<<Web数据挖掘>>

图书基本信息

书名：<<Web数据挖掘>>

13位ISBN编号：9787302193388

10位ISBN编号：730219338X

出版时间：2009-4

出版时间：清华大学出版社

作者：刘兵

页数：375

字数：594000

译者：俞勇

前言

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。

<<Web数据挖掘>>

内容概要

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。

作者简介

刘兵（Bing Liu），伊利诺伊大学芝加哥分校（tnc）教授，他在爱丁堡大学获得人工智能博士学位。刘兵教授是Web挖掘研究领域的国际知名专家，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（如KDD、www、AAAl

<<Web数据挖掘>>

书籍目录

章节摘录

插图：第一部分 数据挖掘基础第1章 概述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，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（包括一个浏览器和一个编辑器World Wide Web）。
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讯的协议——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 Andreesen和他的团队发布了UNIX操作系统上图形界面的 网络浏览器一Mosaic for X。

编辑推荐

《Web数据挖掘》为世界著名计算机教材精选之一。

版权说明