Springer

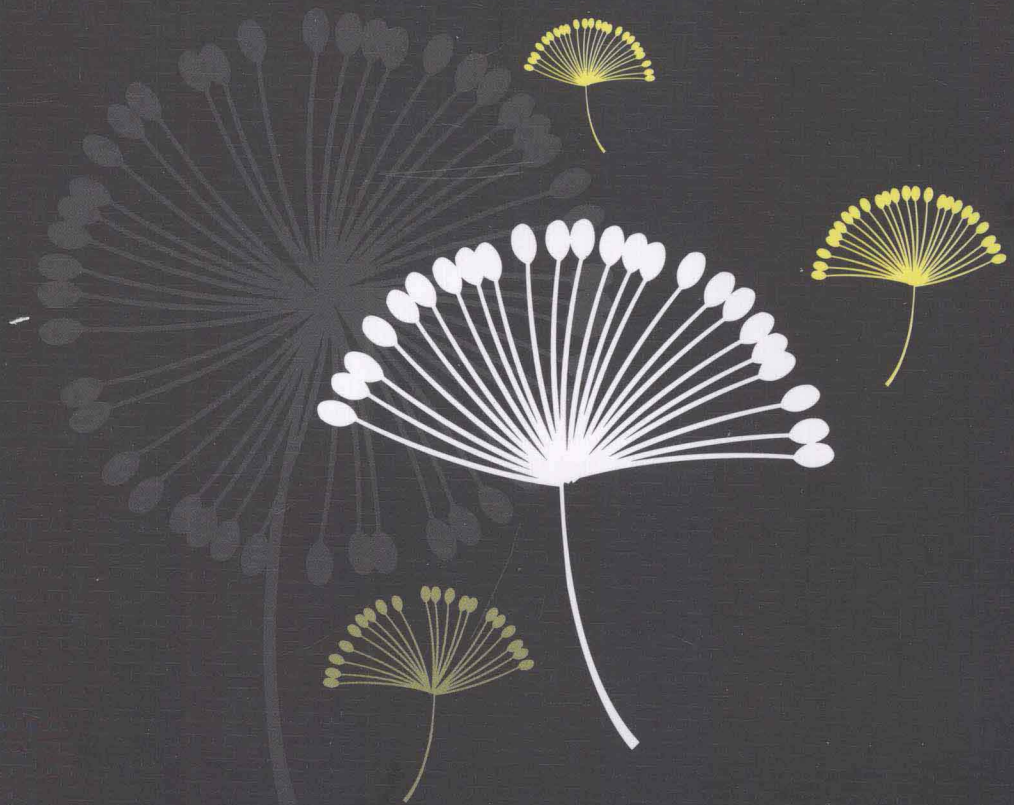
世界著名计算机教材精选

Web数据挖掘

（第2版）

Bing Liu 著

俞 勇 等译



WEB DATA MINING Second Edition

清华大学出版社

世界著名计算机教材精选

Web 数据挖掘

（第 2 版）

Bing Liu 著
俞 勇 等译

清华大学出版社
北 京

图书在版编目（CIP）数据

Web 数据挖掘（第 2 版）/（美）刘兵著；俞勇等译. —北京：清华大学出版社，2013.1
书名原文：Web Data Mining, Second Edition
世界著名计算机教材精选
ISBN 978-7-302-29870-0

Ⅰ. ①W…　Ⅱ. ①刘…　②俞…　Ⅲ. ①数据采集–教材　Ⅳ. ①TP311.13

中国版本图书馆 CIP 数据核字（2012）第 197385 号

责任编辑：龙启铭
封面设计：傅瑞学
责任校对：李建庄
责任印制：何　芊

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn,010-62795954
印 装 者：三河市李旗庄少明印装厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张：**28.25　　**字　　数：**702 千字
版　　次：2009 年 4 月第 1 版　2013 年 1 月第 2 版　　**印　　次：**2013 年 1 月第 1 次印刷
印　　数：1～3000
定　　价：59.50 元

