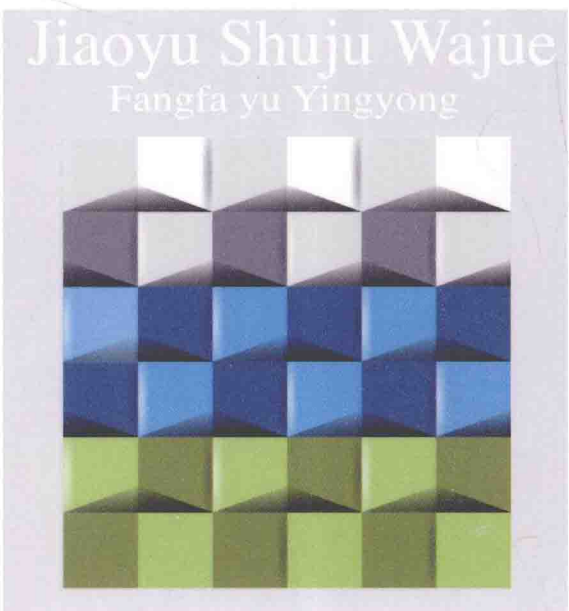
教育部科学技术研究重点项目成果

教育数据挖掘

方法与应用

葛道凯　张少刚　魏顺平　著



教育科学出版社
ESPH Educational Science Publishing House

教育部科学技术研究重点项目成果

教育数据挖掘

方法与应用

葛道凯　张少刚　魏顺平 著

教育科学出版社
·北京·

出 版 人　所广一
责任编辑　殷　欢
版式设计　杨玲玲
责任校对　贾静芳
责任印制　曲凤玲

图书在版编目（CIP）数据

教育数据挖掘：方法与应用/葛道凯，张少刚，魏顺平著. —北京：教育科学出版社，2012.12
ISBN 978-7-5041-7160-3

Ⅰ.①教… Ⅱ.①葛… ②张… ③魏… Ⅲ.①数据采集-计算机应用-教育-研究 Ⅳ.①G434

中国版本图书馆 CIP 数据核字（2012）第 282206 号

教育数据挖掘：方法与应用
JIAOYU SHUJU WAJUE：FANGFA YU YINGYONG

出版发行	教育科学出版社		
社　　址	北京·朝阳区安慧北里安园甲 9 号	市场部电话	010-64989009
邮　　编	100101	编辑部电话	010-64981269
传　　真	010-64891796	网　　址	http://www.esph.com.cn
经　　销	各地新华书店		
制　　作	国民灰色图文中心		
印　　刷	北京人卫印刷厂		
开　　本	169 毫米×239 毫米　16 开	版　　次	2012 年 12 月第 1 版
印　　张	12.75	印　　次	2012 年 12 月第 1 次印刷
字　　数	182 千	定　　价	28.80 元

序　言

21 世纪以来，随着信息化进程在教育领域的推进，特别是数字化校园建设和网络高等教育的大力推进，教育领域已经部署了众多的软件系统，在这些软件系统中存储着海量的教育数据。如何利用这些教育数据，使这些数据转变为信息、知识，并为教育决策、教学优化服务，而不至于“淹没在数据的海洋中，却又忍受着信息的饥渴”，已成为教育工作者特别是教育决策者所关注的内容。此时，致力于从大量数据中提取或“挖掘”知识的数据挖掘将有助于发挥教育数据的价值。数据以及数据挖掘可以作为审慎决策的依据。

本书以“教育数据挖掘”为主题，围绕教育数据挖掘的方法和应用两大方面展开论述，根据实际任务情境构建了若干数据挖掘模式，并结合研究和实践中的实际问题开展了大量实证研究，检验了教育数据挖掘的重要价值。

教育数据挖掘是一个将来自各种教育系统的原始数据转换为有用信息的过程，这些有用信息可为教师、学生及其家长、教育研究人员以及教育软件系统开发人员所利用。教学、管理、科研是教育机构的基本活动，根据数据挖掘在这三个业务领域的具体应用，可以将教育数据挖掘进一步细分为 E-Learning 数据挖掘、E-Management 数据挖掘和 E-Research 数据挖掘等。

