世界著名计算机教材精选

Web数据挖掘

Bing Liu 著

俞勇 薛贵荣 韩定一 译

WEB DATA MINING

清华大学出版社

世界著名计算机教材精选

Web Data Mining

Web 数据挖掘

Bing Liu 著
俞 勇
薛贵荣 译
韩定一

清华大学出版社
北 京

图书在版编目（CIP）数据

Web 数据挖掘/(美)刘兵(Liu, B.)著;俞勇等译. —北京：清华大学出版社，2009.4
(世界著名计算机教材精选)
书名原文：Web Data Mining
ISBN 978-7-302-19338-8

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

中国版本图书馆 CIP 数据核字(2009)第 010593 号

责任编辑：龙啟铭
责任校对：徐俊伟
责任印制：李红英

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**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

印 刷 者：北京密云胶印厂
装 订 者：三河市兴旺装订有限公司
经　　销：全国新华书店
开　　本：185×260　　**印　　张**：24.5　　**字　　数**：594 千字
版　　次：2009 年 4 月第 1 版　　**印　　次**：2009 年 4 月第 1 次印刷
印　　数：1～3000
定　　价：49.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。
联系电话：010-62770177 转 3103　　产品编号：027154-01

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，它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 数据挖掘便应运而生，并且伴随 Web 的发展而备受关注。Web 数据挖掘它建立在信息检索、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。反过来，可以说，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 数据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。

全书主要包括前言和 12 个章节。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 章和第 2 章，薛贵荣负责第 3～7 章，韩定一负责第 8～12 章。参加翻译工作的还有韩定一（前言、第 1 章、第 8 章）、徐生良（第 2 章）、凌霄（第 3 章）、郭晋文（第 4 章、第 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)等。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龙启铭编辑,是他的远见,使得本书能够尽快与读者见面。

由于本书所涉及到内容非常广泛,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

序言

过去几十年里，Web 的迅速发展使其成为世界上规模最大的公共数据源。Web 数据挖掘的目标是从 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 挖掘的任务可以被划分为三种主要类型：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 使用挖掘则从记录每位用户点击情况的使用日志中挖掘用户的访问模式。

本书旨在讲述这些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。其中结构化数据的抽取、信息整合、观点挖掘和 Web 使用挖掘等 4 章是本书的特色，这些内容在已有书籍中没有提及，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作了详细描述。

本书尽管题为“Web 数据挖掘”，却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）这三大最重要的数据挖掘任务，以及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。第一部分，包括第 2～5 章，介绍数据挖掘的基础。第二部分，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，这本书也许永远无法成为现实。我最深切的感谢要给予 Filippo Menczer 和 Bamshad Mobasher，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 爬取这一章，Bamshad 负责 Web 使用挖掘这一章。我还要感谢 Wee Sun Lee(李伟上)，他帮助完成第 5 章半监督学习的很大一部分。

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。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)对链接分析提供了帮助。Clement Yu(余德)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(余德)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的成形。这些人包括 Amir Ashkenazi、Imran Aziz、Roberto Bayardo、Wendell Baker、Ling Bao、Jeffrey Benkler、AnHai Doan、Byron Dom、Michael Gamon、Robert Grossman、Jiawei Han(韩家炜)、Wynne Hsu、Ronny Kohavi、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。

我的学生们(不论已毕业或是在读)检查了许多算法的正确性并且作出了许多修正。他们包括 Gao Cong(从高)、Minqing Hu、Nitin Jindal、Xin Li、Yiming Ma、Yanhong Zhai 和 Kaidi Zhao。本书中一些章节是我在伊利诺伊斯大学芝加哥分校的研究生课程讲义。我要感谢这些课程的学生帮我实现了一部分算法。他们提出的问题在某些情况下也帮助我修正算法。在这里我不可能完全列出他们的名字，但我要特别感谢 John Castano、Xiaowen Ding、Murthy Ganapathibhotla、Cynthia Kersey、Hari Prasad Divyakotti、Ravikanth Turlapati、Srikanth Tadikonda、Makio Tamura、Haisheng Wang 和 Chad Williams，他们指出讲义中文本、举例或算法的错误。来自德保罗大学的 Michael Bombyk 也指出了不少笔误。

与 Springer 出版社的员工一起工作是一段令人愉快的经历。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得秩序井然，使我可以将充分的时间和精力用在这本书上。谨以此书献给他们！

Bing Liu(刘兵)

目录

第一部分　数据挖掘基础

第二部分 Web 挖掘

第一部分

数据挖掘基础