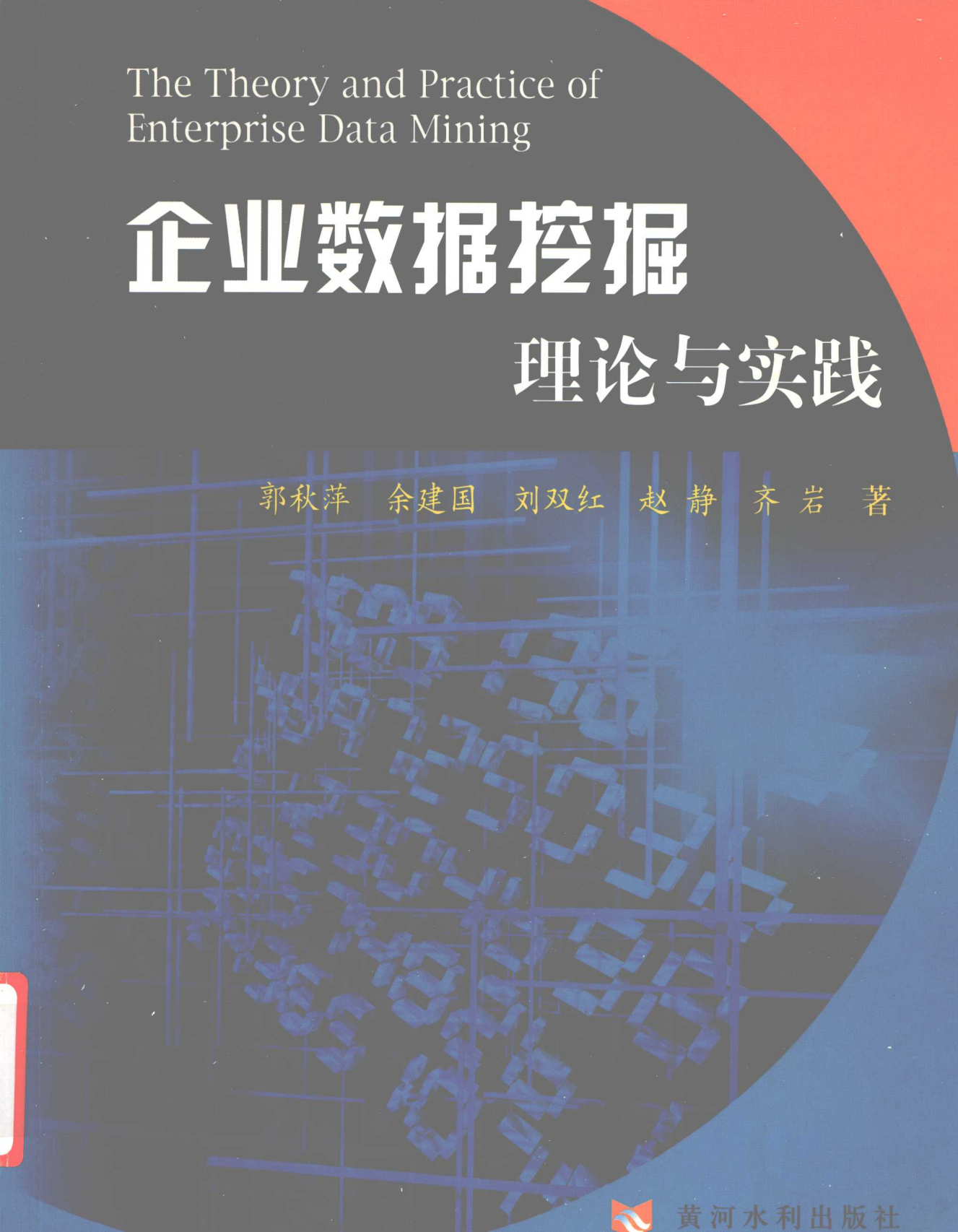
The Theory and Practice of Enterprise Data Mining

企业数据挖掘理论与实践

郭秋萍 余建国 刘双红 赵 静 齐 岩 著



黄河水利出版社

企业数据挖掘理论与实践

郭秋萍　余建国　刘双红　赵静　齐岩　著

黄河水利出版社

内 容 提 要

本书概述了数据挖掘的基本概念、应用领域、相关学科、发展趋势以及数据仓库和OLAP技术，着重讨论了数据仓库和数据挖掘在企业管理中的应用及其构建策略，并基于SQL Server提供一个具体实例，阐述了企业数据仓库和数据挖掘的实施过程，最后对前端数据挖掘应用程序开发时可选用的开发工具——Delphi在数据挖掘应用方面提供的支持进行了介绍。本书可供从事数据仓库和数据挖掘开发与设计的人员以及高等院校师生阅读和参考。

