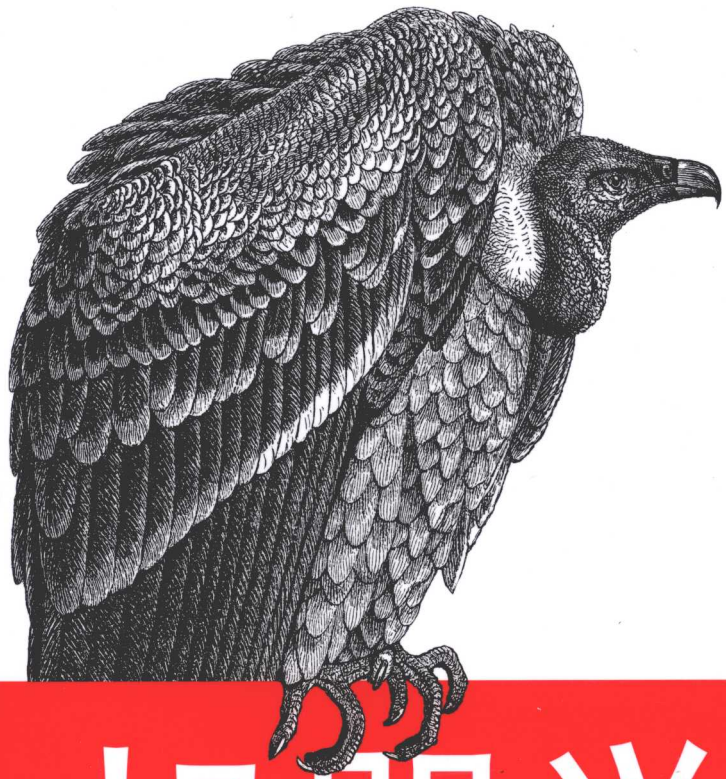
Machine Learning for Hackers

HZ BOOKS
华章科技



机器学习

实用案例解析

Drew Conway & John Myles White 著
陈开江 刘逸哲 孟晓楠 译
罗森林 审校

O'REILLY®
机械工业出版社
China Machine Press

机器学习：实用案例解析

Drew Conway & John Myles White 著
陈开江 刘逸哲 孟晓楠 译
罗森林 审校

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权机械工业出版社出版

机械工业出版社

图书在版编目（CIP）数据

机器学习：实用案例解析/（美）康威（Conway, D.）等著；陈开江，刘逸哲，孟晓楠译. —北京：机械工业出版社，2013.3
（O'Reilly精品图书系列）

书名原文：Machine Learning for Hackers
ISBN 978-7-111-41731-6

I. 机… II. ①康… ②陈… ③刘… ④孟… III. 机器学习 IV. TP181
中国版本图书馆CIP数据核字（2013）第042873号

北京市版权局著作权合同登记
图字：01-2012-4851号

书　　名/　机器学习：实用案例解析
书　　号/　ISBN 978-7-111-41731-6
责任编辑/　秦健
封面设计/　Karen Montgomery，张健
出版发行/　机械工业出版社
地　　址/　北京市西城区百万庄大街22号（邮政编码 100037）
印　　刷/　藁城市京瑞印刷有限公司印刷
开　　本/　178毫米×233毫米　16开本　20印张（含1印张彩插）
版　　次/　2013年4月第1版　2013年4月第1次印刷
定　　价/　69.00元（册）

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010)88378991　88361066
购书热线：(010)68326294　88379649　68995259
投稿热线：(010)88379604
读者信箱：hzjsj@hzbook.com

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了Make杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——Wired

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

译者序

当今各行业，尤其是互联网，数据规模越来越大，要从中有效地发现模式来提高生产力，用传统的方式已经几乎不可能，只能借助计算机来完成诸多使命。因此，机器学习这一新兴的学科变得越来越重要，它已经在搜索、推荐、数据挖掘等多个领域闪耀光芒。机器学习是一门交叉学科，内容涉及概率论、统计学、高等数学、计算机科学等多门学科。该学科致力于设计一种让计算机具有“学习”能力的算法，通过发现经验数据中隐藏的模式，实现对未知数据的预测。

