

 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

Translation from English language edition:
Web Data Mining
by Bing Liu
Copyright © 2011, Springer Berlin Heidelberg
Springer Berlin Heidelberg is a part of Springer Science+Business Media
All Rights Reserved

本书为英文版 *Web Data Mining* 的简体中文翻译版, 作者 Bing Liu, 由 Springer 出版社授权清华大学出版社出版发行。

北京市版权局著作权合同登记号 图字: 01-2012-0339 号

本书封面贴有清华大学出版社防伪标签, 无标签者不得销售。

版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

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 Mukherjee、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 章 概述	1	1.3.2 什么是 Web 数据挖掘	5
1.1 什么是万维网	1	1.4 各章概要	6
1.2 万维网和互联网的历史简述	2	1.5 如何阅读本书	8
1.3 Web 数据挖掘	3	文献评注	9
1.3.1 什么是数据挖掘	4	参考文献	9

第 1 部分 数据挖掘基础

第 2 章 关联规则和序列模式	13	2.9 从序列模式中产生规则	42
2.1 关联规则的基本概念	13	2.9.1 序列规则	42
2.2 Apriori 算法	15	2.9.2 标签序列规则	42
2.2.1 频繁项目集生成	15	2.9.3 分类序列规则	43
2.2.2 关联规则生成	18	文献评注	43
2.3 关联规则挖掘的数据格式	20	参考文献	45
2.4 多最小支持度的关联规则挖掘	21	第 3 章 监督学习	49
2.4.1 扩展模型	22	3.1 基本概念	49
2.4.2 挖掘算法	23	3.2 决策树归纳	52
2.4.3 规则生成	27	3.2.1 学习算法	53
2.5 分类关联规则挖掘	28	3.2.2 混杂度函数	54
2.5.1 问题描述	28	3.2.3 处理连续属性	57
2.5.2 挖掘算法	29	3.2.4 其他一些问题	58
2.5.3 多最小支持度分类关联规则挖掘	31	3.3 评估分类器	60
2.6 序列模式的基本概念	32	3.3.1 评估方法	61
2.7 基于 GSP 挖掘序列模式	34	3.3.2 查准率、查全率、F-score 和平衡点 (Breakeven Point)	62
2.7.1 GSP 算法	34	3.3.3 受试者工作特征曲线	63
2.7.2 多最小支持度挖掘	35	3.3.4 提升曲线	65
2.8 基于 PrefixSpan 算法的序列模式挖掘	38	3.4 规则归纳	66
2.8.1 PrefixSpan 算法	39	3.4.1 顺序化覆盖	66
2.8.2 多最小支持度挖掘	40	3.4.2 规则学习: Learn-	

One-Rule 函数	68	4.4 层次聚类	107
3.4.3 讨论	70	4.4.1 单连结方法	108
3.5 基于关联规则的分类	71	4.4.2 全连结方法	108
3.5.1 使用类关联规则 进行分类	71	4.4.3 平均连结方法	109
3.5.2 使用类关联规则 作为分类属性	74	4.4.4 优势和劣势	109
3.5.3 使用古典的关联 规则分类	74	4.5 距离函数	110
3.6 朴素贝叶斯分类	75	4.5.1 数字属性	110
3.7 朴素贝叶斯文本分类	78	4.5.2 布尔属性和名词性 属性	110
3.7.1 概率框架	78	4.5.3 文本文档	112
3.7.2 朴素贝叶斯模型	79	4.6 数据标准化	112
3.7.3 讨论	81	4.7 混合属性的处理	114
3.8 支持向量机	81	4.8 采用哪种聚类算法	115
3.8.1 线性支持向量机: 可分的情况	82	4.9 聚类的评估	115
3.8.2 线性支持向量机: 数据不可分的情况	86	4.10 发现数据区域和 数据空洞	118
3.8.3 非线性支持向量机: 核方法	88	文献评注	119
总结	90	参考文献	121
3.9 k -近邻学习	91	第 5 章 部分监督学习	124
3.10 分类器的集成	92	5.1 从已标注数据和无标注 数据中学习	124
3.10.1 Bagging	92	5.1.1 使用朴素贝叶斯 分类器的 EM 算法	125
3.10.2 Boosting	92	5.1.2 Co-Training	128
文献评注	93	5.1.3 自学习	129
参考文献	94	5.1.4 直推式支持向量机	130
第 4 章 无监督学习	98	5.1.5 基于图的方法	131
4.1 基本概念	98	5.1.6 讨论	133
4.2 k -均值聚类	100	5.2 从正例和无标注数据中 学习	133
4.2.1 k -均值算法	100	5.2.1 PU 学习的应用	134
4.2.2 k -均值算法的 硬盘版本	102	5.2.2 理论基础	135
4.2.3 优势和劣势	102	5.2.3 建立分类器: 两步方法	137
4.3 聚类的表示	105	5.2.4 建立分类器: 偏置 SVM	142
4.3.1 聚类的一般表示 方法	106	5.2.5 建立分类器: 概率估计	144
4.3.2 任意形状的聚类	106	5.2.6 讨论	145

