

<<Web数据挖掘>>

图书基本信息

书名：<<Web数据挖掘>>

13位ISBN编号：9787115194046

10位ISBN编号：7115194041

出版时间：2009-2

出版时间：人民邮电出版社

作者：查凯莱巴蒂

页数：344

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## 前言

This book is about finding significant statistical patterns relating hypertext documents, topics, hyperlinks, and queries and using these patterns to connect users to information they seek. The Web has become a vast storehouse of knowledge.

## <<Web数据挖掘>>

### 内容概要

本书是信息检索领域的名著，深入讲解了从大量非结构化Web数据中提取和产生知识的技术。书中首先论述了Web的基础(包括Web信息采集机制、Web标引机制以及基于关键字或基于相似性搜索机制)，然后系统地描述了Web挖掘的基础知识，着重介绍基于超文本的机器学习和数据挖掘方法，如聚类、协同过滤、监督学习、半监督学习，最后讲述了这些基本原理在Web挖掘中的应用。本书为读者提供了坚实的技术背景和最新的知识。

本书是从事数据挖掘学术研究和开发的专业人员理想的参考书，同时也适合作为高等院校计算机及相关专业研究生的教材。

## 作者简介

Soumen Chakrabarti, Web搜索与挖掘领域的知名专家, ACM Transactions on the Web副主编。加州大学伯克利分校博士, 目前是印度理工学院计算机科学与工程系副教授。曾经供职于IBM Almaden研究中心, 从事超文本数据库和数据挖掘方面的工作。他有丰富的实际项目开发经验, 开发了多个Web挖掘系统, 并获得了多项美国专利。

<<Web数据挖掘>>

书籍目录

1 INTRODUCTION	1.1 Crawling and Indexing	1.2 Topic Directories	1.3 Clustering and Classification
1.4 Hyperlink Analysis	1.5 Resource Discovery and Vertical Portals	1.6 Structured vs. Unstructured Data Mining	1.7 Bibliographic Notes
PART INFRASTRUCTURE	2 CRAWLING THE WEB	2.1 HTML and HTTP Basics	2.2 Crawling Basics
2.3 Engineering Large-Scale Crawlers	2.3.1 DNS Caching, Prefetching, and Resolution	2.3.2 Multiple Concurrent Fetches	2.3.3 Link Extraction and Normalization
2.3.4 Robot Exclusion	2.3.5 Eliminating Already-Visited URLs	2.3.6 Spider Traps	2.3.7 Avoiding Repeated Expansion of Links on Duplicate Pages
2.3.8 Load Monitor and Manager	2.3.9 Per-Server Work-Queues	2.3.10 Text Repository	2.3.11 Refreshing Crawled Pages
2.4 Putting Together a Crawler	2.4.1 Design of the Core Components	2.4.2 Case Study: Using w3c-libwww	2.5 Bibliographic Notes
3 WEB SEARCH AND INFORMATION RETRIEVAL	3.1 Boolean Queries and the Inverted Index	3.1.1 Stopwords and Stemming	3.1.2 Batch Indexing and Updates
3.1.3 Index Compression Techniques	3.2 Relevance Ranking	3.2.1 Recall and Precision	3.2.2 The Vector-Space Model
3.2.3 Relevance Feedback and Rocchio's Method	3.2.4 Probabilistic Relevance Feedback Models	3.2.5 Advanced Issues	3.3 Similarity Search
3.3.1 Handling à Find-Similar ó Queries	3.3.2 Eliminating Near Duplicates via Shingling	3.3.3 Detecting Locally Similar Subgraphs of the Web	3.4 Bibliographic Notes
PART LEARNING PART	APPLICATIONS	References	Index

章节摘录

插图：

## 媒体关注与评论

本书是Web挖掘与搜索引擎领域的经典著作，自出版以来深受好评，已经被斯坦福、普林斯顿、卡内基梅隆等世界名校采用为教材。

书中首先介绍了Web爬行和搜索等许多基础性的问题，并以此为基础，深入阐述了解决Web挖掘各种难题所涉及的机器学习技术，提出了机器学习在系统获取、存储和分析数据中的许多应用，并探讨了这些应用的优劣和发展前景。

全书分析透彻，富于前瞻性，为构建Web挖掘创新性应用奠定了理论和实践基础，既适用于信息检索和机器学习领域的研究人员和高校师生，也是广大Web开发人员的优秀参考书。

“本书深入揭示了搜索引擎的技术内幕！

有了它，你甚至能够自己开发一个搜索引擎。

” ——searchenginewatch.com网站 “本书系统、全面而且深入，广大Web技术开发人员都能很好地理解和掌握其中内容。

作者是该研究领域的领军人物之一，在超文本信息挖掘和检索方面有着渊博的知识和独到的见解。

” ——Joydeep Ghosh，得克萨斯大学奥斯汀分校教授，IEEE会士 “作者将该领域的所有重要工作融合到这部杰作中，并以一种通俗易懂的方式介绍了原本非常深奥的内容。

有了这本书，Web挖掘终于有可能成为大学的一门课程了。

” ——Jaideep Srivastava，明尼苏达大学教授，IEEE会士

编辑推荐

《Web数据挖掘:超文本数据的知识发现(英文版)》是从事数据挖掘学术研究和开发的专业人员理想的参考书,同时也适合作为高等院校计算机及相关专业研究生的教材。

《Web数据挖掘》是Web挖掘与搜索引擎领域的经典著作,自出版以来深受好评,已经被斯坦福、普林斯顿、卡内基梅隆等世界名校采用为教材。

书中首先介绍了Web爬行和搜索等许多基础性的问题,并以此为基础,深入阐述了解决Web挖掘各种难题所涉及的机器学习技术,提出了机器学习在系统获取、存储和分析数据中的许多应用,并探讨了这些应用的优劣和发展前景。

《Web数据挖掘》分析透彻,富于前瞻性,为构建Web挖掘创新性应用奠定了理论和实践基础,既适用于信息检索和机器学习领域的研究人员和高校师生,也是广大Web开发人员的优秀参考书。

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>