大数据时代是机器学习最美好的时代，因为数据不再是问题，各类问题都可以收集到海量的数据。但是，对于很多人来说，这一门交叉学科本身却神秘而陌生，对于没有系统学习过相关基础学科的人来说尤其感到“高不可攀”。如今已出版的机器学习相关书籍中，很多都有这个特点：公式多，晦涩难懂。这让很多程序员出身的人望而却步。然而，在第一次读到本书的英文版时，译者就彻底相信：机器学习完全可以讲解得通俗易懂，让知识的传递实现“润物细无声”。

本书秉承的原则是：实践出真知，只要多动手，没有攻克不了的技术难题。因此作者预期的阅读对象是如电脑黑客般的人，要求对技术有发自内心的求知欲和好奇心，愿意自己动手而非纸上谈兵。全书精心选择了12个机器学习案例，由浅入深，面面俱到，既有基础知识（如数据分析），也有当前热门的社交网站推荐案例。书中的每一个案例都由作者娓娓道来，逐一剖析关键算法的代码，没有丝毫学究气息，触动每个机器学习初学者的内心最深处。

书中所有算法都采用R语言实现。R语言是一门用于统计学的开源脚本语言，基于它的开源性，有来自世界各地的开源拥护者贡献的各种统计学相关的程序包，稳定且方便，尤其是它对数据可视化的支持，更是一柄利器，既轻巧又实用。书中所有源代码和数据在

原书的官方网站上都可以免费下载。在阅读过程中，犹如作者亲至身侧，为你讲解代码和思路，为你排除错误和优化效果。

全书案例既有分类问题，也有回归问题；既包含监督学习，也涵盖无监督学习。所选择的案例妙趣横生，如分析UFO目击记录、破译密码、预测股票、分析美国参议员“结党”的情况，等等，这里就不“剧透”了，大家自己去享受学习的乐趣吧。

书中12个案例之间的依赖关系不是特别强（除R语言基础知识外，其余某几章仅有个别知识点之间存在依赖性），可以像连续剧一样，逐一播放，也可以像一个个小品一般，挑感兴趣的内容分别播放。学习完这些案例之后，相信你会窥见机器学习的一斑，然后再根据自己的实际情况更深入地学习。

本书翻译工作由三位来自互联网世界的工程师通力协作完成，其中，来自新浪微博的陈开江负责完成前言及第1～4章的翻译；来自阿里B2B的刘逸哲负责完成第5、8、9和11章的翻译；来自阿里一淘的孟晓楠负责完成第6、7、10和12章的翻译；同时，全书审校工作由来自北京理工大学的罗森林教授义务承担。

本书能够得以出版，首先要感谢机械工业出版社的吴怡编辑，是她给了我们三位工程师这个学习知识并传递知识的机会，她经验丰富，在翻译过程中给予了我们许多建设性的指导意见。其次，要感谢罗森林教授，他在百忙之中为我们担任全书的审校工作，从而让国内的机器学习者能感受到这本书应有的魅力。最后，我们要感谢互联网，因为译者与本书的缘分始于互联网，从看到原书、报名翻译、组成翻译团队、翻译过程中的讨论，所有这样都是通过互联网完成的。

虽然经过罗森林教授认真审校并且给我们提出了宝贵意见，但是由于译者本身水平有限，书中译文势必还存在不妥甚至错误之处，恳请机器学习界的广大前辈、同仁们不吝赐教，促使我们继续为大家更好地传递先进技术，让更多机器学习爱好者成为机器学习的黑客。