针对不同的人群，教育数据挖掘有其特定的价值。对于学习者而言，教育数据挖掘的作用体现在：向学习者推荐有助于改进他们学习的学习活动、学习资源和学习任务；向学习者推荐好的学习经验；等等。这些建议可以通过分析这些学习者完成的行为以及与之相似的学习者完成的行为来取得。对于教育工作者而言，教育数据挖掘的作用体现在：向他们提供更多更客观的反馈信息，使他们能够更好地调整和优化教育决策、改进教育过程、完善课程开发；根据学习者的学习状态来组织教学内容、重构教学计划；等等。

教育数据挖掘方法可分为五类：统计分析与可视化；聚类（聚类、离群点分析）；预测（决策树、回归分析、时序分析）；关系挖掘（关联规则挖掘、序列模式挖掘、社会网络分析）；文本挖掘。其中的关系挖掘方法、预测挖掘方法应用十分广泛。

本书通过 E-Learning 数据挖掘、E-Management 数据挖掘和 E-Research 数据挖掘三大领域数据挖掘的七项实证研究，对于远程开放教育领域可获得的数据种类、可采用的数据挖掘方法和工具以及可挖掘得到的知识模式进行了较为完整的介绍，并得出以下基本结论。

一是恰当运用数据挖掘技术能够为优化教育规划和管理、提高教育教学质量、改进教育软件设计与开发提供有益的帮助。

二是对于多数教育机构来说，在教育教学过程中适时应用数据挖掘技术不仅是必要的，也是可能的。

三是研究人员借助数据挖掘方法，基于各种专业数据库，在一定程度上可以更全面、快速、准确地了解某一研究领域的现状，并预测未来的发展方向。

四是注重教育教学过程、管理过程以及研究过程中相关信息的采集和存储是一项有价值的活动。

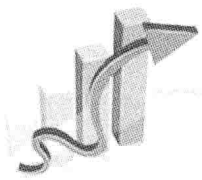
本书的学术价值体现在两个方面：一是拓展数据挖掘的应用范围，创新优化教学和教育决策的系统集成方法，为远程教育的改革发展和质量提高提供新的分析方法与支持手段；二是拓展远程教育研究的新领域，提升研究工作的针对性和有效性。

本书在撰写过程中，得到了中央广播电视大学多个部门的大力支持和帮助。华中师范大学傅德荣教授、北京师范大学何克抗教授等的思想方法使作者受益匪浅，教育部原副部长赵沁平教授、中国医学科学院院长刘德培院士、解放军总医院尹岭教授以及教育部发展规划司谢焕忠司长、科学技术司娄晶副司长等为本书提出了许多宝贵的意见和建议，教育科学出版社及责任编辑殷欢同志为本书的顺利出版付出了辛勤的努力，在此表示衷心的感谢。

受作者水平所限，书中不足和疏漏在所难免，敬请读者不吝赐教。

葛道凯
2012 年 9 月

目　录



第一章
教育信息化：历程与成效

21 世纪人类社会全面进入信息时代，信息技术已深度融入国民经济和社会生活的各个方面，人类的生产方式、学习方式、生活方式正在发生着深刻的变化，人类社会的进步和经济发展对信息资源、信息技术和信息产业的依赖程度越来越大。信息化程度的高低已成为衡量一个国家现代化水平的重要标志。信息作为当今世界推动社会生产力发展的新动力，正日益受到人们的重视。信息同能源、材料一起被看作是人类生产和生活必不可少的三大资源。在当今世界上，谁先掌握信息，谁掌握的信息多，谁就能立于不败之地。①

当今社会处于信息爆炸的时代，虽然我们渴求信息，但往往是“淹没在数据的海洋中，却又忍受着信息的饥渴”。可见，我们不仅需要海量数据，更需要将海量数据转换为有用的信息、知识并进而产生效益的方法和技术。据美国物理学家组织网、《科学》杂志等网站报道，美国科学家的研究表明，目前人类能存储至少 295 安比特（exabytes，1 安比特 = 10^{18} 比特）的信息量，这是全世界沙粒数量的 315 倍，但还不到存储于所有人类 DNA 分子中信息量的百分之一。如果把这 295 安比特信息存储在只读光盘（CD - ROMs）上，这些光盘将会从地球堆到月球。② 由此可见，我们所拥有的数据不可谓不多。

21 世纪以来，随着信息化进程在教育领域的推进，特别是数字化校园建

① 韩晓华．迎接信息化社会的挑战［J］．北方经济，1999（11）：39 - 40.

② 常丽君．美公布人类掌控信息总体能力［N］．科技日报，2011 - 02 - 14（2）.

