TURING 图灵原版计算机科学系列

ELSEVIER 爱思唯尔

Mining the Web

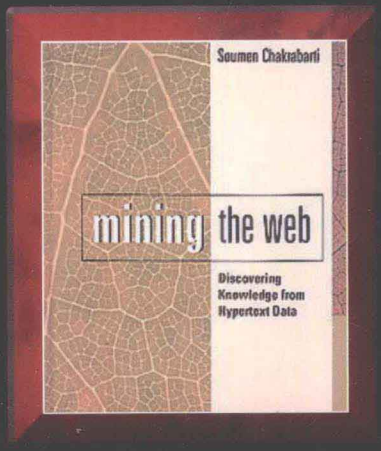
Discovering Knowledge from Hypertext Data

Web数据挖掘

超文本数据的知识发现

（英文版）

[印度] Soumen Chakrabarti 著



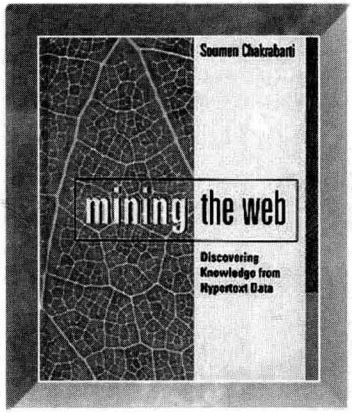
人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵原版计算机科学系列

Mining the Web
Discovering Knowledge from Hypertext Data

Web数据挖掘
超文本数据的知识发现
（英文版）

[印度] Soumen Chakrabarti 著



人民邮电出版社
北　京

图书在版编目（CIP）数据

Web数据挖掘：超文本数据的知识发现：英文/（印）查凯莱巴蒂（Chakrabarti, S.）著. —北京：人民邮电出版社，2009.2

（图灵原版计算机科学系列）

书名原文：Mining the Web: Discovering Knowledge from Hypertext Data

ISBN 978-7-115-19404-6

I. W… II. 查… III. 数据采集－英文 IV. TP311.13

中国版本图书馆CIP数据核字（2008）第198510号

内 容 提 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。

图灵原版计算机科学系列

Web数据挖掘：超文本数据的知识发现（英文版）

◆ 著 [印度] Soumen Chakrabarti

责任编辑 杨海玲

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号

邮编 100061 电子函件 315@ptpress.com.cn

网址 http://www.ptpress.com.cn

北京铭成印刷有限公司印刷

◆ 开本：800×1000 1/16

印张：22.5

字数：432千字 2009年2月第1版

印数：1-2 000册 2009年2月北京第1次印刷

著作权合同登记号 图字：01-2008-5821号

ISBN 978-7-115-19404-6/TP

定价：59.00元

读者服务热线：（010）88593802 印装质量热线：（010）67129223

反盗版热线：（010）67171154

版权声明

ISBN: 978-981-272-321-5

Elsevier (Singapore) Pte Ltd.
3 Killiney Road
#08-01 Winsland House I
Singapore 239519
Tel: (65)6349-0200
Fax: (65)6733-1817

First Published 2009
2009年初版

FOREWORD

Jiawei Han
University of Illinois, Urbana-Champaign

The World Wide Web overwhelms us with immense amounts of widely distributed, interconnected, rich, and dynamic hypertext information. It has profoundly influenced many aspects of our lives, changing the ways we communicate, conduct business, shop, entertain, and so on. However, the abundant information on the Web is not stored in any systematically structured way, a situation which poses great challenges to those seeking to effectively search for high quality information and to uncover the knowledge buried in billions of Web pages. Web mining-or the automatic discovery of interesting and valuable information from the Web-has therefore become an important theme in data mining.

As a prominent researcher on Web mining, Soumen Chakrabarti has presented tutorials and surveys on this exciting topic at many international conferences. Now, after years of dedication, he presents us with this excellent book. *Mining the Web: Discovering Knowledge from Hypertext Data* is the first book solely dedicated to the theme of Web mining and if offers comprehensive coverage and a rigorous treatment. Chakrabarti starts with a thorough introduction to the infrastructure of the Web, including the mechanisms for Web crawling, Web page indexing, and keyword or similarity-based searching of Web contents. He then gives a systematic description of the foundations of Web mining, focusing on hypertext-based machine learning and data mining methods, such as clustering, collaborative filtering, supervised learning, and semi-supervised learning. After that, he presents the application of these fundamental principles to Web mining itself-especially Web linkage analysis-introducing the popular PageRank and HITS algorithms that substantially enhance the quality of keyword-based Web searches.