产品编号：044329-01

第 1 版译者序

作为互联网上最重要的应用之一，Web（万维网）提供了便捷的文档发布与获取机制，并逐步成为各类信息资源的聚集地。据 Google 于 2008 年发布的官方报告，他们已经在互联网上发现超过 1 万亿个 Web 文档，而且这个数字还在以每天新增几十亿的速度持续增长。面对如此巨大的信息量，普通 Web 用户往往迷失其中，他们迫切需要一种机制快速定位到所需信息。Web 挖掘便应运而生，伴随 Web 的发展而备受关注。它建立在信息检索、数据挖掘以及知识管理等技术的基础上，通过对大量 Web 文档进行分析来获得隐含的知识和模式，从而帮助人们更好地进行信息搜索和决策制定。也正是 Web 挖掘技术的不断进展推动了 Web 的进一步蓬勃发展。

目前 Web 挖掘已经引起了学术界、工业界、社会学家的广泛关注，也吸引了众多研究人员与开发人员投身其中。国内外很多大学与研究机构先后开设了 Web 挖掘课程。但长期以来并没有专门针对 Web 挖掘的教材与专著。刘兵教授 2006 年出版的这本著作填补了该领域的空白。该教材针对 Web 挖掘中众多关键主题进行了深入分析。清华大学出版社独具慧眼，决定将该书翻译成中文版在国内出版，这必将对我国 Web 挖掘的教学与研究产生积极的推动作用，有幸承担该书的翻译工作，我们感到十分荣幸。

本书是由伊利诺伊大学芝加哥分校（UIC）的刘兵（Bing Liu）教授历经一年的时间所著的“Web Data Mining”的翻译版。刘兵教授是 Web 挖掘研究领域的国际知名专家，曾担任多个国际期刊的编辑，也是多个国际学术会议（如 WWW, KDD 与 AAAI 等）的程序委员会委员。刘兵教授在 Web 内容挖掘、互联网观点挖掘、数据挖掘等领域有非常高的造诣。他先后在国际著名学术期刊与重要国际学术会议上发表论文一百多篇。本教材中的部分章节也融入了刘兵教授从事 Web 挖掘研究多年的心血。

全书主要包括前言与 12 章节。本书的翻译和审校由俞勇、薛贵荣和韩定一共同完成。其中，俞勇负责前言、第 1 至 2 章，薛贵荣负责第 3 至 7 章，韩定一负责第 8 至 12 章。参加翻译工作的还有韩定一（前言、第 1、8 章）、徐生良（第 2 章）、凌霄（第 3 章）、郭晋文（第 4、5 章）、王亮（第 6 章）、陈林虎（第 7 章）、傅临云（第 9 章）、第 7 张迪（第 10 章）、包胜华（第 11 章）和王乐天（第 12 章）。上海交通大学 APEX 数据和知识管理实验室的全体同学参加了本书的校对工作。

在本书的翻译过程中，得到了刘兵教授的大力支持。他向译者提供了全文书稿的最终版本，并对翻译工作提出了指导性建议。同时，感谢微软亚洲研究院李航博士的引荐，使我们有机会学习和翻译此书。最后，感谢清华大学出版社的编辑们，是他们使得本书能够尽快与读者见面。

由于本书所涉及内容非常广泛，许多术语目前尚无固定译法，翻译难度相对较大。尽管我们对某些术语进行了推敲，但仍然可能出现词不达意的地方。此外，由于译者水平有限，译文中不当之处也在所难免。我们也真诚地希望同行与读者朋友们不吝赐教。如果您能将您的意见与建议发往 yyu@apex.sjtu.edu.cn，我们将不胜感激。

译者

2009 年 3 月

第 2 版译者序

伊利诺伊大学芝加哥分校（UIC）的刘兵（Bing Liu）教授历所著的“Web Data Mining”是少有的专门针对 Web 挖掘领域的教材与专著，笔者曾有幸学习和翻译此书的第 1 版。在过去的几年里，Web 挖掘领域取得了许多重大的进展，为了反应这些进展，刘兵教授出版了“Web Data Mining”第 2 版。笔者很荣幸能够受到刘兵教授认可，继续承担本书第 2 版的翻译工作。