图书在版编目(CIP)数据

企业数据挖掘理论与实践/郭秋萍等著. —郑州：黄河水利出版社，2005.4

ISBN 7-80621-901-3

Ⅰ.企… Ⅱ.郭… Ⅲ.数据采集-计算机应用-企业管理 Ⅳ.F270.7

中国版本图书馆CIP数据核字(2005)第018740号

出 版 社：黄河水利出版社

地址：河南省郑州市金水路11号 邮政编码：450003

发行单位：黄河水利出版社

发行部电话及传真：0371-66022620

E-mail：yrcp@public.zz.ha.cn

承印单位：黄河水利委员会印刷厂

开本：787mm×1 092mm 1/16

印张：17.5

字数：312千字 印数：1—1 000

版次：2005年4月第1版 印次：2005年4月第1次印刷

书号：ISBN 7-80621-901-3/F·61 定价：33.00元

前　言

随着计算机技术和信息技术的发展，企业在生产、销售等各种企业活动中产生并积累了大量的数据，信息量的增长速度呈现指数上升，在超级市场、银行、电信等企业表现得更为突出，数据的容量已达到了TB级甚至PB级。这些海量数据中隐藏着大量具有潜在价值的信息，传统获取和分析知识的方法已远远不能满足企业获取这些信息的需要。“数据丰富，知识贫乏”的矛盾进一步加剧。如何从数据中发现有价值的知识或信息，就成为企业一项非常艰巨的任务。人们迫切需要一种能够从海量数据中提取知识和信息的技术，以便能够智能地、自动地把信息和数据转换成知识。这样，数据挖掘技术就应运而生了。

我国数据挖掘技术研究开始于20世纪90年代，经过十几年的发展，这一领域目前正处于蓬勃发展时期。课题组开始酝酿和申报此课题时，关于数据挖掘的中文资料，无论是图书还是期刊都比较少；关于企业数据挖掘应用的资料更是少之又少。但短短两三年的时间，数据挖掘技术在我国得到了迅速的发展，同时在计算机界、信息管理界得到广泛的重视和研究，多种国家科研基金如国家自然科学基金、863计划、“九五”计划等都对数据挖掘项目进行了资助，取得了许多研究成果。

企业数据应用的目的性很强，这就决定了基于数据仓库进行数据挖掘成为企业数据挖掘应用的主流。但由于数据仓库、数据挖掘技术都是数据处理及分析领域出现的新技术，大部分人都把目光投向了基于这两项技术基础上的基础理论的研究，特别是具体技术、算法的实现，而忽略了对数据挖掘理论与实践相结合的研究，使得这方面的书籍和论文很少，而且研究者大多集中在高校，研究成果很多是对国外成果的介绍、引进、补充、改进或翻译。许多企业、机构已经认识到数据挖掘的先进性和必要性，希望构建自己的数据挖掘系统，少数大型企业已经开始这方面的实施工作，如海尔集团和小天鹅集团等企业已经利用数据挖掘技术进行客户关系管理，并取得了较好的效益。但是，绝大部分企业还缺少构建一个完整的数据挖掘系统的理论和实践指导体系。因此，本书内容不注重数据挖掘概念、技术、算法的具体研究和介绍，而是希望能够在一个更高的层次上，为企业高层管理者及相关人员在企业数据挖掘系统的构建方面，提供一个完整的理论和实践体系，为推动数据挖掘在企业的应用做些贡献。

本书是河南省科技厅自然科学基金项目(项目编号为0311012000)的研究成果,是课题组全体成员集体研究的成果。课题组成员在进行本课题的研究中分工协作,既注重个人特长的发挥,又注重团队的合作力量;既查阅了大量的中外参考资料,又到企业进行实地调研和考察;同时我们的团队在研究工作中保持不断的学习,既对原有的理论和实践进行总结,又不断地将新学到的知识运用到实践中去,并对原有的理论进行完善和补充。在课题组成员共同努力下,完成了本书稿。在这次合作著书的过程中,课题组成员还发展了友谊,建立了一个具有良好结构的学习型团队,今后大家将继续努力合作,创造出更多的成果。

