普通高等院校电子商务“十二五”规划重点教材

总主编 杨坚争

数据挖掘原理与商务应用

朱小栋 徐 欣 编著

立信会计出版社
LIXIN ACCOUNTING PUBLISHING HOUSE

普通高等院校电子商务“十二五”规划重点教
总主编　杨坚争

数据挖掘原理与商务应用

朱小栋　徐　欣　编著

立信会计出版社
LIXIN ACCOUNTING PUBLISHING HOUSE

图书在版编目(CIP)数据

数据挖掘原理与商务应用/朱小栋,徐欣编著.—
上海:立信会计出版社,2013.3
普通高等院校电子商务"十二五"规划重点教材
ISBN 978-7-5429-3816-9

Ⅰ.①数… Ⅱ.①朱… ②徐… Ⅲ.①数据采集—高等学校—教材 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2013)第 066703 号

策划编辑 窦瀚修
责任编辑 徐小霞
封面设计 周崇文

数据挖掘原理与商务应用

出版发行	立信会计出版社		
地　　址	上海市中山西路 2230 号	邮政编码	200235
电　　话	(021)64411389	传　　真	(021)64411325
网　　址	www.lixinaph.com	电子邮箱	lxaph@sh163.net
网上书店	www.shlx.net	电　　话	(021)64411071
经　　销	各地新华书店		
印　　刷	常熟市梅李印刷有限公司		
开　　本	787 毫米×960 毫米	1/16	
印　　张	12	插　　页	2
字　　数	209 千字		
版　　次	2013 年 3 月第 1 版		
印　　次	2013 年 3 月第 1 次		
印　　数	1—3 100		
书　　号	ISBN 978-7-5429-3816-9/TP		
定　　价	26.00 元		

编委会名单

总　序

电子商务作为20世纪末出现的新兴产业，经过10余年的发展，已经成为世界经济中增长最快的行业之一。截至2011年3月31日，全球互联网用户达到20.95亿人，已经占到世界人口的30.2%①。截至2011年6月，全球网站总量已经达到3.46亿个，在经历了2008年的小幅挫折之后又有了大幅度的增长②。2009年，全球电子商务交易额达到16万亿美元，同比增长25%。我国2008年电子商务交易额达到31 427亿元，同比增长44.8%；2009年电子商务交易额达到38 251亿元，同比增长21.7%③；2010年电子商务交易额已经突破4万亿元，达到4.5万亿元。

电子商务的高速发展引起国家最高领导层的高度重视。2011年3月5日，在第十一届全国人民代表大会第四次会议上温家宝总理明确提出："积极发展电子商务、网络购物、地理信息等新型服务业。"国家电子商务发展"十二五"规划，各省市电子商务发展"十二五"规划都在积极制定中。

电子商务作为一种新兴行业，有以下4个鲜明特点：

(1) 电子商务是以重大技术突破和重大发展需求为基础的新兴行业。互联网技术的开发是20世纪影响力最大的技术突破。但在其开发的前30年，一直被禁锢在军事和研究领域，没有在社会上得到很好的推广。20世纪90年代，商业机构跻身于互联网世界，立即发现它的巨大潜力，并在短短的20年间形成了巨大的社会需求。电子商务正是以现代网络信息技术为基础而发展起来的一个新兴行业。

(2) 电子商务对经济社会全局和长远发展具有重大引领带动作用。实体市场与虚拟市场两者并行的局面造就了21世纪世界市场的新格局。电子商务是

---

① Internetworldstats. Com. World Internet Users and Population Stats [EB/OL]. (2011-03-31)[2011-07-07]. Internet World Stats Website: http://www.internetworldstats.com/stats.html.

② Netcraft. June 2011 Web Server Survey [EB/OL]. (2011-06-07)[2011-07-07]. Netcraft Website: http://news.netcraft.com/archives/2011/06/07/june-2011-web-server-survey.html.

③ 商务部. 中国电子商务报告[M]. 北京：清华大学出版社，2010.

以电子商务为代表，包括即时通讯、搜索引擎、网络游戏、网络广告等多种形式的互联网经济模式。电子商务正在对经济社会的全局和长远发展产生巨大的推动作用。2010年平息的腾讯和360公司的争端竟然波及10多亿网络用户，不仅影响到虚拟经济，甚至影响到整个社会的稳定①，其影响力甚至超过其他新技术。

(3) 电子商务是知识技术密集、物质资源消耗少的产业。商业活动最显著的特点就是追求高效率和低成本。20年的实践证明，最先进的信息网络技术都是首先在电子商务领域找到最好的用武之地。电子商务已经成为先进技术的聚集地和协同枢纽。特别是在交易安全领域，电子商务对技术的要求是最高的。正是因为先进技术的广泛应用，使得电子商务的交易成本远远低于传统的实体市场交易成本，从而将贝塔斯曼从中国"挤"了出去②，将最后一家传统书店从十里南京路"挤"了出去③。可以预见，未来还有更多的传统产业将步传统书店的后尘。