相比于第一版而言，本书的第 2 版根据目前的研究前沿对于已有的章节进行了内容的增加和修正。特别的，本书的 11、12 章加入了社会网络分析、推荐系统、协同过滤等方向的进展。本书的英文第 2 版也从原来的 532 页变成了 622 页。读者能够通过本书了解到 Web 挖掘领域的基础知识以及最新动向。

第 2 版的翻译是在第一版翻译的基础上翻译审校完成的。第 2 版的翻译和审校由俞勇、薛贵荣、韩定一和陈天奇共同完成。参加翻译工作的还有陈相如（第 1、2、6 章）、潘俊峰（第 3、4、5 章）、孙辛若（第 7、8、9 章）、牛星（第 10 章）、陈凯龙和殷力昂（前言、第 11 章）、张伟楠和陆秋霞（第 12 章）。戎术、陈柏良、潘晔等参加了本书的校对工作。感谢上海交通大学 APEX 数据和知识管理实验室的全体同学对本书的翻译及校对工作的大力支持。

由于译者水平有限，译文中不当之处也在所难免。我们也真诚地希望同行与读者朋友们不吝赐教。如果您能将您的意见与建议发往 yyu@apex.sjtu.edu.cn，我们将不胜感激。

译者

2012 年 12 月

序　言

在过去的 20 年里，Web 的迅速发展使其成为世界上规模最大的公共数据源。Web 挖掘的目标是从 Web 超链接、网页内容和使用日志中探寻有用的信息。依据在挖掘过程中使用的数据类别，Web 挖掘任务可以被划分为 3 种主要类型：Web 结构挖掘、Web 内容挖掘和 Web 使用挖掘。Web 结构挖掘从表征 Web 结构的超链接中寻找知识。Web 内容挖掘从网页内容中抽取有用的信息和知识。而 Web 使用挖掘则从使用日志和其他形式的用户交互记录中挖掘用户的活动模式。从本书在 2006 年底的第 1 版发行之后，很多领域已经有了重大的进展。大部分的章节都已经添加了新的材料来反应这些进展。主要的改动在第 11 章和第 12 章中，这两章已经被重新撰写并做了重要的扩展。在撰写第 1 章的时候，观点挖掘（第 11 章）的研究仍处于初步阶段。从那以后，搜索社区对这个问题已经拥有了一个更好的理解并提出了许多新颖的技术来解决问题的各个方面。为了将 Web 使用挖掘（第 12 章）的最新进展包含进来，关于推荐系统、协同过滤、用户日志挖掘和计算广告学的话题已经被添加进来。新版比原来长了很多。

本书旨在讲述上述的互联网数据挖掘任务以及它们的核心挖掘算法；尽可能涵盖每个话题的广泛内容，给出足够多的细节，以便读者无须借助额外的阅读，即可获得相对完整的关于算法和技术的知识。其中第 5 章——监督学习的部分内容、结构化数据的抽取、信息整合、观点挖掘和 Web 使用挖掘——是本书的特色，这些内容在其他书籍中没有提及，但它们在 Web 数据挖掘中却占有非常重要的地位。当然，传统的 Web 挖掘主题，如搜索、页面爬取和资源探索以及链接分析在书中也做了详细描述。

本书尽管题为“Web 数据挖掘”，但依然涵盖了数据挖掘和信息检索的核心主题；因为 Web 挖掘大量使用了它们的算法和技术。数据挖掘部分主要由关联规则和序列模式、监督学习（分类）、无监督学习（聚类）这三大重要的数据挖掘任务，和半监督学习这个相对深入的主题组成。而信息检索对于 Web 挖掘而言最重要的核心主题都有所阐述。因此，本书自然的分为两大部分，第 1 部分包括第 2～5 章，介绍数据挖掘的基础，第 2 部分包括第 6～12 章，介绍 Web 相关的挖掘任务。

