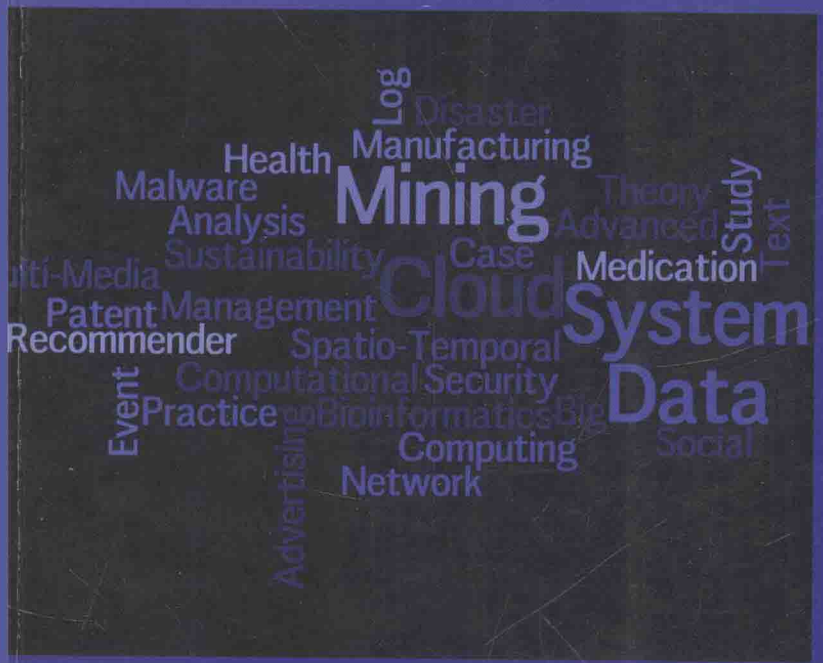
DATA MINING
Where Theory Meets Practice

数据挖掘的应用与实践

——大数据时代的案例分析



国际数据挖掘领域知名专家

李涛 等 著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

DATA MINING
Where Theory Meets Practice

数据挖掘的应用与实践

——大数据时代的案例分析

李涛 等 著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

数据挖掘的应用与实践:大数据时代的案例分析/李涛等著. —厦门:厦门大学出版社,2013.10
ISBN 978-7-5615-4294-1

Ⅰ. ①数… Ⅱ. ①李… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2012)第 101326 号

厦门大学出版社出版发行
(地址:厦门市软件园二期望海路 39 号 邮编:361008)
http://www.xmupress.com
xmup @ xmupress.com
厦门市明亮彩印有限公司印刷
2013 年 10 月第 1 版 2013 年 10 月第 1 次印刷
开本:787×1092 1/16 印张:27.25 插页:2
字数:736 千字 印数:1~2 000 册
定价:49.00 元
如有印装质量问题请与承印厂调换

内容简介

本书以笔者所带领团队的数据挖掘工作为基石，架设起研究和应用的桥梁，帮助读者们从应用实例中学习数据挖掘。

本书的宗旨是以各个领域的实际应用为导向，始终以实际案例来讲解应用之下的技术和理论。具体而言，本书对每个案例都有详细的解析，全面介绍了如何将一个实际问题抽象和转化为数据挖掘的问题，然后利用数据挖掘的理论和方法加以解决，让读者明白来龙去脉。目的是切实指导数据挖掘的应用实践，建立起研究和应用的桥梁。

本书注重原理和思想，不过多纠缠技术细节，尽量简化数学公式和模型，强调其背后的基本思想和出发点。本书不按理论和技术来划分章节，而是以实际的应用案例来贯穿始终，通过数据挖掘应用的实例来介绍如何应用和学习数据挖掘技术。

本书包括16 个章节，第一章对数据挖掘做一个简单的介绍，其余的15 章分为四个部分：

1. 第一部分：数据挖掘在计算机系统方面的应用，包括系统日志和事件的挖掘、数据挖掘在云计算中的应用、恶意软件智能检测；
2. 第二部分：数据挖掘在社会服务方面的应用，包括社交媒体挖掘、推荐系统、智能广告、灾难信息管理；
3. 第三部分：不同的数据类型下的典型应用，包括文本挖掘、多媒体数据挖掘、空间数据挖掘；
4. 第四部分：数据挖掘的一些综合应用，包括生物信息学和健康医疗、数据挖掘在建筑业中的应用、数据挖掘在高端制造业的应用、数据挖掘在可持续发展的应用和在专利领域中的应用。