我们坚信集体智慧是再高的个人智慧都无法企及的，因此真诚希望大家一起来贡献自己的智慧。三位译者的微博分别为：*http://weibo.com/kaijiangidan*（陈开江，@刑无刀）、*http://weibo.com/liuyizhe10*（刘逸哲，@刘逸哲）、*http://weibo.com/u/1911115643*（孟晓楠，*@XiaonanMeng*）。无论是对翻译本身有任何意见或建议，还是对机器学习方面有心得，都欢迎大家到我们的微博上交流、切磋，我们一起贡献自己的智慧，在集体智慧中互相学习，共同进步。

作者介绍

Drew Conway 机器学习专家，拥有丰富的数据分析与处理工作经验。目前主要利用数学、统计学和计算机技术研究国际关系、冲突和恐怖主义等。他曾作为研究员在美国情报和国防部门供职数年。他拥有纽约大学政治系博士学位，曾为多种杂志撰写文章，是机器学习领域的著名学者。

John Myles White 机器学习专家，拥有丰富的数据分析与处理工作经验。目前主要从理论和实验的角度来研究人类如何做出决定，同时还是几个流行的R语言程序包的主要维护者，包括ProjectTemplate和log4r。他拥有普林斯顿大学哲学系博士学位，曾为多家技术杂志撰稿，发表过许多关于机器学习的论文，并在众多国际会议上发表演讲。

译者介绍

罗森林 博士，教授，博导。现任北京理工大学信息系统及安全对抗实验中心主任、专业责任教授。国防科技工业局科学技术委员会成员；《中国医学影像技术杂志》、《中国介入影像与治疗学》编委会委员；全国大学生信息安全技术专题邀请赛专家组副组长；中国人工智能学会智能信息安全专业委员会委员等。主要研究方向为信息安全、数据挖掘、媒体计算、中文信息处理等。负责或参加完成国家自然科学基金、国家科技支撑计划、863计划、国家242计划等省部级以上项目40余项。已发表学术论文90余篇，出版著作8部，出版译著1部，获授权专利3项。

陈开江 新浪微博搜索部研发工程师，曾独立负责微博内容反垃圾系统、微博精选内容挖掘算法、自助客服系统（包括自动回复、主动挖掘、舆情监测）等项目，目前主要从事社交挖掘、推荐算法研究、机器学习、自然语言处理相关工作，研究兴趣是社交网络的个性化推荐。

刘逸哲 阿里巴巴，CBU基础平台部搜索与推荐团队核心技术与query分析方向负责人，机器学习技术领域及圈子负责人。曾任中国雅虎相关性团队、自然语言处理团队算法工程师；AvePoint.inc开发工程师，从事企业级搜索引擎开发。研究兴趣是机器学习、自然语言处理及个性化推荐等算法在大规模数据上的应用。

孟晓楠 一淘广告技术，阿里非搜索广告算法负责人，负责用户行为分析、建模与细分，RTB竞价算法，展示广告CTR预估与SEM优化。曾工作于网易杭州研究院，参与过分布式全文检索系统和网易博客产品的数据挖掘算法开发。研究兴趣是计算广告技术、机器学习、大数据技术、信息检索等。

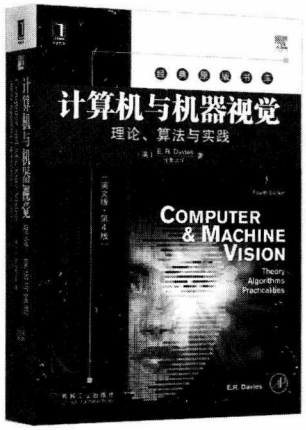
封面介绍

本书封面动物是兀鹫（griffon vulture，鹰科）。这种庞然大鸟分布在旧大陆（即欧、亚、非）较暖和的地区，也就是说地中海附近。

这类鸟头部的羽毛呈白色且稀少，翅膀宽大，尾巴短小。成年兀鹫——身高在0.9~1.1m、翅宽平均在2.3~2.8m——通常身体羽毛呈黄棕色，间杂黑色，颈部周围羽毛呈白色。兀鹫是一种食腐动物，只捕食死尸。

