广泛应用到财经、电信、军事、商贸、股票市场等几乎所有的领域。不同的应用领域时常需要相关方法的集成。因此，一个通用的数据挖掘系统可能不适合于特定领域的挖掘任务，量身制作的数据挖掘系统应更为实用。

总之，数据挖掘工作是一个复杂而系统化的过程，它的发展是信息技术革命的必然趋势。在它发展过程中，产生了许多新概念和技术，并且随着它的研究深入，一些概念和方法趋于清晰。同时它的出现为许多技术和方法提供了丰富的应用土壤。本文从数据挖掘的概念、体系结构和主要方法3个方面较全面地介绍了数据挖掘的研究成果。

参 考 文 献:

[1] Han J. Towards on-line analytical mining in lare database sysytems[C]. ACM-SIGMOD,1998. 97-107.

[2] Chen M.S.,et al. Data mining: An overview from a database perspective[J]. IEEE Transactions on knowledge and Data Engineering. 1996,8(6.):866-883.

[3] Fayyad U.M., et al. Advances in Knoeledge Discovery and Data Mining[C]. Cambridge, MA:AAAI/MIT Press,1996.

[4] Han J., et al. Data Mining: Concepts and Techniques.Morgan Kaufmann Publishers[M].(高等教育出版社，影印版),2001.

[5] Agrawal R., et al. Mining association rule between sets of items in large database. Proceedings of the ACM SIGMOD International Conference on Management of Data[C]. Washington, D.C., ACM Press, 1993. 207-216.

[6] Agrawal R. et al. A tree projection algorithm for generation of frequent itemsets[J]. In Ijournal of parallel and Distributed Computing. 2000.

[7] Ankerst M., et al. Visual classification: An interactive approach to decision tree construction[C]. In Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data. Philadephia, 1999. 46-60.

[8] Wile R. Restructuring lattice theory: An approach based on hierarchies of concepts[C]. In: Rival I ed. Ordered Sets. Dordrecht: Reidel. 1982. 445-470.

[9] Han J., et al. Mining Frequent Patterns without Candidate Generation[C]. SIGMOD Conference, 2000. 1-12.

[10] Bayardo R.J., et al. Efficiently mining long patterns from databases[C]. In: proc. 1998 ACM-SIGMOD Int. Conf. Management of Data. Seattle, 1998. 85-93.

(上接第 12 页)

[3] Govindarajan R, Nemawarkar S, LeNir P. Design and performance evaluation of a multithreaded architecture [J]. In First IEEE Symposium on High-Performance Computer Architecture, 1995,(1): 298-307.

[4] Hammond L, Nayfeh B A, Olukotun K. A single-chip multiprocessor[J]. IEEE Computer, 30(9):79-85, September 1997.

[5] Hirata H, Kimura K, Nagamine S, et al. An elementary processor architecture with simultaneous instruction issuing from multiple threads[J]. In 19th Annual International Symposium on ComputeArchitecture, 1992,(5): 136-145.

[6] Laudon J, Gupta A, Horowitz M. Interleaving: A multithreading technique targeting multiprocessors and workstations [J]. In Sixth International Conference on Ar-chitectural Support for Programming Languages and Operating Systems, 1994(10): 308-318.

[7] Palacharla S, Jouppi N P, Smith J E . Complexity-effective superscalar processors[J]. In 24th Annual International Symposium on Computer Architecture, 1997(6): 206-218.

[8] Tullsen D M, Eggers S J, Emer J S ,et al. Exploiting choice: Instruction fetch and issue on an implementable simultaneous multithreading processor[J]. In 23rd Annual International Symposium on Computer Architecture, 1996, 5: 191-202.

[9] Tullsen D M, Eggers S J, Levy H M. Simultaneous multithreading: Maximizing on-chip parallelism[J]. In 22nd Annual International Symposium on Computer Architecture, 1995, (6):392-403.

[10] John S Seng, Dean M Tullsen, George Z N Cai. Power-Sensitive Multithreaded Architecture [R]. Publiudhed in proceedings of the 2000 International Conference on Computer Design, 2000.