If you are a researcher, a Web technology developer, or just an interested reader curious about how to explore the endless potential of the Web, you will find this book provides both a solid technical background and state-of-the-art knowledge on this fascinating topic. It is a jewel in the collection of data mining and Web technology books. I hope you enjoy it.

ABOUT THE AUTHOR

Soumen Chakrabarti is assistant professor in Computer Science and Engineering at the Indian Institute of Technology, Bombay. Prior to joining IIT, he worked on hypertext information retrieval and mining at IBM Almaden Research Center. He has developed several systems for Web mining, published extensively, and acquired eight U.S. patents on his inventions to date. Chakrabarti has served as a vice-chair or program committee member for many conferences, including WWW, SIGIR, ICDE, VLDB, KDD, and SODA, and was a guest editor of the IEEE TKDE special issue on mining and searching the Web. His work on focused crawling received the Best Paper award at the 1999 WWW Conference. He holds a Ph.D. from the University of California, Berkeley.

PREFACE

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, built in a decentralized yet collaborative manner. It is a living, growing, populist, and participatory medium of expression with no central editorship. This has positive and negative implications. On the positive side, there is widespread participation in authoring content. Compared to print or broadcast media, the ratio of content creators to the audience is more equitable. On the negative side, the heterogeneity and lack of structure makes it hard to frame queries and satisfy information needs. For many queries posed with the help of words and phrases, there are thousands of apparently relevant responses, but on closer inspection these turn out to be disappointing for all but the simplest queries. Queries involving nouns and noun phrases, where the information need is to find out about the named entity, are the simplest sort of information-hunting tasks. Only sophisticated users succeed with more complex queries—for instance, those that involve articles and prepositions to relate named objects, actions, and agents. If you are a regular seeker and user of Web information, this state of affairs needs no further description.

Detecting and exploiting *statistical dependencies* between terms, Web pages, and hyperlinks will be the central theme in this book. Such dependencies are also called *patterns*, and the act of searching for such patterns is called *machine learning*, or *data mining*. Here are some examples of machine learning for Web applications. Given a crawl of a substantial portion of the Web, we may be interested in constructing a topic directory like Yahoo!, perhaps detecting the emergence and decline of prominent topics with passing time. Once a topic directory is available, we may wish to assign freshly crawled pages and sites to suitable positions in the directory.

In this book, the data that we will "mine" will be very rich, comprising text, hypertext markup, hyperlinks, sites, and topic directories. This distinguishes the area of Web mining as a new and exciting field, although it also borrows liberally from traditional data analysis. As we shall see, useful information on the Web is accompanied by incredible levels of noise, but thankfully, the law of large numbers kicks in often enough that statistical analysis can make sense of the confusion. Our

goal is to provide both the technical background and tools and tricks of the trade of Web content mining, which was developed roughly between 1995 and 2002, although it continues to advance. This book is addressed to those who are, or would like to become, researchers and innovative developers in this area.

Prerequisites and Contents

The contents of this book are targeted at fresh graduate students but are also quite suitable for senior undergraduates. The book is partly based on tutorials at SIGMOD 1999 and KDD 2000, a survey article in SIGKDD *Explorations*, invited lectures at ACL 1999 and ICDT 2001, and teaching a graduate elective at IIT Bombay in the spring of 2001. The general style is a mix of scientific and statistical programming with system engineering and optimizations. A background in elementary undergraduate statistics, algorithms, and networking should suffice to follow the material. The exposition also assumes that the reader is a regular user of search engines, topic directories, and Web content in general, and has some appreciation for the limitations of basic Web access based on clicking on links and typing keyword queries.

The chapters fall into three major parts. For concreteness, we start with some engineering issues: crawling, indexing, and keyword search. This part also gives us some basic know-how for efficiently representing, manipulating, and analyzing hypertext documents with computer programs. In the second part, which is the bulk of the book, we focus on machine learning for hypertext: the art of creating programs that seek out statistical relations between attributes extracted from Web documents. Such relations can be used to discover topic-based clusters from a collection of Web pages, assign a Web page to a predefined topic, or match a user's interest to Web sites. The third part is a collection of applications that draw upon the techniques discussed in the first two parts.

