

# 浅谈数据仓库与数据挖掘的本科教学

## A Brief Discussion on Database and Data Mining for Undergraduate Education

胡建军

HU Jian-jun

(广东商学院信息学院, 广东广州 510320)

(School of Information, Guangdong University of Business Studies, Guangzhou, Guangdong, 510320, China)

**摘要:**根据在大学计算机类和经济管理类本科生中教授数据仓库与数据挖掘课程中存在的问题及特点,提出在教学中要根据计算机类和经济管理类本科生的不同特点展开教学,对计算机类的学生应侧重算法和设计,而对管理类的学生则应侧重具体应用。

**关键词:**数据仓库 数据挖掘 本科教学

**中图分类号:**TP311 **文献标识码:**A **文章编号:**1002-7378(2007)03-0209-02

**Abstract:** Based on the problems and the features in database and data mining Courses taught to undergraduates majoring in computer science and economics & administration, the article suggests that teaching should be carried out in accordance with the different characteristics of students majoring in computer science and economics & administration. Emphasis should be placed on algorithms and design for students of computer science, while for students of economics & administration, stress should be laid on the specific application of algorithms.

**Key words:** database, data mining, undergraduate education

数据仓库和数据挖掘是商业智能系统的核心内容,是目前计算机领域中的热点研究方向之一。以数据仓库和数据挖掘为核心的商业智能的应用在我国才刚刚起步,具有广阔的发展前景。数据仓库和数据挖掘技术之间既有区别又有联系,前者是后者的基础,后者是前者的发展和运用,它们共同构成了现代商业智能的基础。随着商业竞争的加剧,社会对商业智能的需求越来越高。为了适应科技发展的需要,很多高校把数据仓库和数据挖掘技术作为选修课在计算机类和经济管理类的本科高年级中开展,由于该课程原属于研究生阶段的专业课程,在本科生中讲授尚无经验借鉴。本文针对本科教育中计算机类和经济管理类学生的不同特点,对数据仓库和数据挖掘课程的授课方法进行初步探讨。

### 1 数据仓库和数据挖掘课程教学中存在的问题及特点

数据仓库和数据挖掘课程涉及知识广泛,不仅用到很多传统的数学和计算机技术,如概率统计、矩阵、信息论、数据库、离散数学、数据结构、决策支持、人工智能、专家系统等,而且还要涉及到最新的计算技术,如机器学习、遗传算法、神经网络、粗糙集、模糊数学等<sup>[1]</sup>。因此要求学生必须有比较扎实的数据基础和较深的计算机专业知识。然而对于本科生来说,其中很多专业知识还没有学到,而且由于课时所限,又不能对数据仓库和数据挖掘课程中所涉及的知识一一讲解。因此,对于课堂授课内容就需要教师根据学生特点精心准备。

由于数据仓库和数据挖掘课程原本是研究生阶段才学的,因此市面上的课本和参考资料也基本上是针对研究生和专业技术人员来编写的。其中所涉及的内容大多比较深,很多都超出了本科生的接受范围。而且数据仓库和数据挖掘原本就是两门不同的课程,多数有关这两方面的课本和参考书籍也都

收稿日期:2007-06-25

作者简介:胡建军(1970-),男,讲师,主要从事数据挖掘、商业智能方面的研究与教学工作。

是分开来编写的。因此,基本上找不出一本适合本科生教学的数据仓库和数据挖掘参考书作教材。

数据挖掘中所使用的技术是基于计算机技术的,而数据挖掘的使用者却多是管理人员。因此,学习这门课程的学生既要有计算机类专业的知识,又要有管理类专业的知识,而不同专业方向的学生的课程基础、学习目标和教学目的是不相同的。这就造成选修这门课程的学生基础水平各不相同,增加了教师的教学难度。

综上所述,数据仓库和数据挖掘课程的教学具有以下特点:(1)涉及内容广泛。既要有数学和计算机方面的知识,又要有商业管理方面的基础知识。(2)内容新,难度大。其中用到很多最新的计算技术都超出了本科生的学习范围。(3)选修本课程的学生水平基础水平各不相同。