设和网络高等教育的大力推进，教育领域已经部署了众多的软件系统，在这些软件系统中存储着海量的教育数据。如何利用这些教育数据，使这些数据转变为信息、知识，并为教育决策、教学优化服务，已成为教育工作者特别是教育决策者所关注的内容。

教育信息化充分发挥现代信息技术优势，注重信息技术与教育的全面深度融合，在促进教育公平和实现优质教育资源广泛共享、提高教育质量和建设学习型社会、推动教育理念变革和培养具有国际竞争力的创新人才等方面具有独特的重要作用，是实现我国教育现代化宏伟目标不可或缺的动力与支撑。① 党中央、国务院一直十分重视教育信息化工作。自 1994 年启动中国教育和科研计算机网（CERNET，简称中国教育科研网）建设以来，经过“211 工程”“985 工程”“面向 21 世纪教育振兴行动计划”“农村中小学现代远程教育工程”“西部大学校园计算机网络建设工程”和“校校通”等一系列重大工程建设，教育信息化得到了快速发展，教育信息化的重要性得到了全社会越来越广泛的认可。

下面简要回顾我国教育信息化发展的历程和成效。

一、教育信息化的发展历程

总的来说，我国教育信息化的发展轨迹可以这样描述：“九五”期间是多媒体教学发展期和网络教育启蒙期，“十五”期间是多媒体应用期和网络建设发展期，“十一五”期间则是网络持续建设和应用普及期②。经过十多年的建设，我国教育信息化在基础设施建设、重大应用、资源建设、标准化建设、法律法规建设和相应的管理等方面均取得了可喜的进展。

参考《我国教育信息化建设与应用专题研究报告》③ 和《中国教育信息

① 教育部．教育信息化十年发展规划（2011—2020 年）．

② 祝智庭．中国教育信息化十年［J］．中国电化教育，2011（1）：20－25.

③ “教育信息化建设与应用研究”课题组．我国教育信息化建设与应用专题研究报告［M］．北京：高等教育出版社，2010.

化十年》①，从基础教育、高等教育、职业教育三个层面梳理，教育信息化的发展历程可以概括如下。

1. 基础教育信息化

2000 年 10 月召开的全国中小学信息技术教育工作会议，是我国基础教育信息化的一个里程碑。时任教育部部长陈至立在会议上提出了“全面启动中小学‘校校通’计划，为中小学普及信息技术教育、推动教育信息化建设奠定基础”的目标，计划用 5～10 年时间，使全国 90% 左右的独立建制的中小学校能够与互联网或中国教育卫星宽带网联通，使中小学师生都能共享网上教育资源，提高所有中小学的教育教学质量，使全体教师能普遍接受旨在提高实施素质教育水平和能力的继续教育。与此同时，全国中小学普及信息技术教育计划也开始全面实施，分三步在全国中小学中普及信息技术课程：2001 年前全国所有高中开设信息技术课程，2005 年前全国所有初中开设信息技术课程，2010 年前全国所有小学开设信息技术课程。

到 2008 年，全国基础教育学校平均拥有计算机 37.2 台，联网率达 53.4%，生机比为 19∶1，师机比为 3∶1。基础教育中，71.8% 的学校拥有数字教学资源，其中 26.48% 的学校还建立了统一的教学资源管理平台；87% 的学校反映信息技术的应用对于改进教学效果有较大的帮助，其中 25% 的学校认为有非常大的帮助。已有 67.5% 的学校开设了信息技术课程，学习信息技术课程的学生比例达到了 69%，每年有 1 亿多名中小学生接受信息技术教育；信息技术专任教师数占所有专任教师的 4.5%，平均每所学校有 1.5 名信息技术教师。基础教育学校也开始引入电子校务，有 13.9% 的学校建立了主页，10% 的学校建有集成各种信息资源的内部信息门户。此外，信息化建设的投入占到学校建设投入总额的 28%，48% 的学校制定了信息化发展规划。

① 祝智庭．中国教育信息化十年［J］．中国电化教育，2011（1）：20－25.