To make the presentation concrete, specific URLs are indicated throughout, but there is no saying how long they will remain accessible on the Web. Luckily, the Internet Archive will let you view old versions of pages at *www.archive.org/*, provided *this* URL does not get dated.

Omissions

The field of research underlying this book is in rapid flux. A book written at this juncture is guaranteed to miss out on important areas. At some point a snapshot

must be taken to complete the project. A few omissions, however, are deliberate. Beyond bare necessities, I have not engaged in a study of protocols for representing and transferring content on the Internet and the Web. Readers are assumed to be reasonably familiar with HTML. For the purposes of this book, you do not need to understand the *XML* (*Extensible Markup Language*) standard much more deeply than HTML. There is also no treatment of Web application services, dynamic site management, or associated networking and data-processing technology.

I make no attempt to cover natural language (NL) processing, natural language understanding, or knowledge representation. This is largely because I do not know enough about natural language processing. NL techniques can now parse relatively well-formed sentences in many languages, disambiguate polysemous words with high accuracy, tag words in running text with part-of-speech information, represent NL documents in a canonical machine-usable form, and perform NL translation. Web search engines have been slow to embrace NL processing except as an explicit translation service. In this book, I will make occasional references to what has been called "ankle-deep semantics"—techniques that leverage semantic databases (e.g., as a dictionary or thesaurus) in shallow, efficient ways to improve keyword search.

Another missing area is Web *usage* mining. Optimizing large, high-flux Web sites to be visitor-friendly is nontrivial. Monitoring and analyzing the behavior of visitors in the past may lead to valuable insights into their information needs, and help in continually adapting the design of the site. Several companies have built systems integrated with Web servers, especially the kind that hosts e-commerce sites, to monitor and analyze traffic and propose site organization strategies. The array of techniques brought to bear on usage mining has a large overlap with traditional data mining in the relational data-warehousing scenario, for which excellent texts already exist.

Acknowledgments

I am grateful to many people for making this work possible. I was fortunate to associate with Byron Dom, Inderjit Dhillon, Dharmendra Modha, David Gibson, Dimitrios Gunopulos, Jon Kleinberg, Kevin McCurley, Nimrod Megiddo, and Prabhakar Raghavan at IBM Almaden Research Center, where some of the inventions described in this book were made between 1996 and 1999. I also acknowledge the extremely stimulating discussions I have had with researchers at the then Digital System Research Center in Palo Alto, California: Krishna

Bharat, Andrei Bröder, Monika Henzinger, Hannes Marais, and Mark Najork, some of whom have moved on to Google and AltaVista. Similar gratitude is also due to Gary Flake, C. Lee Giles, Steve Lawrence, and Dave Pennock at NEC Research, Princeton. Thanks also to Pedro Domingos, Susan Dumais, Ravindra Jaju, Ronny Lempel, David Lewis, Tom Mitchell, Mandar Mitra, Kunal Punera, Mehran Sahami, Eric Saund, and Amit Singhal for helpful discussions. Jiawei Han's text on data mining and his encouragement helped me decide to write this book. Krishna Bharat, Lyle Ungar, Karen Watterson, Ian Witten, and other, anonymous, referees have greatly enhanced the quality of the manuscript.

Closer to home, Sunita Sarawagi and S. Sudarshan gave valuable feedback. Together with Pushpak Bhattacharya and Krithi Ramamritham, they kept up my enthusiasm during this long project in the face of many adversities. I am grateful to Tata Consultancy Services for their generous support through the Lab for Intelligent Internet Research during the preparation of the manuscript. T. P. Chandran offered invaluable administrative help. I thank Diane Cerra, Lothlórien Homet, Edward Wade, Mona Buehler, Corina Derman, and all the other members of the Morgan Kaufmann team for their patience with many delays in the schedule and their superb production job. I regret forgetting to express my gratitude to anyone else who has contributed to this work. The gratitude does live on in my heart. Finally, I wish to thank my wife, Sunita Sarawagi, and my parents, Sunil and Arati Chakrabarti, for their constant support and encouragement.