## 2 教学实践与对策

### 2.1 教学内容选择

由于计算机类和经济管理类学生的基础和教学目标不一样,因此所授内容应区别对待。

#### 2.1.1 计算机类本科生的教学内容

对于计算机类的学生,其学习目的是掌握数据仓库的设计方法,掌握基本的数据挖掘算法,能够根据不同的挖掘目的设计出简单的数据挖掘软件。因此,在教学中应侧重于数据仓库设计方面的内容和挖掘算法。然而,由于数据挖掘算法很多,根据挖掘目的不同通常可把挖掘算法分为描述性信息挖掘、关联规则挖掘、聚类分析、分类分析、预测分析、web挖掘等<sup>[2]</sup>,每一类又包含很多使用不同技术和技巧的算法,其中很多技术都超出了本科生的学习范围,因此,需要对讲授的算法精心挑选,突出重点,不能面面俱到,眉毛胡子一把抓。

讲授数据仓库部分时,应在简要介绍建立数据仓库的原因、目的、发展史,以及与数据库系统的相同与不同之处的基础上,详细讲解建立数据仓库过程中所使用的方法、技巧等技术细节。讲授数据挖掘部分时,为了便于学生接受,每一类挖掘算法可以选择一个相对简单易懂的典型算法详细讲解,对于使用复杂技术的算法可以做些简单介绍,并推荐一些相关的参考文章或书籍,便于学生扩大知识面。例如,关联规则算法可以详细讲解 Apriori 算法<sup>[3]</sup>,对于 Apriori 算法的改进算法和学生难以理解的 FP-Tree 算法及其相关的改进算法则可以简单介绍一下它们的思路;聚类算法可以详细讲解最短距离法,

其它使用复杂技术的聚类算法则可以简单介绍<sup>[4]</sup>;分类算法可以详细讲解分类树算法,对于其中所涉及的信息增益概念和公式可稍作解释,其它分类算法可仅简单介绍其算法思路,对于复杂类型的挖掘技术,如多媒体数据挖掘、序列分析、Web 挖掘等,由于涉及到一些复杂的技术,学生难以理解,因此,也应以简单介绍为主。

#### 2.1.2 经济管理类本科生的教学内容

对于经济管理类的学生,其学习目标是通过本课程的学习,能够对数据仓库有初步了解,能够使用数据挖掘工具软件完成特定的挖掘任务。由于经济管理类的学生缺少必要的计算机基础知识,对于计算机算法和专业技术很难理解,而且也不要要求他们掌握挖掘软件的实现。因此,在教学内容上应与计算机专业的学生有所区别,应尽量避免讲授具体的算法实现和技术细节,应把讲授重点放在挖掘的功能、原理、应用场合和数据的预处理上。这样,经济管理类的学生学完本讲和后才能更好地把所学知识应用于将来的工作实践中。

由于经济管理类的学生多数都没有学习过数据库技术,因此对于数据仓库中所涉及的结构设计可能难以理解。所以在讲授这一部分内容时,应把讲解重点放在数据仓库系统与普通数据库系统的区别和联系,建立数据仓库的原因、目的和成功的组织模式,以及建立数据仓库给企业带来的益处上。而把数据仓库的详细设计过程忽略不讲或简单介绍。在讲解数据挖掘部分时,也要针对经济管理类学生的特点和学习目的,把重点放在具体的应用上。如讲解关联规则时,可以先讲一下“啤酒与尿布”的故事,把关联规则的神奇之处以实例的形式展现给学生,从而提高学生的学习兴趣。然后再讲解关联规则产生的原因、发展历史、应用的场合,并以实例的方式讲解产生的关联规则在业务中的具体应用。最后再讲解一下产生关联规则的形式、规则的含义、不足之处、注意事项,以及使用关联规则前需要对数据进行的预处理。讲解分类算法、聚类算法等其它挖掘算法时,也要把重点放在如何在具体业务中应用这些算法上。

### 2.2 实验项目设置

实验是检验和巩固所学知识、增强学生动手实践能力不可或缺的手段之一,因此数据仓库和数据挖掘课程除了加强课堂讲授外还要加强学生的实验环节。由于计算机类学生和经济管理类学生的教学

(下转第214页)

## 参考文献:

- [1] 李克东. 信息技术与课程整合的目标与方法[M]//李克东, 谢幼如, 柯清超, 等. 信息技术与学科教学整合. 北京: 万方数据电子出版社, 2001.
- [2] 曾玉华, 刘朝阳, 刘朝辉. 论信息技术与数学课程的整

合[J]. 江西教育科研, 2005(8): 40-41.