有两大指导性原则贯穿本书始末。其一，本书的基础内容适合本科生阅读，但也包括足够多的深度资料，以满足打算在 Web 数据挖掘和相关领域研读博士学位的研究生。书中对读者的预备知识几乎没有作任何要求，任何对算法和概率知识稍有理解的人都应当能够顺利地读完本书。其二，本书从实践的角度来审视 Web 挖掘的技术。这一点非常重要，因为大多数 Web 挖掘任务都在现实世界中有所应用。在过去的几年中，我有幸直接或间接地与许多研究人员和工程人员一起工作，他们来自于多个搜索引擎、电子商务公司，甚至是对在业务中利用 Web 信息感兴趣的传统公司。在这个过程中，我获得了许多现实世界问题的实践经历和第一手知识。我尽量将其中非机密的信息和知识通过本书传递给读者，因此本书能在理论和实践中有所平衡。我希望本书不仅能够成为学生的教科书，也能成为 Web

挖掘研究人员和实践人员获取知识、信息、甚至是创新想法的一个有效渠道。

致　谢

在撰写本书的过程中，许多研究人员都给予我无私的帮助；没有他们的帮助，这本书也许永远也无法成为现实。我最深切的感谢要给予 Filippo Menczer、Bamshad Mobasher 和 Olfa Nasraoui，他们热情地撰写了本书中重要的两个章节。他们也是相关领域的专家。Filippo 负责 Web 爬取的整一章，Bamshad 和 Olfa 负责 Web 使用挖掘这一章的所有片段，除了推荐系统那一节，但是他们也提供了帮助。我还要感谢 Wee Sun Lee（李伟上），他帮助完成第 5 章的很大一部分。

Jian Pei（裴健）帮助撰写了第 2 章中 PrefixSpan 算法，并且检查了 MS-PS 算法。Eduard Dragut 帮助撰写了第 10 章的最后一节，并且多次阅读并修改这一整章。Yuanlin Zhang 对第 9 章提出很多意见。Simon Funk、Yehuda Koren、Wee Sun Lee、Jing Peng、Arkadiusz Paterek 和 Domonkos Tikk 对第 12 章中的推荐系统的撰写提供了帮助。我对他们所有人都有所亏欠。

还有许多研究人员以各种方式提供了帮助。Yang Dai（戴阳）和 Rudy Setiono 在支持向量机（SVM）上提供帮助。Chris Ding（丁宏强）帮助社交网络分析。Clement Yu（于德）和 ChengXiang Zhai（翟成祥）阅读了第 6 章。Amy Langville 阅读了第 7 章。Kevin C.-C. Chang（张振川）、Ji-Rong Wen（文继荣）和 Clement Yu（于德）帮助了第 10 章的许多方面。Justin Zobel 帮助理清了索引压缩的许多议题。Ion Muslea 帮助理清了包裹简介的一些议题。Divy Agrawal、Yunbo Cao（曹云波）、Edward Fox、Hang Li（李航）、Xiaoli Li（李晓黎）、Zhaohui Tan、Dell Zhang（张德）和 Zijian Zheng 帮助检查了各个章节。在此对他们表示感谢！

和许多研究人员的讨论也帮助本书成形。这些人包括 Amir Ashkenazi、Imran Aziz、Roberto Bayardo、Shenghua Bao（包胜华）、Roberto Bayardo、Wendell Baker、Ling Bao、Jeffrey Benkler、Brian Davison、AnHai Doan、Byron Dom、Juliana Freire、Michael Gamon、Robert Grossman、Natalie Glance、Jiawei Han（韩家炜）、Meichun Hsu、Wynne Hsu、Ronny Kohavi、Birgit König、David D. Lewis、Ian McAllister、Wei-Ying Ma（马维英）、Marco Maggini、Llew Mason、Kamel Nigan、Julian Qian、Yan Qu、Thomas M. Tirpak、Andrew Tomkins、Alexander Tuzhilin、Weimin Xiao、Gu Xu（徐谷）、Philip S. Yu 和 Mohammed Zaki、Yuri Zelenkov 和 Daniel Zeng。

