数据挖掘的技术与商业定义及其研究对象

●朱建平[1] 范霄文[2] 张志强[2]
(1. 厦门大学 统计学系 福建,厦门 361005 2. 山西财经大学 经济学院,山西 太原 030006)

摘 要:本文在对数据挖掘考察的基础上,从技术角度和商业角度对数据挖掘的概念予以界定,并探讨了数据挖掘在知识发现中的地位,明确了数据挖掘的对象和对该领域的研究方向.

关键词:数据挖掘;定义;知识发现;数据库
中图分类号:G40-05 文献标识码:A 文章编号:1005-5762(2004)01-0007-04

一、引 言

如何从大型数据库中发现有用的信息、模式和知识?如何开发有效的挖掘方法? 已成为众多科技工作者共同关注的焦点。在过去几年,一个称为"数据挖掘"(Data Mining)的新领域得到了快速发展, 这是一个介于统计学、模式识别、人工智能、机器学习、数据库技术以及高性能并行计算等领域的交叉新学科,已在经济、商业、金融、天文等行业得到了成功的应用, 在国际上掀起了一股空前的研究热潮。

从总体上,国外在数据挖掘领域中的研究内容十分广泛,已经取得了明显的成果,如 Han, J. and Fu, Y. (1995)等人对于定量关联规则以及其他种类关联规则的发现研究,Mehta, M. (1996)等人针对大型数据库快速分类算法的研究,Owen, A. B. (1999)对分类与回归的管状邻域研究,Friedman, J. H. (1997)对最近邻分类方法的改进,以及所列文献中对聚类规则的研究、数据泛化、简约和特征提取研究等。目前,国内的许多科研单位和高等院校竞相开展数据挖掘的基础理论及其应用研究,例如:模糊方法在知识发现中的应用研究;对数据立方体代数的研究;对关联规则开采算法的优化和改造;非结构化数据的知识发现以及 Web 数据挖掘。为了更好地发展数据挖掘的技术与理论,将其应用于实际,我们有必要对数据挖掘的概念有一个清楚的界定,并详细明确数据挖掘的对象及其与知识发现(Knowledge Discovery in Database)的关系。

二、数据挖掘的技术定义与商业定义

什么是数据挖掘(Data Mining)? Friedman, J. H. 在技术报告 Data Mining and Statistics: What's The Connection? 中总结出了多家关于数据挖掘的定义(也有对知识发现而言的):

Fayyad 提出数据挖掘是一个确定数据中有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程。

Zekulin 的说法是数据挖掘是一个从大型数据库中提取以前未知的、可理解的、可执行的信息,并用它来进行关键的商业决策的过程。

Ferruzza 给出数据挖掘是用在知识发现过程,来辨识存在于数据中的未知关系和模式的一些方法。

Jonn 提到数据挖掘是发现数据中有益模式的过程。

Parsaye 定义数据挖掘是我们为那些未知的信息模式而研究大型数据集的一个决策支持过程。

这些表达方式虽然不同,但从各自的角度描述出了对数据挖掘的理解。这里我们主要从技术和商业的角度给出数据挖掘的定义。

1、数据挖掘的技术定义

从技术的角度看,数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。这个定义包括好几层含义:数据源必须是真实的、大量的、含噪声的;发现的是用户感兴趣的知识;发现的知识要可接受、可理解、可运用;这些知识是相对的,是有特定前提和约束条件的,在特定领域中具有实际应用

收稿日期:2003-11-05
作者简介:朱建平(1962-),男,教授,主要研究方向数理统计、数据挖掘.
范霄文(1962-),女,副教授,主要研究方向经济统计、数据分析
张志强(1964-),女,副教授,主要研究方向数理统计、保险精算.

价值。

那么，什么是知识呢？从广义上理解，数据、信息也是知识的表现形式，但是人们更把概念、规则、模式、规律和约束等看作知识。人们把数据看作是形成知识的源泉，好像从矿石中采矿或淘金一样。原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现的知识可以被用于信息管理、查询优化、决策支持和过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门交叉学科，它把人们对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决策支持。在这种需求牵引下，汇聚了不同领域的研究者，尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员，投身到数据挖掘这一新兴的研究领域，形成新的技术热点。

2、数据挖掘的商业定义

从商业应用角度看，数据挖掘是一种新的商业信息处理技术。其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性知识，即从一个数据库中自动发现相关商业模式。实际上多年来，统计学家就开始手工挖掘数据库，从数据库中寻找符合统计学规律的有意义的模式。这也是统计学类型的数据挖掘技术，是目前数据挖掘技术中最为成熟的重要原因之一。