- [3] 张斌, 李晓平. 关于信息技术与数学课程整合的思考[J]. 信息技术教育, 2003(4): 35-36.

(责任编辑: 韦廷宗)

(上接第210页)

重点不一样,因而在实验设计方面要针对不同的学生,设计不同的实验内容.基本指导思想是,针对计算机类的学生多设计一些开发、设计方面的实验;而针对经济管理类的学生则多设计一些操作应用型的实验.

### 2.2.1 计算机类本科生的实验项目设置

对于计算机类的学生,数据仓库实验可以如下设计:由教师给定一个单位(如超市)不同业务部门的数据库系统<sup>[5]</sup>,让学生根据数据仓库的设计原则和单位的业务特点,设计出整个单位数据仓库系统中的各种数据库表、原数据表和数据集市,使数据可以按不同的粒度和分区合理地存储在数据仓库中.然后编程实现对业务数据库中的数据进行必要的预处理,并转存到数据仓库系统中.由于实验学时所限,可能没有时间实现所有的功能,可以让学生自选其中的一部分来做.数据挖掘部分的实验可以是实现和改进课堂讲授的具体算法,如编程实现 Apriori 算法、最短距离算法和分类树算法,并用所设计的程序对教师给定的数据集进行相应的挖掘.通过以上实验,不仅可以提高计算机类学生的动手编程能力,而且可以加深对所学内容的理解,掌握数据仓库系统和数据挖掘软件系统的设计技巧.

### 2.2.2 经济管理类本科生的实验项目设置

对于经济管理类的学生,数据仓库实验可以做如下设计:使用设计好的 ETL 软件把业务数据库中的数据转存到设计好的数据仓库系统中.通过这个练习,可以使经济管理类的学生更加深刻地体会到数据库系统与数据仓库系统的异同,以及数据从数据库系统转存到数据仓库系统的过程.数据挖掘部分的实验可以设计如下:使用数据挖掘软件系统(如 DBMiner)对教师给定的数据集进行单一的挖掘实

验,让学生掌握每一种挖掘功能的使用方法.最后做一个综合实验,给定一个真实的数据集,让学生对其中的数据进行必要的预处理,然后使用数据挖掘软件进行各种挖掘实验,最后对挖掘结果进行合理的解释并说明如何在实际中运用挖掘结果.通过以上练习,可以让学生掌握数据挖掘软件的使用方法,为其将来在实际工作中使用数据挖掘工具打好基础.

## 3 结束语

在本科生中讲授数据仓库和数据挖掘尚处于探讨阶段,其教学方法和教学内容还很不完善,以上仅是本人在教学中的一些心得体会.总之,在教学中要根据本科生的特点和各类不同学生的特点展开教学,比如对计算机类的学生侧重讲解算法和设计,而对经济管理类的学生则侧重讲解具体应用.

### 参考文献:

- [1] 曾华军,张银奎. 机器学习机[M]. 北京:机械工业出版社, 2005:10-13.
- [2] 韩家炜. 数据挖掘概念与技术[M]. 北京:机械工业出版社, 2001:14-20.
- [3] AGRAWAL R, WRIKANT R. Fast Algorithms for mining association rules in large databases[C]. Proc. 20th int'l Conf Very large databases, 1994:478-499.
- [4] 胡建军,唐常杰,李川,等. 基于最近邻优先的高效聚类算法[J]. 四川大学学报:工程科学版, 2004, 36(6): 93-99.
- [5] 徐洁磐. 数据仓库与决策支持系统[M]. 北京:科学出版社, 2005:28-32.

(责任编辑: 韦廷宗)