附录：朴素贝叶斯 EM 算法 的推导	145
-----------------------------	-----

文献评注	147
参考文献	148

第 2 部分 Web 挖掘

第 6 章 信息检索与 Web 搜索	153
6.1 信息检索中的基本概念	154
6.2 信息检索模型	156
6.2.1 布尔模型	156
6.2.2 向量空间模型	157
6.2.3 统计语言模型	159
6.3 关联性反馈	160
6.4 评估标准	162
6.5 文本和网页的预处理	164
6.5.1 无用词移除	165
6.5.2 词干提取	165
6.5.3 其他文本预处理 步骤	165
6.5.4 网页预处理步骤	166
6.5.5 副本探测	167
6.6 倒排索引及其压缩	168
6.6.1 倒排索引	168
6.6.2 使用倒排索引搜索	169
6.6.3 索引的建立	170
6.6.4 索引的压缩	171
6.7 隐式语义索引	175
6.7.1 奇异值分解 (singular value decomposition)	176
6.7.2 查询和检索	177
6.7.3 实例	178
6.7.4 讨论	181
6.8 Web 搜索	181
6.9 元搜索引擎和组合多种 排序	183
6.9.1 使用相似度分数的 合并	184
6.9.2 使用排名位置的 合并	184

6.10 网络作弊	186
6.10.1 内容作弊	187
6.10.2 链接作弊	187
6.10.3 隐藏技术	188
6.10.4 抵制作弊	189
文献评注	190
参考文献	191
第 7 章 社会网络分析	195
7.1 社会网络分析	196
7.1.1 中心性	196
7.1.2 权威	198
7.2 同引分析和引文耦合	199
7.2.1 同引分析	200
7.2.2 引文耦合	200
7.3 PageRank	201
7.3.1 PageRank 算法	201
7.3.2 PageRank 算法的 优点和缺点	207
7.3.3 Timed PageRank 和 Recency Search	207
7.4 HITS	208
7.4.1 HITS 算法	209
7.4.2 寻找其他的特征 向量	211
7.4.3 同引分析和引文 耦合的关系	211
7.4.4 HITS 算法的优点 和缺点	212
7.5 社区发现	213
7.5.1 问题定义	213
7.5.2 二分核心社区	215
7.5.3 最大流社区	216
7.5.4 基于中介性的 电子邮件社区	218

7.5.5 命名实体的重叠社区	219	标记编码	265
文献评注	220	9.2 包装器归纳	266
参考文献	220	9.2.1 从一张网页抽取	267
第 8 章 Web 爬取	225	9.2.2 学习抽取规则	269
8.1 一个简单爬虫算法	225	9.2.3 识别提供信息的样例	272
8.1.1 宽度优先爬虫	227	9.2.4 包装器维护	273
8.1.2 带偏好的爬虫	227	9.3 基于实例的包装器学习	273
8.2 实现议题	228	9.4 自动包装器生成中的一些问题	276
8.2.1 网页获取	228	9.4.1 两个抽取问题	276
8.2.2 网页解析	228	9.4.2 作为正则表达式的模式	277
8.2.3 删除无用词并提取词干	230	9.5 字符串匹配和树匹配	277
8.2.4 链接提取和规范化	230	9.5.1 字符串编辑距离	278
8.2.5 爬虫陷阱	232	9.5.2 树匹配	279
8.2.6 网页库	232	9.6 多重对齐	282
8.2.7 并发性	233	9.6.1 中星方法	283
8.3 通用爬虫	234	9.6.2 部分树对齐	284
8.3.1 可扩展性	234	9.7 构建 DOM 树	287
8.3.2 覆盖度、新鲜度和重要度	235	9.8 基于列表页的抽取: 平坦数据记录	288
8.4 限定爬虫	236	9.8.1 有关数据记录的两个观察结果	289
8.5 主题爬虫	238	9.8.2 挖掘数据区域	290
8.5.1 主题本地性和线索	240	9.8.3 从数据区域中识别数据记录	294
8.5.2 最优优先变种	243	9.8.4 数据项对齐与抽取	294
8.5.3 自适应	246	9.8.5 利用视觉信息	295
8.6 评价标准	249	9.8.6 一些其他技术	295
8.7 爬虫道德和冲突	253	9.9 基于列表页的抽取: 嵌套数据记录	296
8.8 最新进展	255	9.10 基于多张网页的抽取	301
文献评注	256	9.10.1 采用前几节中的技术	301
参考文献	257	9.10.2 RoadRunner 算法	301
第 9 章 结构化数据抽取: 包装器生成	261	9.11 一些其他问题	303
9.1 预备知识	261	9.11.1 从其他网页中	
9.1.1 两种富含数据的网页	262		
9.1.2 数据模型	263		
9.1.3 数据实例的 HTML			