数据挖掘是利用统计学和机器学习等技术，探求那些符合市场、客户行为的模式。如今数据挖掘已经可使挖掘技术自动化，将数据挖掘与商业数据仓库相结合，以适当的形式将挖掘结果展示给企业经营管理人员。对于数据挖掘的应用不仅依靠良好的算法建立模型，而且更重要的是解决如何将数据挖掘技术集成到信息技术应用环境中。同时，还要有数据挖掘分析人员参与，因为数据挖掘技术不具备人所特有的经验和直观，不能区分哪些挖掘出的模式在现实中是有意义的，哪些是无意义的。因此，数据挖掘分析人员的参与是必不可少的。

简而言之，数据挖掘其实是一类深层次的数据分析方法。数据分析本身已经有很多年的历史，只不过在过去数据收集和分析的目的更多是用于科学研究，另外，由于当时计算能力的限制，对大数据量进行分析的复杂数据分析方法受到很大限制。现在，由于各行业业务自动化的实现，商业领域产生了大量的业务数据，这些数据不再是为了分析的目的而搜集的，而是由于业务处理操作而获取和积累的。分析这些数据也不再是单纯为了研究的需要，更主要是为商业决策提供真正有价值的信息，进而获得利润。但所有企业面临的一个共同问题是：企业数据量非常大，而其中真正有价值的信息却很少，因此从大量的数据中经过深层分析，获得有利于商业运作、提高竞争力的信息，就像从矿石中淘金一样，数据挖掘也因此而得名。

因此，数据挖掘可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，且进一步将其模型化的数据处理方法。

三、数据挖掘与知识发现

1、知识发现过程

数据挖掘和知识发现的关系需要澄清。根据Fayyad 对KDD 的定义“The nontrivial Process of identifying valid，novel，potentially useful，and ultimately understandable pattern in data”，知识发现过程可以粗略的理解为三部曲：数据准备(data preparation)、数据挖掘以及结果的解释评估(interpretation and evaluation)（参见图 1）。

(1)数据准备

数据准备又可分为三个子步骤：数据选取、数据预处理和数据变换。数据选取的目的是确定发现任务的操作对象，即目标数据。数据预处理一般可能包括消除噪声、推导计算缺失值数据、消除重复记录、完成数据类型转换等。数据变换的主要目的是消减数据维数或降维，即从初始特征中找出真正有用的特征以减少数据挖掘时要考虑的特征或变量数。

(2)数据挖掘阶段

数据挖掘阶段主要是确定挖掘的任务，如数据总结、分类、聚类、关联规则发现或序列模式发现等。确定了挖掘任务后，就要决定使用什么样的挖掘算法。选择实现算法有两个需要考虑的因素：一是不同的数据有不同的特点，需要用与之相应的算法来挖掘；二是根据用户或实际运行系统的要求，有的用户可能希望获取描述型的、容易理解的知识，而有的用户或系统的目的是获取预测准确度尽可能高的预测型知识。有了上述的准备工作，就可以实施数据挖掘操作了。

(3)结果解释和评价

数据挖掘阶段发现出的模式，经过用户和机器的评价，可能存在冗余或无关的模式，这时需要将其剔除。如果有的模式不满足用户要求，需要将整个发现过程退回到发现阶段之前。最终结果是要面向用户，有时要对发现的模式进行可视化，或者将结果转化为用户易懂的另一种形式。

2、数据挖掘的地位

从上面的介绍我们可以看出，KDD 是一种知识发现的一连串程序，数据挖掘只是 KDD 的一个重要程序。数据挖掘主要是利用某些特定的知识发现算法，在一定的运算效率的限制内，从数据中发现有关的知识，即隐藏的模式，数据挖掘是 KDD 中最重要的一步，在 KDD 的全过程中起到了至关重要的作用（张尧庭，谢邦昌，朱世武（2001），12－23）。因此，人们往往不加区别地使用数据挖掘和 KDD。