本书各章相对独立，用户可以直接阅读跟自己具体应用领域相关的部分，而不用按照顺序进行阅读。值得一提的是，本书中的绝大部分内容是基于笔者团队的科研项目和研究积累。我们尽量提供书中涉及到原始数据和开源的软件平台，方便读者学习和使用。对笔者来说，本书既是一个阶段性的研究和应用工作总结，也是对未来大数据应用研究的铺垫。我可以无愧地说，写这本书是很严肃和很有诚意的。

前 言

笔者长期从事数据挖掘研究和教学工作，经历了从最初数据挖掘基础研究的兴起到如今数据挖掘应用百花齐放这样一个时代的变迁，深刻体会到研究和应用两者间不可分割的联系：数据挖掘研究源于实践中的实际应用需求，用具体的应用数据作为驱动，以方法、工具和系统作为支撑，最终将发现的知识和信息运用到实践中去，从而提供量化的、合理的、可行的，并且能够产生巨大价值的信息。数据挖掘是理论技术和实际应用的完美结合，所以，数据挖掘践行者们都要时刻坚定——**应用是检验研究的最高标准**这样的理念。

数据挖掘是大数据中最关键和最有价值的工作

国际知名的行业战略咨询公司麦肯锡（McKinsey & Company）在其2011 的大数据研究报告中明确指出，在诸如卫生保健、公共部门管理、个人位置信息服务等领域，有效利用大数据能够带来每年超过千亿美元的经济价值。同时，在零售、制造等行业应用大数据解决方案，能够给企业带来相当巨大的资金效率和生产效率提升。IBM、谷歌、微软、阿里巴巴等IT 巨头也将大数据描述成一种颠覆性的技术，其力量在将来足以影响和改变我们每一个人，甚至一个行业和一个国家。充分发挥大数据的巨大潜力，数据的产生和收集是基本，数据挖掘（知识发现）是工具和手段，是大数据中最关键和最有价值的工作。

在大数据时代，利用数据挖掘提升竞争力已成为各行各业都在追逐和挑战的目标，精彩的故事也层出不穷。在2012年美国总统大选中，奥巴马最终连任，其大数据挖掘与分析团队居功至伟。该团队利用两年时间收集、处理与整合了海量数据，将以往选举数据、居民基本信息、社交网络等数据整合在一个数据仓库中，利用数据挖掘算法与统计模型，预测有效选民、进行精准广告投放、优化资源配置，最终帮助奥巴马团队募集到10 亿美金资金且最终赢得选举。

实践出真知

了解和学习数据挖掘，可以让我们为迎接大数据时代的激烈竞争作好准备。在长期的数据挖掘研究和教学工作中，笔者发现学习数据挖掘主要有两大难点：其一是数据挖掘是一个交叉学科，融合了统计分析、模式识别、机器学习、信息检索、数据库、信息论和最优化算法等领域的学术思想，所以其技术理论基础比较多并且分散。初学者往往很难把握数据挖掘的整个脉络，将技术理论的众多知识点系统有机地联系起来。其二是技术理论和应用实践容易脱节，初学者往往不能很好地将两者相结合。数据挖掘的应用性很强，包括诸如关联规则挖掘、时间

序列模式挖掘、分类预测、聚类分析、链接分析和异常检测等多种功能。与此同时，不同的功能通常有不同的理论和技术基础，而每一个具体的应用案例往往涉及多个不同的功能。对于有兴趣进行数据挖掘应用实践的读者们来说，他们常常有这样的困惑，如何将实际问题和已经学到的方法原理联系起来，如何将数据挖掘技术有效地运用在实际应用中，给使用者带来价值。

现今市面上已经有书籍全面地介绍数据挖掘的技术理论基础，详细解析各种挖掘算法的原理和细节。同时还有书籍专门介绍各种数据挖掘算法的实现和相关工具的使用。但这些书籍侧重于介绍单个数据挖掘功能及其相关算法原理，并没有涉及如何将数据挖掘应用到具体实践。目前，关于数据挖掘技术的应用案例分析都零星分散在一些会议论文、期刊和报告之中，并没有专门的书籍来详述。

鉴于此，笔者希望本书能填补这个空白。本书以笔者所带领团队的数据挖掘工作为基石，架设起研究和应用的桥梁，帮助读者们从应用实例中学习数据挖掘。具体而言，本书以各个领域的实际应用为导向，始终以实际案例来讲解应用之下的技术和理论。本书对每个案例都有详细的解析，全面介绍了如何将一个实际问题抽象和转化为数据挖掘的问题，然后利用数据挖掘的理论和方法加以解决，让读者明白来龙去脉。目的是切实指导数据挖掘的应用实践，建立起研究和应用的桥梁。