我已毕业和在读的学生们 Gao Cong、Xiaowen Ding、Murthy Ga-napathibhotla、Minqing Hu、Nitin Jindal、Xin Li、Yiming Ma、Arjun Muk-herjee、Quang Qiu（浙江大学的访问学生）、William Underwood、Yanhong Zhai、Zhongwu Zhai（清华大学的访问学生）、Lei Zhang 和 Kaidi Zhao 这些年来贡献了非常多的研究思路，而且还检查了很多算法并作出了许多更正。书中的大部分章节已经用在芝加哥大学我的研究生课程里。我感谢那些在客上实现了一些算法的学生。他们的问题帮助我提升并在某些情况下更正了算法。在这里列出他们所有人的名字不太可能。这里，我特别想感谢 John Castano、Hari Prasad Divyakotti、Islam Ismailov、Suhyuk Park、Cynthia Kersey、Po-Hsiu Lin、Srikanth Tadikonda、Makio Tamura、Ravikanth Turlapati、Guillermo Vazquez、Haisheng Wang 和 Chad Williams 指出了文字、例

子或算法的错误。德保尔大学的 Michael Bombyk 也找到了几个打字错误。

与 Springer 出版社的员工一起工作是一段令人愉快的经历。我感谢编辑 Ralf Gerstner 在 2005 年初征询我对撰写一本有关 Web 挖掘的书籍是否感兴趣。从那以后，我们一直保持着愉快的合作经历。我还要感谢校对 Mike Nugent 提高了本书内容的表达质量，以及制作编辑 Michael Reinfarth 引导我顺利完成了本书的出版过程。还有两位匿名评审也给出不少有见解的评论。伊利诺伊斯大学芝加哥分校计算机科学系对本项目提供了计算资源和工作环境的支持。

最后，我要感谢我的父母和兄弟姐妹，他们给予我一贯的支持和鼓励。我将最深刻的感激给予我自己的家庭成员：Yue、Shelley 和 Kate。他们也在许多方面给予支持和帮助。尽管 Shelley 和 Kate 还年幼，但他们阅读了本书的绝大部分，并且找出了不少笔误。我的妻子将家里一切事情打理地秩序井然，使我可以将充分的时间和精力花费在这本书上。谨以此书献给他们！

Bing Liu（刘兵）

目　录

第 1 部分　数据挖掘基础

第 2 部分　Web 挖掘

第1章 概 述

当你捧起这本书的时候，毫无疑问，你已经知道什么是**万维网**（World Wide Web，WWW）并且在广泛使用它了。万维网影响着我们日常生活中的方方面面。它是目前最广为人知的一种信息源，也是规模最大的信息源。它能很方便地被访问和搜索到。在这张网中，包括了几十亿份相互连接的文档（称为**网页**（Web page）），而这些文档的作者就是数以百万计的万维网用户自己。自从万维网诞生以来，人们获取信息的行为方式已经发生了巨大变化。在这之前，我们通常会通过朋友和专家，或者通过购买和借阅书籍来获取信息。但是，万维网出现以后，一切就变成只需在家中或者办公室里轻点几下鼠标那么简单。我们不仅可以从网上找到需要的信息，更能便捷地发布自己的信息和知识以供他人阅读。

万维网还成为进行商业交易的重要渠道。我们可以从在线商店（虚拟商店）购买到几乎所有商品，因此便无须大费周折跑去实体商店。万维网还提供了一个方便的交流渠道，我们可以通过它和世界上任何一个角落的人进行沟通，并且表达自己对任何事件的观点。万维网已成为一个名副其实的**虚拟社会**（Virtual Society）。在这第一个章节中，我们会介绍万维网、它的历史以及将在本书中探讨的各个主题。

1.1 什么是万维网

万维网的官方定义是：一个在世界范围内的超媒体信息获取平台，它提供一种获取海量文档的统一途径。简单来说，万维网就是基于互联网的计算机网络，它允许一台计算机的用户访问存放在另外一台机器上的信息，而这些机器通过一个名为**互联网**（Internet）的全球网络保持互相连通。互联网的实现完全依照标准**客户端-服务器**（Client-Server）模型。在这个模型中，一名用户依赖于被称为**客户端**的程序连接到存放数据的远程机器上，这台机器被称为**服务器**。在万维网上冲浪主要通过名为**浏览器**（Browser）的客户端程序，如Netscape、Internet Explorer、Firefox、Chrome 等。这些浏览器首先向远程服务器发送请求，然后将传回的使用 HTML 编写的文本和图片内容展现在客户端的计算机屏幕上。