在这里，我们也应该清楚地认识到，数据挖掘的质量取

决于两方面的影响：一是所采用数据挖掘技术的有效性；二是用于挖掘的数据的质量数量（数据量的大小）

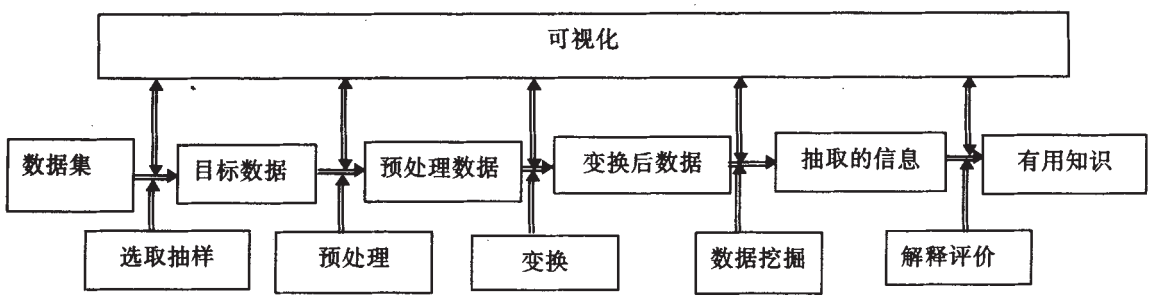

图 1　　知识发现过程示意图

（史忠植（2001），13－16）。如果选择了错误的数据或不恰当的属性，或对数据进行了不恰当的转换，则挖掘的结果就会影响挖掘的质量。

四、数据挖掘的对象

数据挖掘的范围非常广泛，可以是社会科学、经济学、商业数据、科学处理产生的数据和卫星观测得到的数据. 它们的数据结构也各不相同，可以是层次的、网状的、关系的和面向对象的数据。

1、关系数据库

关系数据库是表的集合，每个表都赋予一个唯一的名字。每个表包含一组属性（列或字段），并通常存放大量元组（记录或行）。关系中的每个元组代表一个被唯一的关键字标识的对象，并被一组属性值描述。

关系数据可以通过数据库查询访问。数据库查询使用如 SQL 这样的关系查询语言，或借助于图形用户界面书写。在后一种情形下，用户可以使用菜单指定包含在查询中的属性和属性上的限制。一个给定的查询被转换成一系列关系操作，如连接、选择和投影，并被优化，以便有效地处理。查询可以检索数据的一个指定的子集。

当数据挖掘用于关系数据库时，可以进一步搜索趋势或数据模式。数据挖掘系统也可以检测偏差，如在商业营运中，与以前的年份相比，哪种商品的销售出人预料。这种偏差可以进一步考察，例如包装是否有变化，或价格是否大幅度提高。

关系数据库是数据挖掘最流行的、最丰富的数据源，因此它是我们数据挖掘研究的主要数据形式。

2、事务数据库

一般地说，事务数据库由一个文件组成，其中每个记录代表一个事务。通常一个事务包含一个唯一的事务标识号和一个组成事务的项的列表（如，在商店购买的商品）（史忠植著，（2002）13－16）。事务数据库可能有一些与之相关联的附加表，包含关于 销售的其他信息，如事务的日期、顾客的 ID 号、销售者的 ID 号、销售分店等等。

如果我们想更深地挖掘数据，在商业营运中，问“哪些商品适合一起销售？”这种‘购物篮数据分析’使我们能够将商品捆绑成组，作为一种扩大销售的策略。例如，给定打印机与计算机经常一起销售的知识，你可以向购买选定计算机的顾客提供对一种很贵的打印机打折销售，希望销售更多较贵的打印机。常规的数据检索系统不能回答上面这种查询。然而，通过识别频繁地一起销售的商品，事务数据的数据挖掘系统可以做到。

3、数据仓库

在数据仓库的发展过程中，许多人对此做出了贡献。其中，Devilin 和 Murphy 在 1998 年发表了一篇关于数据仓库论述的最早文章。而 Inmon，W. H. 在 1993 年所写的论著 Building the Data Warehouse 则首先系统性地阐述了关于数据仓库的思想、理论（Inmon，W. H.（1996）），为数据仓库的发展奠定了历史基石。在 Building the Data Warehouse 中，他将数据仓库定义为“一个面向主题的、集成的、随时间变化的非易失性数据的集合，用于支持管理层的决策过程”。关于数据仓库的定义还有：“数据仓库是一种体系结构，一种独立存在的不影响其他已经运行的业务系统的语义一致的数据仓储，可以满足不同的数据存取、文档报告的需要”。数据仓库‘是一个不断发展的过程，将多个异质的原始数据融合在一起，用于支持结构化的在线查询，分析报告和决策支持”。

从 Inmon，W. H. 关于数据仓库的定义中，可以发现数据仓库具有这样一些重要的特性：面向主题性、数据集成性、数据的时变性、数据的非易失性、数据的集合性和支持决策作用。

通常，数据仓库用多维数据库结构建模。其中，每一维对应于模式中的一个或一组属性，每个单元存放某个聚集度量值。数据仓库的实际物理结构可以是关系数据存储或多维数据立方体（data cube）。它提供数据的多维视图，并允许预计算和快速访问汇总的数据。