在编写本书时，笔者制订了两个原则。首先是内容要尽量全面，即覆盖当前数据挖掘的主要应用。在介绍每个应用案例时，详细阐述应用的背景，该领域中数据的来源和特点，数据采集与预处理方式，应用领域中数据挖掘的任务和实施数据挖掘技术的难点。同时提供相应的数据挖掘算法分析、工具设计以及系统实现。其次是要条理清晰、便于理解。一方面本书主要面向的是热爱和关心数据挖掘技术的学术界和工业界读者，帮助他们更好地理解研究的目的和应用的基础；另一方面，笔者也争取让没有太多相关技术背景的读者可以通过阅读本书了解数据挖掘的意义和价值，可以看出数据挖掘是如何被广泛地应用于实际案例并成为解决各种问题的核心工具。

笔者希望这是一本既通俗易懂，适合不同背景的读者阅读，同时又比较全面，且融入最新前沿技术和应用的数据挖掘书籍，也欢迎各大高校的师生把此书作为数据挖掘和机器学习课堂的实践教材和参考书籍。

致谢

本书中涉及的数据挖掘研究项目得到了美国佛罗里达国际大学计算机学院(School of Computer Science, Florida International University)、美国国家自然科学基金(National Science Foundation，NSF)、美国国土安全部(Department of Homeland Security，DHS)、美国军方研究实验室（Army Research Office，ARO）、中国国家自然科学研究基金、中国福建省自然科学基金、美国国际商业机器公司研究中心（IBM Research）、日本电气股份有限公司研究中心（NEC Research）、美国施乐公司研究中心（Xerox Research）、中国四川长虹电子集团公司、中国福建省厦门人才市场、中国福建省厦门市信息技术中心的资助。厦门大学信息科学与技术学院和厦门理工学院计算机与信息工程学院对本书的编写和出版给予了大力

支持，在此一并致谢！

关于作者

笔者2004年7月毕业于美国罗彻斯特大学（University of Rochester），获计算机科学博士学位。现为美国佛罗里达国际大学（Florida International University，FIU）计算机学院终身教授，数据挖掘实验室主任，是国内多家高校的客座教授。笔者长期从事关于大数据分析、数据挖掘和信息检索等方面的研究，在基于矩阵方法的数据挖掘和学习，智能推荐系统，音乐信息检索，系统日志数据挖掘，数据挖掘的各种应用等方面做出了有影响力的工作，在国际著名会议及期刊上已发表超过两百篇文章。同时，笔者是数据挖掘和知识发现的国际权威期刊ACM Transactions on Knowledge Discovery from Data (ACM TKDD), IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), 和Knowledge and Information Systems （KAIS）的副主编。

从2005年至今，笔者作为独立的项目承担人，申请到超过230 万美金的科研项目（包括五项美国自然科学基金项目以及两项美国军方实验室科研项目），作为重要的项目承担人和合作者，申请到超过1000万美金的科研项目。于2006年获美国自然科学基金委颁发的杰出青年教授奖（NSF CAREER Award，2006-2010），并多次获得IBM学院研究奖(IBM Faculty Research Awards) 和2010 IBM大规模数据分析创新奖（2010 IBM Scalable Data Analytics Innovation Award）。

本书由笔者的数据挖掘团队成员执笔编写，欢迎读者积极反馈。各个章节的作者如下：

- 系统日志和事件的挖掘（唐良，李涛）
- 数据挖掘在云计算中的应用（姜页希，李涛）
- 数据挖掘在恶意软件检测中的研究与应用（叶艳芳，李涛）
- 社交媒体挖掘（李晶轩，李涛）
- 推荐系统（李磊，郑思婷，姜姗，洪文兴，李涛）
- 智能广告（李磊，李涛）
- 灾难信息管理（郑理，李涛）
- 文本挖掘（沈超，李涛）
- 多媒体数据挖掘（陆文婷，李晶轩，李涛）
- 空间数据挖掘（周武柏，李泓泰，朱顺痣，李涛）
- 数据挖掘在生物信息和健康医疗中的应用（曾而良，李涛）
- 数据挖掘在建筑业中的应用（周绮凤，杨帆，李涛）
- 数据挖掘在高端制造业的应用（曾春秋，郑理，李磊，李晶轩，李涛）

- 数据挖掘在可持续发展的应用（薛维，李涛）
- 数据挖掘在专利领域中的应用（张龙晖，李涛）

网站和联系方式

与本书配套的网站地址是: http://bigdata-node01.cs.fiu.edu/dm-book （美国网站）和http://bigdata.xmu.edu.cn/dm-book （中国网站）。该网站不仅收录了多个相关资源的链接，还提供了一些相关数据、程序和工具给大家下载使用。我们也欢迎大家将更多的反馈意见和修改建议发邮件到towerlee@xmu.edu.cn。

李涛

2013年09月16日

目 录