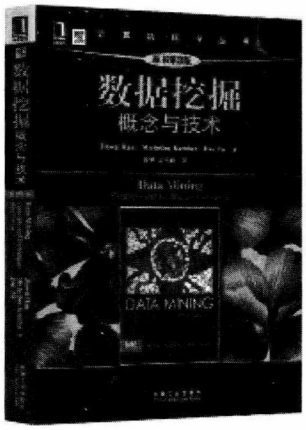
兀鹫最长寿命现存记录是41.4年（养殖场记录）。它们广泛分布在欧洲南部、非洲北部山区，以及亚洲。每次产蛋仅一枚。

推荐阅读



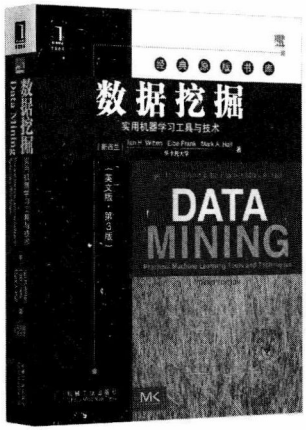
计算机与机器视觉：理论、算法与实践

作者：E. R. Davies ISBN：978-7-111-41232-8 定价：128.00元



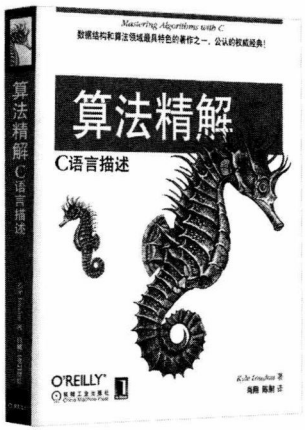
数据挖掘：概念与技术

作者：Jiawei Han ISBN：978-7-111-39140-1 定价：79.00元



数据挖掘：实用机器学习工具与技术

作者：Ian H. Witten等 ISBN：978-7-111-37417-6 定价：108.00元



算法精解：C语言描述

作者：Kyle Loudon ISBN：978-7-111-39426-6 定价：79.00元

推荐阅读

HTML 5应用开发实践指南

作者：Zachary Kessin ISBN：978-7-111-41451-3 定价：49.00元

机器学习：实用案例解析

作者：Drew Conway 等 ISBN：978-7-111-41731-6 定价：69.00元

Visual C++并行编程实战

作者：Colin Campbell 等 ISBN：978-7-111-38806-7 定价：59.00元

PHP精粹：编写高效PHP代码

作者：Davey Shafik 等 ISBN：978-7-111-39907-0 定价：59.00元

程序员度量：改善软件团队的分析学

作者：Jonathan Alexander ISBN：978-7-111-40140-7 定价：59.00元

编写可读代码的艺术

作者：Dustin Boswell 等 ISBN：978-7-111-38544-8 定价：59.00元

Android程序设计

作者：Zigurd Mednieks ISBN：978-7-111-40184-1 定价：79.00元

Android 应用开发攻略

作者：Ian F. Darwin ISBN：978-7-111-41411-7 定价：99.00元

精通搜索分析

作者：Brent Chaters ISBN：978-7-111-39681-9 定价：69.00元

目录

前言

致机器学习的黑客们

为了更好地阐释本书的切入点，很有必要对“机器学习”与“黑客”这两个词语下个定义。

什么是机器学习？简单来说，机器学习就是一套工具和方法，凭借这些工具和方法我们可以从观测到的样本中提炼模式、归纳知识。举个例子，如果我们要让计算机识别信封上的邮政编码，那么需要这样的数据：首先是信封的图片数据，其次信封上必须有收件人的邮政编码。换句话说，在特定情境下，我们可以记录研究对象的行为，从中学习，然后对其行为建模，该模型反过来促进我们对该情境有更深入的理解。在实际项目中，机器学习需要数据，而且对当今的应用程序来说不是一点点数据（有可能达TB级的数据）。大多数机器学习技术不担心没有数据，现代企业运营所产生的数据量之大意味着这些技术应用的春天来了。