(4) 电子商务是成长潜力大、综合效益好的产业。相对于其他产业，电子商务的发展速度令人吃惊。淘宝网、京东商城、1号店、快钱等电子商务网站的成长历程清楚地说明了这一点。电子商务发展的同时也带来了良好的社会效益。2009年，中国邮政1/3的包裹量来自电子商务；2010年11月，淘宝网创造了167万个直接且充分就业机会，而每一人在淘宝网开店实现就业，就将带动2.85个相关产业的就业机会④。

当我们做出了上述分析之后，我们完全有理由将电子商务列为战略性新兴产业并按照战略性新兴产业的思路发展电子商务。本套教材正是从这一战略高度出发，结合电子商务发展的最新模式，为广大电子商务专业学生和电子商务从业者展现了电子商务领域的最新研究成果。

本套教材包括《电子商务原理》、《网络营销教程》、《网络信息检索与利用》、《网络营销调研技术》、《信息系统工程项目管理》、《电子商务安全与支付》、《金融电子商务》、《电子商务物流》、《移动电子商务》、《电子商务安全管理与支付》、《电子商务网站技术基础》、《电子金融学》、《电子商务统计理论与实务》、《数据挖掘

---

① 百度名片. 腾讯360之争[EB/OL]. (2011-06-18)[2011-07-07]. 百科百度：http://baike.baidu.com/view/4633773.html.

② 陈熙涵. 贝塔斯曼将关闭在华36家门店[N]. 文汇报，2008-06-17(9).

③ 许明，房浩. 南京路最后一家新华书店停业[N]. 新民晚报，2010-11-04.

④ 淘宝网数据. 淘宝网：2010年11月，淘宝网创造了167万个直接且充分就业机会. [EB/OL]. (2010-12-02)[2011-07-07]. 阿里巴巴研究中心：http://www.aliresearch.com/data/alibabag/12024/.

原理与商务应用》、《信息系统与电子商务》、《电子商务创业》等10多本，涵盖了电子商务学科的主要领域。

本套教材的特色主要表现在以下4个方面：

(1) 强调教材的先进性。针对国内外电子商务发展的最新动态，调整教材内容，使整套教材能够充分反映电子商务发展中出现的新思维、新技术和新模式；同时，揭示电子商务发展中出现的新情况和新问题，拓展读者的视野，使读者能够站在世界电子商务发展的最前沿进行电子商务发展的战略思考。

(2) 强调教材的科学性。电子商务涉及多学科知识领域的交叉，本套教材注意处理好科学性与系统性、系统性与交叉性之间的关系。结合电子商务应用性和创新性强的特点，设计科学的教学内容和实践体系，突出学生创新能力的培养。

(3) 强调理论与实践的结合。电子商务是一门实践性很强的学科，因此，在本套教材编写过程中，吸收了高校教师、理论工作者、电子商务企业家的参与。理论工作者与实际工作者思想火花的碰撞，使得理论知识与实践应用紧密结合，从而为学以致用、用以促学奠定了良好基础。

(4) 强调实践教学。在本套教材的编写过程中，笔者逐渐完善了“中国电子商务示范平台”。该平台为电子商务专业的学生提供了在线实践的机会，也为本套教材配套了多个内容密切联系的教学实验，注重形象思维和引导性操作，使学生能够在全面了解电子商务的最新发展、理解电子商务基本理论的基础上，具有电子商务应用的实际操作技能。

在组织编写本套教材的过程中，我们参考了国内外大量有关电子商务的专业文献，并得到立信会计出版社的大力支持和帮助，在此表示衷心的感谢。由于电子商务的发展迅速，本套教材从立题、撰写提纲到实际成书，虽经几番修改，仍感到许多地方还需斟酌，错误和不当之处，切望专家和读者批评指正。

杨坚争

2011年8月

前　言

从20世纪90年代初数据挖掘术语的出现到今天近20年的时间里，数据挖掘受到了学术界和产业界的广泛重视，得到了重要的发展。伴随着云时代的到来，海量的数据与强烈的知识需求矛盾更加凸显，数据挖掘还将得到更多的关注。数据挖掘在客户关系管理、电子商务、信息安全、生物科技、医疗、金融、政务、教育等许多领域有着广泛的应用。

目前，有许多数据挖掘相关的书籍陆续出版，但大多数学术性较强，且缺乏习题及实践环节，较适宜于作为研究生阶段的数据挖掘课程参考用书。本书不仅通过丰富的示例讲解数据挖掘的算法理论，而且详细地讲解企业的商务智能解决方案中如何应用数据挖掘产品。本书能培养学生运用数据挖掘技术，以及将已有专业知识综合运用的能力。本书的前导课程包括数据库原理、系统开发与设计、数据结构、软件工程等。