值得一提的是,在企业进行数据挖掘系统的建设过程中,涉及到数据挖掘工具的开发与选择、数据仓库建设策略等关键性问题,课题组对此进行了深入的研究,并把对这些问题的研究成果反映到了本书内容中,希望能给企业构建数据挖掘系统提供一些帮助。

本书共分7章,第一章由郭秋萍撰写;第二、五章由刘双红撰写;第三章由赵静撰写;第四章由郭秋萍、赵静撰写;第六章由余建国撰写;第七章由齐岩撰写;附录由余建国整理。本书由郭秋萍负责全书的策划和大纲制定。另外,在本书第四章的编写过程中,付永华进行了部分资料的收集工作,在此表示感谢。

由于数据挖掘是不断发展的新技术,数据挖掘在企业的应用也在不断地发展和完善中,加之时间比较仓促,以及作者知识和能力的局限,尽管课题组最大努力地精益求精,但书中难免存在错漏之处。恳请各位读者批评指正,以便纠正和提高。

愿数据挖掘成为企业制胜的工具,带领企业走向成功!

作 者

2005年3月

目　录

第一章 导 论

今天的时代是信息爆炸的时代。以计算机、因特网为代表的信息技术迅猛发展，信息的收集、过滤、处理和传输越来越快速，越来越便捷，人们积聚的信息量也越来越大，已经达到了 TB 级甚至 PB 级。但是，由于人类认知能力的有限性，海量信息在给人们带来方便的同时也带来了一大堆问题：第一是信息超载，难以消化，人们已被淹没在数据和信息的汪洋大海中；第二是有效信息难以提取，真假信息难以辨识，人们开始出现信息迷失；第三是信息安全难以保证，有意和无意的行为经常会威胁到信息的安全；第四是信息形式不一致，既有结构化的信息，也有非结构化的信息，难以统一处理。

今天的时代又是知识化的时代。在海量的信息面前，人们束手无策。普通数据库系统虽然可以高效地实现数据的录入、查询、统计和维护等功能，并可以对数据做一些简单分析处理，却无法挖掘和提供数据背后隐含的、人们真正需要的有价值的知识。例如，企业管理者需要快速敏捷地获得具体的决策知识和方案，以应对变幻莫测的市场；超市的经营者希望能从过去的销售记录中分析出顾客的消费习惯和购买规律，以便及时变换营销策略；保险公司希望知道各类客户的特征；医务人员希望能从已有的成千上万本病历中找出某种疾病病人的共同特征，从而为治愈这种疾病提供一些帮助；等等。人们开始感叹“信息丰富而知识贫乏”，人们面对海量的信息而难以做出抉择，越来越需要经过概括和总结提炼的信息——知识。

信息爆炸的时代和知识化的时代给人们提出了新的问题：如何面对信息爆炸而不被信息淹没？如何从海量的信息中提取有价值的知识？如何提高信息利用率？人们迫切需要一种能够智能地、自动地把信息和数据转换成知识的技术与工具，它既能发现数据之间的关联和规则，也能根据现有的数据预测未来的发展趋势。于是，数据挖掘技术应运而生，并显示出强大的生命力，成为未来信息技术发展的方向。它使数据处理技术进入了一个更高级的阶段，不仅能对过去的数据进行查询，并且能够找出过去数据之间的潜在联系，进行更高层次的分析，以帮助企业更好地做出正确的决策，预测未来的发展趋势。

数据挖掘以一种全新的概念改变着人类加工和利用信息的方式，使人们从单纯的信息收集、整理、存储、整理、利用、变无序信息为有序信息，向信息整合、

信息创新、信息再生产以及变信息为知识的深层次加工等转变。

数据挖掘是一门新兴的综合性学科，它融合了其他许多学科领域的技术，包括数据库技术、统计分析、机器学习、高性能计算、模式识别、神经网络、数据可视化、信息检索、图像数据库与信号处理以及空间数据分析等；能从更深层次挖掘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和知识。发现的知识可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。