什么是黑客呢？在我们眼中，黑客就是喜好用新技术进行实验、解决问题的人，而与“网络罪犯”、“不法少年”这些世俗字眼完全无关。如果你曾经手捧O'Reilly最新出版的一本关于一门新计算机语言的图书，跌跌撞撞地敲下代码，并最终调试通过了你的第一个程序，那么你就称得上是一名黑客。或者，你曾把新买来的小机器大卸八块，并最终弄懂了它的整个机械结构，那么你也是个黑客。通常，黑客这样做并无特别的原因，只是为了享受这个过程，只是为了要彻底了解一门未知的技术。

计算机黑客对事物原理有一种与生俱来的好奇心和动手的热情，他们（与之相对应的还有汽车黑客、生活黑客、美食黑客，等等）还有软件设计和开发的经验，他们就是以前写过程序的人，甚至很可能使用过很多种语言。对于一个黑客来说，UNIX不是一个四

个字母的单词，工作时用命令行导航和bash shell操作与用图形用户界面一样熟练。处理文本时，黑客首先想到的就是正则表达式，以及sed、awk、grep这些工具。在本书的写作中，我们也假设读者在这些方面的知识水平比较高。

本书的组织结构

机器学习融合了许多传统领域的理论和技术，诸如数学、统计学和计算机科学等。因此，学习本学科存在许多切入点。由于数学和统计学是机器学习的理论基础，因此新手应该在一定程度上掌握机器学习基础技术的范式。市面上已有很多这方面的优秀书籍，如Hastie、Tibshirani、Friedman三人的经典著作《统计学习基础》（The Elements of Statistical Learning，[HTF09]，完整信息见参考书目）[注1]。但是在黑客们的人生理念里，很重要的一部分就是：边做边学。很多黑客在面对问题的时候，更习惯于在实际操作过程中寻找解决方案，而非从理论基础出发推导解决方案。

从这个角度而言，教授机器学习的另一种方法就是采用案例式教学。例如，在讲解推荐系统时，我们会提供一批训练样本和一版模型，然后看看模型如何使用训练样本。类似的参考书有很多，比较新的一本是Segaran的《集体智慧编程》（Programming Collective Intelligence，[Seg07]）。以上的讨论只是介绍了操作方式，却没有解释为什么这样做。在理解了一个方法的原理时，我们也许还想知道为何这个方法适用于某个情境，或者为何解决了某个特定的问题。

因此，为了给机器学习的黑客们提供一本更全面的参考书，我们必须在学科理论的深度和应用探索的广度之间寻求一个平衡。为了达到这个目的，我们决定采用案例教学的方式来教授机器学习。

最好的学习方法是：首先带着问题思考，然后专心研究解决问题的方法。这也是案例教学能有效执行的机理所在。不同之处在于，我们并不是拿一些还没有成熟解决方法的机器学习问题来举例，而是讨论一些已深入理解和广泛研究的问题，并列举了一些特定的案例辅以说明。对于这些案例，有些方法可以很好地解决问题，而有些方法却根本不适用。

基于上述指导思想，本书的每一章都是基于特定机器学习问题的独立案例研究。本书前几个案例从分类讲到回归（第1章会进一步讨论），然后又讨论了聚类、降维、最优化问题等。需要说明的是，不是所有的问题都可以简单地归为分类问题或者回归问题，书中涉及的一些案例同时包含分类与回归问题（有些比较明显，有些不易察觉）。以下是本书中出现的所有案例研究的简要介绍（按出现先后顺序）。

注1：　这本书也可以从*http://www-stat.stanford.edu/~tibs/ElemStatLearn/*免费下载。

文本分类：垃圾邮件识别

这一章介绍由电子邮件文本数据引起的二分类问题。在此要处理的是机器学习中的经典问题：将某个输入识别为两个类中的一个，在这里指的就是正常邮件（合法的电子邮件）或垃圾邮件（用户不希望看到的邮件）。