2. 高等教育信息化

在基础设施建设方面，1994 年启动中国教育科研网建设，2003 年开始试验第二代互联网（CERNET 2）。2001 年教育部发起数字化校园建设项目，到 2008 年高等学校已经全部建成了校园网络，校园网在学生宿舍、教学、科研与管理楼宇的覆盖率达到 85.32%，学校无线网覆盖学校公共区域的比例也达到 15.82%，学生人均信息点数达 0.677 个。中国教育科研网格（ChinaGrid）也取得重大进展，通过集成全国 13 个省市 20 所重点高校的计算、存储、数据、软件等信息资源，建立了聚合计算能力达到 16 万亿次、存储能力达到 180 TB（万亿字节）的网格环境。

在数字化资源建设方面，一是 2001 年启动中国高等教育文献保障系统（China Academic Library & Information System，简称 CALIS）建设工作，到 2008 年成员图书馆超过 500 家，联合目录数据库数据量达 180 万条，馆藏总量近 700 万条。二是实施大学数字博物馆建设项目，到 2008 年在 30 所高等院校建设了 10 万余件优质教学标本和特色藏品的基础资源。三是实施高等学校精品课程建设工程，充分利用高校名师资源开发精品网络课程并免费开放，以实现优质教学资源共享，到 2010 年全国高校共建成国家级精品课程 3 750 门。另外，绝大多数高校都建有教学资源库，53.4% 的高校建立了全校统一的教学资源管理平台，校均数字教学资源达 618 GB；83.72% 的高校建有电子图书资源，校均电子图书资源达 32.2 万册。

在高校现代远程教育方面，从 1999 年开始教育部先后批准 69 所高等学校开展现代远程教育（也称网络教育）试点，共开设 299 种专业、1 560 个专业点，建设了 2 万多门课程的网络资源和一批网络教学与管理系统平台，设立了 9 000 多个校外学习中心和教学点。截至 2009 年，现代远程教育试点累计招收本专科学生近 1 000 万人，毕业学生 500 多万人，开展非学历教育培训数千万人次。中央广播电视大学顺利实现由传统广播电视教育向远程开放教育的转型。

3. 职业教育信息化

相对于基础教育和高等教育，职业教育信息化建设相对滞后。2005 年 10 月，国务院发布《关于大力发展职业教育的决定》，明确提出加强职业教育信息化建设，相关建设工作开始步入快车道。到 2009 年，全国职业学校拥有计算机 230 万台，平均每 100 人拥有 11 台；70% 的学校建有计算机教室、多媒体教室、电子阅览室，60% 的学校建有不同规模的校园网。从 2000 年开始建设的高等职业教育专业教学资源库，到 2009 年已建设网络课程近 700 门。经批准，基于计算机互联网和卫星网络开展职业教育培训的企业有 20 家。在信息管理方面，投入实际应用的软件系统有全国中等职业学校学生信息管理系统、中等职业学校就业信息服务平台、高等职业学校评估数据采集系统等。

二、教育信息化的成效

20 世纪 90 年代以来，国家实施的一系列重大工程和政策措施，为我国教育信息化发展奠定了坚实的基础。面向全国的教育信息基础设施体系初步形成，城市和经济发达地区各级各类学校已不同程度地建有校园网并以多种方式接入互联网，信息终端正逐步进入农村学校；数字教育资源不断丰富，信息化教学的应用不断拓展和深入；教育管理信息化初见成效；网络远程教育稳步发展，为构建终身学习体系发挥了重要作用。教育信息化对于促进教育公平、提高教育质量、创新教育模式的支撑和带动作用初步显现①。

我们还必须清醒地认识到，加快推进教育信息化还面临诸多的困难和挑战。对教育信息化重要作用的认识还有待深化和提高；加快推进教育信息化发展的政策环境和体制机制尚未形成；基础设施有待普及和提高；数字教育

① 教育部．教育信息化十年发展规划（2011—2020 年）．

资源共建共享的有效机制尚未形成，优质教育资源尤其匮乏；教育管理信息化体系有待整合和集成；教育信息化对于教育变革的促进作用有待进一步发挥[①]。这些问题的出现与下列因素存在密切联系[②]：

① 对信息化战略地位认识不足，缺乏统筹与总体规划；

② 管理体制条块分割，缺乏强有力的协调与管理；

③ 资金投入不足，缺乏长效投入保障机制；

④ 基础设施建设不均衡，管理水平和使用效率低；

⑤ 信息化人才队伍短缺，信息素养有待提高；

⑥ 标准建设与应用明显滞后，标准采用率低。

三、数据的价值

《教育信息化十年发展规划（2011—2020 年）》对未来十年的教育信息化提出了新要求，具体包括以下四个方面。

① 面向未来，育人为本。面向建设人力资源强国的目标要求，面向未来国力竞争和创新人才成长的需要，努力为每一位学习者提供个性化学习、终身学习的信息化环境和服务。

② 应用驱动，共建共享。以人才培养、教育改革和发展需求为导向，开发应用优质数字教育资源，构建信息化学习和教学环境，建立政府引导、多方参与、共建共享的开放合作机制。