第一节 数据挖掘的概念

数据挖掘(Data Mining)也叫数据开采、数据采掘等，是从大量的、不完整的、有噪声的、模糊的和随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在而又有用的信息和知识的过程。也有一些文献把数据挖掘称为知识抽取(Knowledge Extraction)、数据考古学(Data Archaeology)、数据捕捞(Data Dredging)等。通俗地讲，数据挖掘就是通过对大量业务数据进行抽取、转换、分析和模型化处理，将数据转化为有价值的知识的过程。这个定义包括以下几层含义：①数据源必须是真实的、大量的、含噪声的；②发现的是用户感兴趣的知识；③发现的知识是可接受的、可理解的、可运用的；④发现的知识是相对的、有特定前提和约束条件的、面向特定领域的，并不要求放之四海而皆准，并且最好能用自然语言表达。

这个定义主要包含三方面的内容，即数据、信息和知识、过程。数据是指有关事实的集合，它记录了事物有关方面的原始信息，是进一步挖掘知识的原材料，如员工档案数据、商场和超市销售数据、银行客户数据等。由于数据挖掘处理的数据是现实世界的客观反映，因而并不能保证所有数据都非常规范，一般需要对数据进行预处理，使之适于知识提取。信息和知识是指通过数据挖掘从当前数据中发现的信息和知识，它们源于数据，但又高于数据。这些信息和知识必须是有用的，否则数据挖掘毫无意义。同时，这些信息和知识还应该是潜在的，这是因为虽然数据挖掘可以对已有的知识进行验证，但发现新的知识或者对已有的知识进行拓展，得到更全面、更具有实际意义的知识往往更重要。过程是指数据挖掘是一个多步骤的、对大量数据进行分析处理的过程，包括数据的选择、预处理、转换、挖掘、结果的解释和评价等，是一个人机交互、螺旋上升的过程，并且往往需要经过多次反复调整，从而挖掘出质量更高、更有效的知识。如在分析影响信用风险因素时，先假设几种可能的因素，

然后通过不断反复的试验，不断增加或删除因素，最终得到对信用风险最具影响的因素。

简单地讲，数据挖掘是先有了数据才兴起的行业，是人们长期对数据库技术进行研究和开发的结果。起初各种数据是存储在计算机数据库中的，然后发展到可对数据库进行查询和访问，进而发展到对数据库的即时遍历。数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联系，对过去既有的数据集合进行统计分析。和传统的数据分析不同，数据挖掘能够找出过去数据之间的潜在联系，从而呈现给人们隐藏在数据背后的知识信息。

通过数据挖掘，有价值的知识、规则或深层次的信息就能从数据库的相关数据集合中抽取出来，并从不同角度显示，从而使大型数据库作为一个丰富可靠的资源，更好地为企业服务，使用户可以在信息的荒漠中更容易地找到知识的绿洲，解决了用户“信息丰富而知识贫乏”的尴尬。

第二节 数据挖掘的特点

根据数据挖掘的定义可以发现，与传统的信息处理方法相比，数据挖掘具有如下特点：

(1)规模性。要从数据中挖掘出规律，数据源的规模必须是海量的。数据挖掘要处理的数据集合往往是一个或多个大量或者海量的数据库。从如此巨量的数据中有效地提取有用的信息，保持信息的时效性，需要数据挖掘技术在一定运算效率的约束下进行。

(2)快速性。在市场和竞争环境瞬息万变的今天，数据变化频繁迅速，甚至有些数据很快就会过时。因此，这就要求数据挖掘能够动态地处理数据，快速地做出反应，以提供用于决策的信息和知识。

(3)动态性。数据挖掘是根据历史数据提取规则，发现潜在规则，管理和维护规则，用于指导现在的行为，并预测未来。但是，这些规则是动态变化的，当前的规则只能反映当前的数据特征。随着数据的不断产生和更新，新数据不断加入，建立规则所用的数据与当前情况的吻合程度可能降低，因此规则需要动态更新。

(4)交互性。一般用户由于检索知识的局限性，提出的即时随机查询往往不能形成精确的查询要求，需要在查询过程中依靠数据挖掘技术进行实时交互，寻找其可能感兴趣的东西，使用户的思维保持连续，以便挖掘出更深入、更有价值

的知识。

(5)适用性。数据挖掘的目标在于发现知识,而不是要求发现放之四海而皆准的真理,也不是要求去发现新的自然科学定理和数学公式。同时,由于数据挖掘发现的规则,主要是基于大样本的统计规律,因此所有发现的知识都是相对的,是有特定前提和约束条件的,是面向特定领域的,并非所有数据都必定满足此规则,即规则具有一定的适用性。但一般情况下,只要达到某一阈值,便可认为数据有此规则。