数据仓库工具对于支持数据分析是有帮助的，但是仍

需要更多的数据挖掘工具，以便进行更深入的自动分析。

4、高级数据库系统

关系数据库系统广泛地用于商务应用。随着数据库技术的发展，各种高级数据库系统已经出现并在开发中，以适应新的数据库应用需要。新的数据库应用包括处理空间数据（如地图）、工程设计数据（如建筑设计、系统部件、集成电路）、超文本和多媒体数据（包括文本、影象、图象和声音数据）、时间相关的数据（如历史数据或股票交易数据）和 Web（通过 Internet 可以使巨大的、广泛分布的信息存储）。这些应用需要有效的数据结构和可伸缩的方法，处理复杂的对象结构、变长记录、半结构化或无结构的数据以及文本和多媒体数据，并具有复杂结构和动态变化的数据库模式。

为响应这些需求，开发了高级数据库系统和面向特殊应用的数据库系统。这些包括面向对象和对象－关系数据库系统（Han, J. W. and Kamber, M. (2001) 12－16）、空间数据库系统（史忠植, (2002) 13－16）、时间和时间序列数据库系统、文本（Hahn, U. et al. (1997)）和多媒体数据库系统、异种和遗产数据库系统、基于 Web 的全球信息系统（Bern, S. (1998)）。虽然这样的数据库或信息存储需要复杂的机制，以便有效地存储、检索和更新大量复杂的数据，但它们也为数据挖掘提供了肥沃的土壤，提出了挑战性地研究和实现问题。

参考文献：

[1] 张尧庭, 谢邦昌, 朱世武．数据采掘入门及应用——从统计技术看数据采掘[J]．北京：中国统计出版社, (2001.6)．

[2] 史忠植．知识发现[M]．北京：清华大学出版社, (2002.1)．

[3] 朱建平，Data Mining 中的统计方法及其应用[J]，博士论 文(2003.3)．

注：

国家社会科学基金资助(03BTJ014)；

国家统计局统计科学研究重点项目基金资助(LX2002－2)；

(上接第 6 页)

基础课程、专业主干课程应大体一致，在专业方向课程上应有所差别，通过专业方向课的差别来突出专业特色。

专业培养方案是实现专业培养目标的载体，是人才培养规格的具体体现。要在研究人才市场的最新变化的基础上，及时调整和修订专业教学计划，突出专业特色。在目前统计学本科教育中要加强数学、概率统计、英语、计算机以及相关实质性学科的教育，加强统计学与相关实质性科学的交叉与融合，加强对学生的素质教育和能力培养。同时要注意统计学专业人才的后续培养，如精算师、特许金融分析师、数据分析师和统计师的培训与考试等。

2. 全面深化课程改革与建设

课程建设是专业建设的基本单元，要使统计学专业所有课程在新的起点上达到合格或优秀的标准：有合理、可行的建设规划、有新的教学大纲和较适用的教材；有与教学大纲、教材相适应的新型的教学方法与手段；有科学的课程考核和课程质量评估办法 有结构合理的师资队伍。

3. 加强师资队伍建设

师资队伍是专业建设和学科发展、建设专业特色的前提条件。没有一支数量充足、素质较高的师资队伍，一个专业的发展是不可想象的。要加大优秀师资引进的力度，充实师资数量，改善师资队伍结构；抓好教师的培养和提高，继续搞好青年教师的传、帮、带，提高青年教师的教学水平，早过教学关；鼓励教师外出学习交流和进一步深造提高；着力培养学术带头人和学科带头人，培养一批在国内知名的专家、教授。

4. 加大实验室建设力度

实验室建设是专业建设的重要组成部分，也是办出专业特色的重要物质条件。要结合专业特色的建设加大、加快实验室建设步伐，使实验室的也具有明显的特色。

5. 抓好教材建设

教材建设是教学内容和课程体系改革的重要内容之一，是建设特色专业的必要条件，也是教学成果的集中体现。要认真规划，做好教材的选、编工作，要组织编写一批高质量的、特色鲜明的专业教材。同时，要积极抓好电子课件、实习实验教材、学生参考教材等配套教材的编写。

6. 加大科研投入，创造科研优势，以科研促教学，以科研创特色

特色专业的创立是以科研为先导的。为了达到科研促教、科研促学、科研促改，提高办学水平的目的，要切实鼓励教师从事科研活动，注重培养青年教师参与搞科研，形成了较为合理的科研梯队；加大科研成果奖励力度，鼓励教师多出、快出科研成果。

参考文献：

[1] 贺铿．关于统计学的性质与发展问题[J]．统计教育，2001, 6.

[2] 胡学锋．立足财经类院校，培养管理型统计学专业人才[J]．统计教育，2003, 6.