抽取	303	10.9.2 词汇恰当	330
9.11.2 析取还是可选	303	10.9.3 实例恰当	331
9.11.3 集合类型还是 元组类型	304	文献评注	331
9.11.4 标注与整合	304	参考文献	331
9.11.5 领域相关的 抽取	305	第 11 章 观点挖掘与情感分析	335
9.12 讨论	305	11.1 观点挖掘问题	335
文献评注	305	11.1.1 问题定义	336
参考文献	306	11.1.2 基于方面的观点 摘要	340
第 10 章 信息集成	310	11.2 文本情感分类	341
10.1 什么是模式匹配	310	11.2.1 基于监督学习的 分类	342
10.2 模式匹配的预处理 工作	312	11.2.2 基于无监督学习的 分类	343
10.3 模式层的匹配	313	11.3 句子主观性与情感分类	345
10.3.1 基于语言学的 算法	313	11.4 观点词汇扩展	347
10.3.2 基于模式约束的 算法	314	11.5 基于方面的观点挖掘	349
10.4 基于域和实例层的 匹配	315	11.5.1 基于方面的 情感分类	349
10.5 综合多种相似度	317	11.5.2 观点的基本规则	351
10.6 1:m 匹配	317	11.5.3 方面抽取	353
10.7 一些其他问题	318	11.5.4 同时扩展观点 词汇和抽取方面	355
10.7.1 重用已有的匹配 结果	318	11.6 比较性观点挖掘	358
10.7.2 大量模式的匹配	319	11.6.1 问题定义	358
10.7.3 模式匹配的结果	319	11.6.2 等级比较性语句的 识别	360
10.7.4 用户交互	320	11.6.3 偏好实体识别	360
10.8 Web 查询界面的集成	320	11.7 其他的一些问题	362
10.8.1 一个基于聚类的 方法	322	11.8 观点搜索	365
10.8.2 基于相互关系的 方法	324	11.9 观点欺诈检测	367
10.8.3 基于实例的方法	326	11.9.1 观点欺诈的目标和 行为	367
10.9 构建一个统一的全局 查询界面	328	11.9.2 隐藏技巧	368
10.9.1 结构恰当和 合并算法	328	11.9.3 基于监督学习的 欺诈检测	369
		11.9.4 基于异常行为的 欺诈检测	370
		11.9.5 群组欺诈检测	372

11.10 评论的效用	372	12.4.2 基于内容的推荐	403
文献评注	373	12.4.3 协同过滤： k -近邻 (kNN)	404
参考文献	374	12.4.4 协同过滤：使用 关联规则	406
第 12 章 Web 使用挖掘	384	12.4.5 协同过滤：矩阵 分解	408
12.1 数据收集和预处理	385	12.5 查询日志挖掘	412
12.1.1 数据的来源和 类型	385	12.5.1 数据源、特征和 挑战	413
12.1.2 Web 使用记录数据 预处理的关键元素 ..	388	12.5.2 查询日志数据 准备	414
12.2 Web 使用挖掘的数据 建模	392	12.5.3 查询日志数据 模型	416
12.3 Web 使用模式的发现和 分析	395	12.5.4 查询日志特征 提取	419
12.3.1 会话和访问者 分析	395	12.5.5 查询日志挖掘 应用	419
12.3.2 聚类分析和访问者 分割	396	12.5.6 查询日志挖掘 方法	421
12.3.3 关联及相关度 分析	399	12.6 计算广告学	423
12.3.4 序列和导航模式 分析	399	12.7 讨论和展望	426
12.3.5 基于 Web 用户事务 的分类和预测	402	文献评注	426
12.4 推荐系统和协同过滤	402	参考文献	427
12.4.1 推荐问题	402		

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。