(6)知识性。数据(仓)库仅仅提供决策所需要的数据,而数据挖掘提供决策所需要的深层次知识。利用这些知识不仅可以辅助决策,更重要的是能够直接给出各种决策备选方案,甚至是给出对各个备选方案的评价。

(7)个性化。在有些情况下,用户不知道他们的数据(仓)库中,什么类型的模式是有价值的,希望并行搜索多种不同的模式。数据挖掘可以适应不同用户的需求或不同应用,挖掘多种类型和不同粒度的模式,能提供个性化服务。

(8)发掘性。对于那些实际并没有发生或很少发生的行为,或者所隐藏的有用规则和规律,并没有在数据库中直接体现出来。数据挖掘能够发掘并提取这些有用规则和规律,并提出预测。

第三节 数据挖掘的分类

数据挖掘是一门新兴的、综合性的交叉学科,它所涉及的学科领域和方法有很多,因而有多种分类方法。但是这些分类方法都从不同角度刻画了数据挖掘研究的策略和范畴,它们是互相交叉、互相补充的。

一、根据数据库类型分类

如果数据挖掘是基于关系数据库的,则称为关系数据挖掘;如果是基于面向对象数据库的,则称为面向对象数据挖掘;以此类推,还有事务数据挖掘、演绎数据挖掘、时态数据挖掘、文本数据挖掘、多媒体数据挖掘、主动数据挖掘、空间数据挖掘、异质数据挖掘、Web 数据挖掘和遗留系统挖掘等。

二、根据挖掘的知识类型分类

根据挖掘的知识类型,可分为关联规则挖掘、特征规则挖掘、分类规则挖掘、聚类规则挖掘、判别式规则挖掘、时序规则挖掘、预测性知识挖掘、偏差分析挖掘和不确定性知识挖掘等。在挖掘不同类型的知识时,数据挖掘系统将采用不同

的方法和技术。

三、根据数据挖掘方法和技术分类

根据数据挖掘采用的方法和技术，可将数据挖掘分为归纳学习类数据挖掘、仿生物技术类数据挖掘、公式发现类数据挖掘、统计分析类数据挖掘、模糊数学类数据挖掘和可视化技术类数据挖掘等。

(一)归纳学习类

归纳学习类方法包括信息论方法，如基于互信息理论的 ID3(Interactive Dicremiser versions 3)、ID4(Interactive Dicremiser versions 4)方法，基于信道容量的 IBLE 方法等和集合论方法，如覆盖正例排斥反例方法、概念树方法、粗糙集方法等。

(二)仿生物技术类

仿生物技术类方法包括神经网络法，如前馈式网络、反馈式网络、自组织式网络等和遗传算法如选择、重组、突变三个基本算子。

(三)公式发现类

公式发现类方法有物理定律发现系统 BACON、经验公式发现系统 FDD (Formula Discovery from Data)等。

(四)统计分析类

统计分析类方法包括常用统计、相关分析、回归分析(多元回归、自回归)、差异分析、聚类分析(系统聚类、动态聚类等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别)等。

(五)模糊数学类

模糊数学类方法包括模糊模式识别、模糊聚类、模糊分类和模糊关联规则等方法。

(六)可视化技术类

可视化技术类包括提取几何图元、绘制、显示和播放几个步骤，常用的绘制方法有几何法、彩色法、多媒体法和光学法等。

四、根据挖掘知识的深度分类

按挖掘知识的深度可将数据挖掘分为原始层次的数据挖掘、高层次的数据挖掘和多层次的数据挖掘等。一个灵活的数据挖掘系统应能在多个层次上发现知识。在较浅层次上，可以利用现有数据库管理系统的查询、检索及报表功能，与多维分析、统计分析方法相结合，进行联机分析处理(OLAP)，从而得出可供

决策参考的统计分析数据。在较深层次上,可以从数据库中挖掘出前所未知的、隐含的、潜在的知识。OLAP 和数据挖掘都可以从数据库中发现有用信息,只是两者挖掘出的信息的层次不同,在决策过程中,两者相辅相成,共同为决策提供支持。

五、根据对需要理解数据的需求分类

数据挖掘具有交互性,需要人机结合,根据对需要理解数据的需求,数据挖掘可以划分为两大类:数据的计算机理解和数据的人理解。具体地说,数据的计算机理解是根据数据建立一个可计算的模型,计算机可以利用这个模型求解新的问题;数据的人理解是将一本很厚的、使用数据或符号书写的书,简化并翻译为自然语言,便于人的理解,从而增加人的知识。