CONTENTS

1 **INTRODUCTION**

1.1 Crawling and Indexing 6
1.2 Topic Directories 7
1.3 Clustering and Classification 8
1.4 Hyperlink Analysis 9
1.5 Resource Discovery and Vertical Portals 11
1.6 Structured vs. Unstructured Data Mining 11
1.7 Bibliographic Notes 13

PART I **INFRASTRUCTURE**

2 **CRAWLING THE WEB**

2.1 HTML and HTTP Basics 18
2.2 Crawling Basics 19
2.3 Engineering Large-Scale Crawlers 21
2.3.1 DNS Caching, Prefetching, and Resolution 22
2.3.2 Multiple Concurrent Fetches 23
2.3.3 Link Extraction and Normalization 25

2.3.4 Robot Exclusion 26
2.3.5 Eliminating Already-Visited URLs 26
2.3.6 Spider Traps 28
2.3.7 Avoiding Repeated Expansion of Links on Duplicate Pages 29
2.3.8 Load Monitor and Manager 29
2.3.9 Per-Server Work-Queues 30
2.3.10 Text Repository 31
2.3.11 Refreshing Crawled Pages 33
2.4 Putting Together a Crawler 35
2.4.1 Design of the Core Components 35
2.4.2 Case Study: Using `w3c-libwww` 40
2.5 Bibliographic Notes 40

3 **WEB SEARCH AND INFORMATION RETRIEVAL**

3.1 Boolean Queries and the Inverted Index 45
3.1.1 Stopwords and Stemming 48
3.1.2 Batch Indexing and Updates 49
3.1.3 Index Compression Techniques 51
3.2 Relevance Ranking 53
3.2.1 Recall and Precision 53
3.2.2 The Vector-Space Model 56
3.2.3 Relevance Feedback and Rocchio's Method 57
3.2.4 Probabilistic Relevance Feedback Models 58
3.2.5 Advanced Issues 61
3.3 Similarity Search 67
3.3.1 Handling "Find-Similar" Queries 68
3.3.2 Eliminating Near Duplicates via Shingling 71
3.3.3 Detecting Locally Similar Subgraphs of the Web 73
3.4 Bibliographic Notes 75

PART II **LEARNING**

4 **SIMILARITY AND CLUSTERING**

4.1 Formulations and Approaches 81
4.1.1 Partitioning Approaches 81
4.1.2 Geometric Embedding Approaches 82
4.1.3 Generative Models and Probabilistic Approaches 83
4.2 Bottom-Up and Top-Down Partitioning Paradigms 84
4.2.1 Agglomerative Clustering 84
4.2.2 The *k*-Means Algorithm 87
4.3 Clustering and Visualization via Embeddings 89
4.3.1 Self-Organizing Maps (SOMs) 90
4.3.2 Multidimensional Scaling (MDS) and FastMap 91
4.3.3 Projections and Subspaces 94
4.3.4 Latent Semantic Indexing (LSI) 96
4.4 Probabilistic Approaches to Clustering 99
4.4.1 Generative Distributions for Documents 101
4.4.2 Mixture Models and Expectation Maximization (EM) 103
4.4.3 Multiple Cause Mixture Model (MCMM) 108
4.4.4 Aspect Models and Probabilistic LSI 109
4.4.5 Model and Feature Selection 112
4.5 Collaborative Filtering 115
4.5.1 Probabilistic Models 115
4.5.2 Combining Content-Based and Collaborative Features 117
4.6 Bibliographic Notes 121

5 **SUPERVISED LEARNING**

5.1 The Supervised Learning Scenario 126
5.2 Overview of Classification Strategies 128

5.3 Evaluating Text Classifiers 129
5.3.1 Benchmarks 130
5.3.2 Measures of Accuracy 131
5.4 Nearest Neighbor Learners 133
5.4.1 Pros and Cons 134
5.4.2 Is TFIDF Appropriate? 135
5.5 Feature Selection 136
5.5.1 Greedy Inclusion Algorithms 137
5.5.2 Truncation Algorithms 144
5.5.3 Comparison and Discussion 145
5.6 Bayesian Learners 147
5.6.1 Naive Bayes Learners 148
5.6.2 Small-Degree Bayesian Networks 152
5.7 Exploiting Hierarchy among Topics 155
5.7.1 Feature Selection 155
5.7.2 Enhanced Parameter Estimation 155
5.7.3 Training and Search Strategies 157
5.8 Maximum Entropy Learners 160
5.9 Discriminative Classification 163
5.9.1 Linear Least-Square Regression 163
5.9.2 Support Vector Machines 164
5.10 Hypertext Classification 169
5.10.1 Representing Hypertext for Supervised Learning 169
5.10.2 Rule Induction 171
5.11 Bibliographic Notes 173