结合本书，有关人员已开始筹建数据挖掘的校级课程网站，将相关电子教案和课件上网，并将进一步建设校级和市级精品课程网站。相关课程的教学方法包括用多媒体教室教学和使用机房进行实验教学。作者从事数据挖掘的相关研究多年，已将数据挖掘最新的发展和一些先进的研究成果纳入本书中。

本书的主要特色：

(1) 注重内容的实践性。本书的内容涵盖如何利用相关软件产品实现数据挖掘的经典算法和技术，还涵盖数据挖掘技术在商务领域中的应用。

(2) 注重本书的应用范围。本书既适合计算机应用技术专业，也适合经管类信息管理与电子商务专业的学生学习。书中既注重从计算机应用角度来讲解数据挖掘，又注重数据挖掘与商务智能、管理科学、决策支持系统的结合。

本书实践篇中的相关实践内容与理论篇是相对应的，这样便于理论联系实际地进行教学。本书的教学组织方式如下：

方式1：48课时，按照每周3课时×16周完成，其中包括实践课16课时。

方式2：32课时，按照每周2课时×16周完成，其中包括实践课10课时。

本书的教学设计思想：根据学生的知识背景安排教学内容。可以在期中之前讲解教材的理论篇部分(数据仓库、关联分析、分类分析和聚类分析)，在期中之后讲解教材的实践篇部分。在教学中，也可以实践和设计同步进行，如在理论

课中讲解理论篇，在实践课中安排对应的实验。

本书由朱小栋和徐欣共同编著，朱小栋负责第1、第2、第3、第4、第5、第6、第7、第8、第11章的撰写，徐欣负责第8、第9、第10章的撰写。全书分为理论篇和实践篇两大模块，第1模块从理论的角度阐述数据挖掘的基本原理，第2模块从实践的角度阐述数据挖掘的商务应用。本书获得了IBM教育合作项目、上海市教委科研创新基金项目(12YZ103)、教育部人文社会科学青年基金项目(12YJC870037)和教育部高等学校博士学科点基金项目(20123120120004)的资助。本书参考引用了国内外数据挖掘研究领域的专家学者的文献资料，在此对他们的工作表示衷心的感谢。

本书可以作为全日制高等学校本专科高年级专业课教材，也可以作为研究生和有关研究人员的参考资料。由于作者水平有限，书中疏漏和错误之处在所难免，敬请广大读者批评指正。

作　者

2013年2月

目　　录

第1篇　理　论　篇

第2篇 实 践 篇

第 1 篇　理论篇

第 1 章

绪　论

1.1　数据挖掘的基础概念

人类历史上每一次工业革命都带来令人震撼的生产力发展，推动人类文明一次新的跨越。如果说 1946 年电子计算机的诞生是第三次工业革命的一次冲击波，那么可以说计算机网络的出现和发展是这个冲击波的又一次里程碑。仍然可以说，在过去 100 年的时间中，人类社会所创造的生产力比过去一切时代创造的全部生产力还要多。

21 世纪的第一个 10 年，互联网技术已趋于成熟，基于互联网的新兴产业，如无线通信、电子商务、数字电视以及物联网等得到前所未有的飞速发展。人类已迈入信息时代，第三产业占据社会产业的大部分比重，这一比重还将进一步增加。人们不仅从传统的报纸、广播电视，而且从更先进的媒介，如互联网、手机和微博等获取丰富的信息，这种信息量仍将与日俱增。例如，当前《人民日报》周一至周五版面达到 24 版，其他众多报纸的信息量也是相当丰富。从信息经济学的角度看，知道如何获取那些更有价值的信息，将获得市场先机。从数据挖掘的角度看，则要加工处理纷繁复杂并且参差不齐的信息，从中发现有价值的规律和知识，并将它们提炼出来。

数据挖掘是从大量的数据中发现隐含模式和知识，并应用这些模式和知识来进行预测以指导决策的过程。自 20 世纪 70 年代关系数据库理论推出以来，数据挖掘受到了学术界和工业界的广泛关注。20 世纪 80 年代以来，为解决“数据丰富，知识贫乏”的困境，数据库中的知识发现 KDD(Knowledge Discovery in Database)和数据挖掘技术作为数据库与统计学、人工智能、机器学习等技术的交叉学科和技术，获得了巨大成功和持续发展。

1.1.1　数据

数据或称资料，它涉及事物的存在形式。它是关于事件的一组离散、客观的事实描述，是构成信息和知识的原始材料，是载荷或记录信息的按一定规则排