六、根据数据挖掘的应用领域分类

根据数据挖掘的应用领域可将数据挖掘分为通用单任务类、通用多任务类和专用领域类。

(一)通用单任务类

通用单任务类仅支持知识发现 KDD 的数据采掘步骤,并且需要大量的预处理和善后处理工作。主要采用决策树、神经网络、基于例子和规则的方法,挖掘任务大多属于分类范畴。

(二)通用多任务类

通用多任务类可执行多个领域的知识发现任务,集成了分类、可视化、聚集、概括等多种策略。主要特点是:集成多种数据挖掘算法,如:关联规则、决策树、神经网络、统计等;支持多种数据源,并有多种数据转换功能;支持多种操作平台。

(三)专用领域类

专用领域类的许多数据挖掘系统是专为特定目的开发的,是针对某个特定领域问题提供的解决方案,用于专用领域的知识发现,针对性比较强,只能用于一种应用。在设计算法时,该类系统充分考虑到了数据需求的特殊化,并作了优化。对任何领域,都可以开发特定的数据挖掘系统。正因为其针对性强,可以处理特殊数据,实现特殊目的,其发现的知识的可靠性也比较高,但它对采掘的数据库有语义要求,发现的知识也比较单一。

第四节 数据挖掘的过程

数据挖掘是一个完整且具有过程性的工作，该过程能从大型数据(仓)库中挖掘先前未知的、有效的、实用的信息，并使用这些信息做出决策或形成丰富知识。其整个过程都处于一定的数据挖掘环境中，数据挖掘环境如图 1-1 所示。

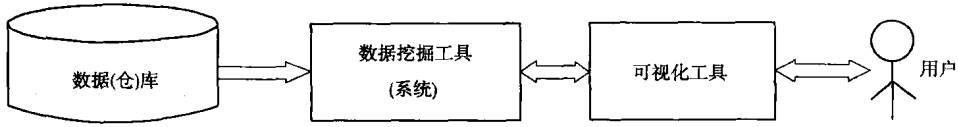


图 1-1 数据挖掘环境

数据挖掘过程一般由三个阶段组成：数据准备阶段、数据挖掘阶段、结果的解释与评价阶段。图 1-2 描述了数据挖掘的基本过程和主要步骤。尽管整个过程看起来是线性的，每个步骤是按一定顺序完成的，但在实践中，它是一个不断反复和循环的过程，整个过程中都存在着步骤间的重复和反馈。如在分析数据的时候，发现某个变量更能精化问题的定义，就可能要对业务问题进行重新定义；也有可能在结果的解释与评价以后，发现有些变量可能影响业务问题，就需要在原来的基础上，把整个数据挖掘过程重新进行一遍。

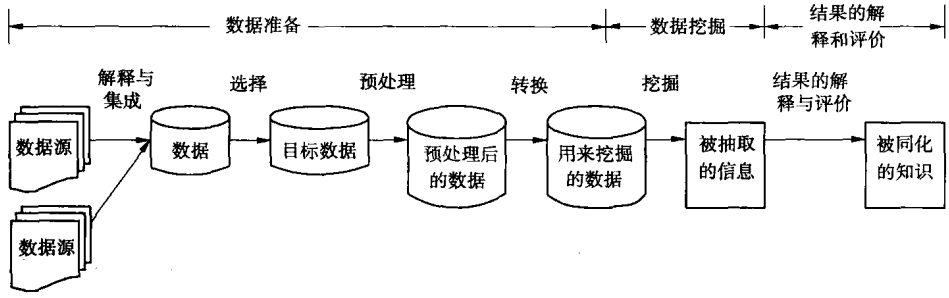


图 1-2 数据挖掘的基本过程和主要步骤

一、数据准备阶段

数据准备对于数据挖掘的成功应用至关重要，一般需要消耗整个数据挖掘过程中 50%～90%的时间和精力。事实上，数据挖掘就是对既有数据集的分析处理，如果没有数据的预处理阶段，就不可能得到一个高质量的数据集，数据挖掘也将成为一个盲目的搜索过程，可能会得出毫无意义或错误的结果。目前，对