③ 统筹规划，分类推进。根据各级各类教育的特点和不同地区的经济社会发展水平，统筹作好教育信息化的整体规划和顶层设计，明确发展重点，坚持分类指导，鼓励形成特色。

④ 深度融合，引领创新。探索现代信息技术与教育的全面深度融合，以信息化引领教育理念和教育模式的创新，充分发挥教育信息化在教育改革和

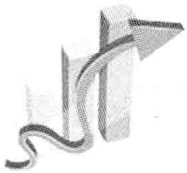
---

① 教育部．教育信息化十年发展规划（2011—2020 年）．

② “教育信息化建设与应用研究”课题组．我国教育信息化建设与应用现状调研与战略研究报告［M］．北京：高等教育出版社，2010：18－21.

发展中的支撑与引领作用。

要实现上述四个方面的要求，不断反思工作推进中存在的问题是必要的手段。应用数据挖掘，对教育信息化建设和应用中产生的数据进行收集、分析并辅助决策，可为提高教育信息化决策的科学化水平提供重要支撑。



第二章
教育数据挖掘概述

2007年，欧洲技术促进学习协会（European Association of Technology Enhanced Learning，EATEL）在希腊克里特岛举办第二届欧洲技术促进学习会议（EC-TEL2007），其间举办了“Applying Data Mining in E-Learning”研讨会（ADML’2007）。在这一研讨会之后，该领域研究者组成国际教育数据挖掘工作组（http：//www. educationaldatamining. org/），创办在线学术期刊《教育数据挖掘杂志》，并从2008年开始每年召开“教育数据挖掘国际会议”（EDM conference）。2011年，国际教育数据挖掘协会（International Educational Data Mining Society，IEDMS）成立。

一、数据挖掘

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。从商业的角度看，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数据。

从上述描述中不难发现，“知识”是其中的核心词汇，正确理解“知识”

的内涵是把握“数据挖掘”概念的关键。

亚历山大（Alexander）等人（1991）将知识分为三类：陈述性知识、程序性知识和条件性知识（转引自：盛群力，等，1998）[①]。梅耶于1987年在综合加涅和安德森的知识和技能观的基础上提出了一个广义的知识观。他将知识分为三大类：① 语义知识，指个人关于世界的知识；② 程序性知识，指用于具体情境的算法或一套步骤；③ 策略性知识，指如何学习、记忆或解决问题的一般方法，包括应用策略进行自我监控（转引自：皮连生，1996）[②]。

在数据挖掘领域，知识表现为各种有趣的数据模式，包括“概念/类描述”“频繁模式、关联和相关”“分类和预测”“聚类”“离群点”和“演变”，而不是我们传统意义上的各种知识。而这些数据模式又与数据挖掘算法有着对应关系，如关联规则算法与“频繁模式、关联和相关”相对应，贝叶斯算法、决策树算法、神经网络算法与“分类和预测”相对应，聚类算法、序列聚类算法与“聚类”相对应，时序算法与“演变”相对应。这些对应关系构成了数据挖掘的外延，有了这些对应关系以及实现这些算法的工具，数据挖掘就成为一种实用的、可操作的方法。

要完整地实现数据挖掘的价值，在数据挖掘进行前后，需要许多前期准备工作和后期分析工作。作为一个完整的工作过程，一次数据挖掘及其前期准备工作和后期分析工作，统称为一个数据挖掘项目。完成一个数据挖掘项目的过程称为数据挖掘项目实施流程，如图2-1所示。

① 数据准备：了解数据挖掘应用领域的有关情况，包括熟悉相关的背景知识，清楚使用者的需求。

② 数据选取：数据选取的目的是确定目标数据，根据使用者的需要从原始数据库中选取相关数据或样本。在此过程中，需要利用一些数据库操作对数据库进行相关处理。

---

① 盛群力，李志强．现代教学设计论［M］．杭州：浙江教育出版社，1998：58.

② 皮连生．智育心理学［M］．北京：人民教育出版社，1996：39.