互联网的运作依赖于**超文本**（Hypertext）文档的结构。超文本允许网页作者创建指向世界上任何计算机上相关文档的链接。用户只需简单点击这些链接（称为**超链接**（Hyperlinks））便可访问这些相关文档。超文本的想法由 Ted Nelson 在 1965 年首先提出[14]，他创造了著名的超文本系统 Xanadu（http://xanadu.com/）。超文本还可以包含其他媒体（如图片、声音、视频文件），这些媒体被统称为**超媒体**（Hypermedia）。

1.2 万维网和互联网的历史简述

万维网的创立：万维网最初是由 Tim Berners-Lee 于 1989 年发明的。当时，他在位于瑞士的欧洲粒子物理实验室（Centre European pour la Recherche Nucleaire，或 European Laboratory for Particle Physics，简称 CERN）工作。他给万维网命名，并且编写了世界上首个万维网服务器——httpd 和世界上首个客户端程序“**WorldWideWeb**”（包括一个浏览器和一个编辑器）。

事件起源于 1989 年 3 月，当时 Tim Berners-Lee 向他在 CERN 的导师提交了一份名为“信息管理提议”的提议书。在这份提议中，他讨论了层次化信息组织的缺点，并且描绘出基于超文本系统的优点。提议书建议设计一套简单的协议，使得用户可以通过网络请求存放在远端系统上的信息；并创立一套使信息可以用相同格式被互相交换，并且用户可以通过超链接把相关文档链接起来的机制。其中还提到如何使用当时在 CERN 的一些文本阅读和图形显示的技术。提议书完整地描述了**分布式超文本系统**（Distributed Hypertext System），也就是当今万维网的基础构架。

起初，这份提议书并没有获得足够的支持。然而，在 1990 年，Berners-Lee 重新分发了提议书，并获得了足够的支持来展开工作。在这个项目中，Berners-Lee 和他在 CERN 的团队为最终把万维网发展成为分布式超文本系统铺平了道路。他们设计了服务器、浏览器、用于在客户端和服务器之间进行通信的协议——**超本文传输协议**（HyperText Transfer Protocol，HTTP）、用于编辑网络文档的**超文本标记语言**（HyperText Markup Language，HTML），以及**统一资源定位符**（Universal Resource Locator，URL）。万维网从此开始迅速发展。

Mosaic 和 Netscape：下一个万维网的重要事件是 **Mosaic** 的出现。1993 年 2 月，来自美国伊利诺伊斯大学国家超级计算应用中心（National Center for Supercomputing Applications, NCSA）的 Marc Andreessen 和他的团队发布了 UNIX 操作系统上图形界面的网络浏览器——Mosaic for X。几个月以后，面向 Macintosh 和 Windows 操作系统的不同版本 Mosaic 相继发布。这是一个相当重要的事件，历史上第一次出现了一个具有一致而简单点击的图形用户界面的网络客户端，并且同时在当时三大最流行的操作系统上均有实现。此事迅速在学术圈外引起轩然大波。1994 年中期，Silicon Graphics（SGI）公司的创立人 Jim Clark 和 Marc Andreessen 展开合作，成立了 **Mosaic** 通信公司（后来改名为**网景通信**（Netscape Communications））。在短短几个月后，**Netscape** 浏览器就被发布了，万维网的爆炸式发展也由此开始。1995 年 8 月，微软公司的 **Internet Explorer** 也进入了市场，开始挑战 Netscape。

Tim Berners-Lee 创建万维网和之后 Mosaic 浏览器的发布通常都被认为是万维网的流行和成功的两大最重要事件。

互联网[1]：如果没有互联网，那么万维网就无法成为现实。因为互联网提供了万维网运

1 在原书中，作者将互联网（Internet）视为万维网（World Wide Web）的底层；互联网提供底层通讯机制，万维网则是构架在互联网之上的各种应用。因此本书严格区分互联网和万维网——译者注。