6 **SEMISUPERVISED LEARNING**

6.1 Expectation Maximization 178
6.1.1 Experimental Results 179
6.1.2 Reducing the Belief in Unlabeled Documents 181
6.1.3 Modeling Labels Using Many Mixture Components 183

6.2 Labeling Hypertext Graphs 184
6.2.1 Absorbing Features from Neighboring Pages 185
6.2.2 A Relaxation Labeling Algorithm 188
6.2.3 A Metric Graph-Labeling Problem 193
6.3 Co-training 195
6.4 Bibliographic Notes 198

PART III **APPLICATIONS**

7 **SOCIAL NETWORK ANALYSIS**

7.1 Social Sciences and Bibliometry 205
7.1.1 Prestige 205
7.1.2 Centrality 206
7.1.3 Co-citation 207
7.2 PageRank and HITS 209
7.2.1 PageRank 209
7.2.2 HITS 212
7.2.3 Stochastic HITS and Other Variants 216
7.3 Shortcomings of the Coarse-Grained Graph Model 219
7.3.1 Artifacts of Web Authorship 219
7.3.2 Topic Contamination and Drift 223
7.4 Enhanced Models and Techniques 225
7.4.1 Avoiding Two-Party Nepotism 225
7.4.2 Outlier Elimination 226
7.4.3 Exploiting Anchor Text 227
7.4.4 Exploiting Document Markup Structure 228
7.5 Evaluation of Topic Distillation 235
7.5.1 HITS and Related Algorithms 235
7.5.2 Effect of Exploiting Other Hypertext Features 238
7.6 Measuring and Modeling the Web 243
7.6.1 Power-Law Degree Distributions 243

7.6.2 The "Bow Tie" Structure and Bipartite Cores 246
7.6.3 Sampling Web Pages at Random 246
7.7 Bibliographic Notes 254

8 **RESOURCE DISCOVERY**

8.1 Collecting Important Pages Preferentially 257
8.1.1 Crawling as Guided Search in a Graph 257
8.1.2 Keyword-Based Graph Search 259
8.2 Similarity Search Using Link Topology 264
8.3 Topical Locality and Focused Crawling 268
8.3.1 Focused Crawling 270
8.3.2 Identifying and Exploiting Hubs 277
8.3.3 Learning Context Graphs 279
8.3.4 Reinforcement Learning 280
8.4 Discovering Communities 284
8.4.1 Bipartite Cores as Communities 284
8.4.2 Network Flow/Cut-Based Notions of Communities 285
8.5 Bibliographic Notes 288

9 **THE FUTURE OF WEB MINING**

9.1 Information Extraction 290
9.2 Natural Language Processing 295
9.2.1 Lexical Networks and Ontologies 296
9.2.2 Part-of-Speech and Sense Tagging 297
9.2.3 Parsing and Knowledge Representation 299
9.3 Question Answering 302
9.4 Profiles, Personalization, and Collaboration 305

References 307
Index 327

CHAPTER 1

INTRODUCTION

The World Wide Web is the largest and most widely known repository of hypertext. Hypertext documents contain text and generally embed hyperlinks to other documents distributed across the Web. Today, the Web comprises billions of documents, authored by millions of diverse people, edited by no one in particular, and distributed over millions of computers that are connected by telephone lines, optical fibers, and radio modems. It is a wonder that the Web works at all. Yet it is rapidly assisting and supplementing newspapers, radio, television, and telephone, the postal system, schools and colleges, libraries, physical workplaces, and even the sites of commerce and governance.

A brief history of hypertext and the Web. Citation, a form of hyperlinking, is as old as written language itself. The Talmud, with its heavy use of annotations and nested commentary, and the Ramayana and Mahabharata, with their branching, nonlinear discourse, are ancient examples of hypertext. Dictionaries and encyclopedias can be viewed as a self-contained network of textual nodes joined by referential links. Words and concepts are described by appealing to other words and concepts. In modern times (1945), Vannevar Bush is credited with the first design of a photo-electrical-mechanical storage and computing device called a Memex (for "memory extension"), which could create and help follow hyperlinks across documents. Doug Engelbart and Ted Nelson were other early pioneers; Ted Nelson coined the term *hypertext* in 1965 [160] and created the Xanadu hypertext system with robust two-way hyperlinks, version management, controversy management, annotation, and copyright management.