数据挖掘的研究仍然主要集中在数据挖掘技术上，数据准备一直未得到应有的重视。因此，加强对数据准备工作的研究是十分重要的。这一阶段通常包括如下工作内容。

(一)确定业务对象

数据准备阶段的第一步就是必须了解使用数据挖掘的企业或部门的业务内容，这样才可能做到有的放矢。了解业务内容、确定业务对象是整个数据挖掘过程的基础，它驱动着整个数据挖掘过程，也是检验挖掘结果和指引分析人员完成数据解释与评价的依据。缺少了这些背景知识，就没办法明确定义要解决的问题，就不能为数据挖掘准备适用的数据，也很难正确解释得到的结果。本步骤主要包括以下工作内容。

1. 定义业务目标

主要是描述用户的业务需求，制定评价目标实现的标准。

2. 可行性评价

根据定义的业务目标，评价目前可使用的所有资源，包括人力资源、软硬件设备、资金等，对任务的可行性从技术、经济、管理等方面进行评估。

3. 定义数据挖掘目标

并非所有的业务目标数据挖掘都可以实现，并且对同一个目标，业务目标和数据挖掘目标不一定是一致的。因此，需要根据定义的业务目标，从技术角度定义数据挖掘目标，并进一步制定评价目标实现的标准。

4. 生成项目计划

根据定义的数据挖掘目标，结合企业和数据挖掘项目组的实际情况，制定数据挖掘的项目计划。

(二)数据采集和理解

数据挖掘必须基于大量数据基础之上，因此必须针对定义的业务对象进行广泛而全面的数据采集。但只有大量的数据是没有任何作用的，在进行信息采集和以后的数据挖掘过程中，企业如果不理解数据的含义，数据挖掘的结果只能是一大堆垃圾而已，对企业没任何作用，数据挖掘将变得毫无意义。所以，必须全面获取和理解数据。本步骤主要包括以下几个过程。

1. 数据采集

根据定义的业务对象，确定要挖掘的数据源，搜索所有与业务对象有关的企业内部和外部数据，获得可供挖掘分析的海量数据。

2. 数据描述

对采集到数据的格式、总体质量等进行描述。

3.数据浏览

通过浏览描述后的数据,确定这些数据包含了什么含义,并使用专用工具进行展示,用更直观的方式发现更多的与主题相关的数据和变量,从而继续在更大范围内进行数据采集,保证数据的全面性。

4.数据验证

并非所有采集到的数据都完全适用,并非所有数据的各个数据项都是正确的、完整的,并且数据间也可能存在着不一致的情况,而这些情况都会影响到数据挖掘结果的正确性和适用性。因此,必须对数据的一致性、完整性和正确性进行验证,对错误的数据类型和空值进行处理,保证数据的综合性、易用性以及数据的质量和时效性。

(三)数据集成

杂乱无章、随处乱放和存在大量重复的数据即使具有正确性、一致性和完整性,对数据挖掘而言还是远远不够的。在挖掘数据前,应该把经过采集、描述和验证之后要挖掘的数据进行集成,即将多个异质操作型数据库、文件或遗留系统运行环境中的数据提取并集成,解决语义模糊性,统一数据格式,消除冗余,清洗数据(包括对噪声数据、缺失数据及异常数据等的处理)。最好把这些数据集成到数据库或数据仓库中。但这并不是说必须使用一个数据库管理系统,而是应该根据挖掘数据量的大小、数据的复杂程度来确定数据管理的方式,有时一个简单的文件或电子表格就足够了。

(四)数据选择

数据选择是指在相关领域知识和专家知识的指导下,从集成后的逻辑数据库或数据仓库中辨别出适用数据挖掘应用的数据集合,缩小处理范围,避免盲目搜索,提高数据挖掘的质量和效率。一般来讲,数据选择包括变量选择和记录选择。

对变量的选择,理想情况下可以选择全部变量,并把它们输入到数据挖掘工具中,让它们来帮助选择哪些是最适用的变量。然而在实践中,这只能是人们的美好愿望。由于随着变量个数的增加,建立模型的时间也会随之增加,并且盲目地把所有变量都加进去,将会导致建立错误的模型。因此,必须通过数据选择来优化数据质量,如通过研究变量间的依赖关系,把那些诸如用生日来预测年龄等依赖于目标变量的变量剔除出去。

与选择变量类似,如果拥有的数据量非常巨大,并试图用所有的记录来建立模型,结果要么花费很长的时间来建立模型,要么买一台计算能力非常强大的机器。这在实践中通常都是比较棘手的问题。因此,在实际工作中,如果数据量特