项目排序：智能收件箱

与上个案例一样，这里还是采用电子邮件文本数据，但是不再研究二分类问题，而是上升到研究一组具体的类别。具体来说，就是要在某一电子邮件中识别并抽取适当的特征，这些特征使该邮件在所有邮件中处于优先阅读的位置。

回归模型：预测网页访问量

现在介绍机器学习的第二个基本工具——线性回归。这里要处理的是关系大致逼近一条直线的数据。在这个案例研究中，目的是预测互联网上排名前1000（2011年）的网页访问量。

正则化：文本回归

有时候，我们并不能用一条直线很好地描述数据间的关系。为了描述这个关系就要用另一种函数来拟合，但同时又要防止出现过拟合。正则化的方法可以克服这一问题，同时通过一个案例加以说明，主要目的是理解O'Reilly图书中词与词之间的关系。

最优化：密码破解

机器学习中几乎每一个算法都可以看成是最优化问题，比如，将预测错误率最小化。这里绍一个经典的算法来实现最优化，并尝试用这个算法破解一段字母密码。

无监督学习：构建股票市场指数

到目前为止我们的讨论还局限于有监督的学习技术。在此要介绍机器学习方法上的另一半：无监督学习。两者最主要的不同是：有监督学习方法是使用结构化数据进行预测，而无监督学习是为了在数据中发现结构。因此，我们将用一批股票市场的数据来构建一个指数，这个指数可以衡量整体市场行情的好坏。

空间相似度：用投票记录对美国参议员聚类

这里介绍这样一个概念：样本点的空间距离。为了实现对参议员聚类，需要设计距离的测算方法，以及样本点基于空间距离的聚类方法。我们用美国参议员记名投票的数据，根据其所得投票记录对这些立法者进行聚类。

推荐系统：给用户推荐R语言包

为深入讨论空间相似度，我们将讨论如何搭建一个基于样本空间密度的推荐系统。这一章介绍K近邻算法，并根据程序员安装的R语言函数包，用这个算法来给他们推荐其他的R语言包。

社交网络分析：在Twitter上感兴趣的人

这一章会结合之前讨论过的许多概念，并引入一些新的概念与方法，用Twitter数据设计并搭建一个“可能感兴趣的人”的推荐系统。在这个例子中，将搭建一个系统用于下载Twitter数据，发现其中的圈子，然后用基本的社交网络分析技术向用户推荐可能感兴趣的人。

模型比较：给你的问题找到最佳算法

最后一章讨论的是用于选择解决问题的最佳机器学习方法的技巧。这一章将介绍最后一个算法——支持向量机，并采用在第3章中介绍的垃圾邮件数据来比较其与其他算法的优劣。

我们在探索这些案例的过程中用到的基本工具就是R统计编程语言（*http://www.r-project.org/*）。R语言非常适合于机器学习的案例研究，因为它是一种用于数据分析的高水平、功能性脚本语言。很多基础算法框架已经内置在R语言中，或者已经在一些R语言包中实现了，这些包可以在综合R档案网（Comprehensive R Archive Network，CRAN）[注2]上找到。这可以避免为每一个实际项目写基础功能代码，从而把我们从重复劳动中解放出来，把精力放在思考问题的本身上。

本书约定

本书使用了以下排版约定：

斜体

用于新术语、URL、电子邮件地址、文件名与文件扩展名。

等宽字体（`Constant width`）

用于表明程序清单，以及在段落中引用的程序中的元素，如变量、函数名、数据库、数据类型、环境变量、语句、关键字。

等宽粗体（**`Constant width bold`**）

用于表明命令，或者需要读者逐字输入的文本内容。

等宽斜体（*`Constant width italic`*）

用于表示需要使用用户提供的值或者由上下文决定的值来替代的文本内容。

注意： 这个图标代表一个技巧、建议或一般性说明。

注2： 关于CRAN的更多信息，请浏览*http://cran.r-project.org/*。