

Mining of Massive Datasets

Second Edition

The popularity of the Web and Internet commerce provides many extremely large datasets from which information can be gleaned by data mining. This book focuses on practical algorithms that have been used to solve key problems in data mining and can be used on even the largest datasets.

It begins with a discussion of the map-reduce framework, an important tool for parallelizing algorithms automatically. The tricks of locality-sensitive hashing are explained. This body of knowledge, which deserves to be more widely known, is essential when seeking similar objects in a very large collection without having to compare each pair of objects. Stream processing algorithms for mining data that arrives too fast for exhaustive processing are also explained. The PageRank idea and related tricks for organizing the Web are covered next. Other chapters cover the problems of finding frequent itemsets and clustering, each from the point of view that the data is too large to fit in main memory, and two applications: recommendation systems and Web advertising, each vital in e-commerce.

This second edition includes new and extended coverage on social networks, machine learning and dimensionality reduction. Written by leading authorities in database and web technologies, it is essential reading for students and practitioners alike

Cambridge University Press
978-1-107-07723-2 - Mining of Massive Datasets: Second Edition
Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman
Frontmatter
[More information](#)

Cambridge University Press
978-1-107-07723-2 - Mining of Massive Datasets: Second Edition
Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman
Frontmatter
[More information](#)

Mining of Massive Datasets

Second Edition

JURE LESKOVEC

Stanford University

ANAND RAJARAMAN

Milliways Labs

JEFFREY DAVID ULLMAN

Stanford University



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press
978-1-107-07723-2 - Mining of Massive Datasets: Second Edition
Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman
Frontmatter
[More information](#)

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107077232

First edition © A. Rajaraman and J. D. Ullman 2012

Second edition © J. Leskovec, A. Rajaraman and J. D. Ullman 2014

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2012

Second edition 2014

Reprinted 2015

Printed in the United States of America by Sheridan Books, Inc.

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-07723-2 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

	<i>Preface</i>	<i>page ix</i>
1	Data Mining	1
	1.1 What is Data Mining?	1
	1.2 Statistical Limits on Data Mining	4
	1.3 Things Useful to Know	7
	1.4 Outline of the Book	15
	1.5 Summary of Chapter 1	17
	1.6 References for Chapter 1	17
2	MapReduce and the New Software Stack	19
	2.1 Distributed File Systems	20
	2.2 MapReduce	22
	2.3 Algorithms Using MapReduce	28
	2.4 Extensions to MapReduce	38
	2.5 The Communication Cost Model	44
	2.6 Complexity Theory for MapReduce	50
	2.7 Summary of Chapter 2	64
	2.8 References for Chapter 2	66
3	Finding Similar Items	68
	3.1 Applications of Near-Neighbor Search	68
	3.2 Shingling of Documents	72
	3.3 Similarity-Preserving Summaries of Sets	75
	3.4 Locality-Sensitive Hashing for Documents	82
	3.5 Distance Measures	87
	3.6 The Theory of Locality-Sensitive Functions	93
	3.7 LSH Families for Other Distance Measures	98
	3.8 Applications of Locality-Sensitive Hashing	104
	3.9 Methods for High Degrees of Similarity	111
	3.10 Summary of Chapter 3	119
	3.11 References for Chapter 3	122
4	Mining Data Streams	123

4.1	The Stream Data Model	123
4.2	Sampling Data in a Stream	127
4.3	Filtering Streams	130
4.4	Counting Distinct Elements in a Stream	133
4.5	Estimating Moments	137
4.6	Counting Ones in a Window	142
4.7	Decaying Windows	148
4.8	Summary of Chapter 4	150
4.9	References for Chapter 4	152
5	Link Analysis	154
5.1	PageRank	154
5.2	Efficient Computation of PageRank	168
5.3	Topic-Sensitive PageRank	174
5.4	Link Spam	178
5.5	Hubs and Authorities	182
5.6	Summary of Chapter 5	187
5.7	References for Chapter 5	190
6	Frequent Itemsets	191
6.1	The Market-Basket Model	191
6.2	Market Baskets and the A-Priori Algorithm	198
6.3	Handling Larger Datasets in Main Memory	207
6.4	Limited-Pass Algorithms	214
6.5	Counting Frequent Items in a Stream	220
6.6	Summary of Chapter 6	224
6.7	References for Chapter 6	226
7	Clustering	228
7.1	Introduction to Clustering Techniques	228
7.2	Hierarchical Clustering	232
7.3	K-means Algorithms	241
7.4	The CURE Algorithm	249
7.5	Clustering in Non-Euclidean Spaces	252
7.6	Clustering for Streams and Parallelism	256
7.7	Summary of Chapter 7	262
7.8	References for Chapter 7	265
8	Advertising on the Web	267
8.1	Issues in On-Line Advertising	267
8.2	On-Line Algorithms	270
8.3	The Matching Problem	273
8.4	The Adwords Problem	276
8.5	Adwords Implementation	285

	8.6 Summary of Chapter 8	289
	8.7 References for Chapter 8	290
9	Recommendation Systems	292
	9.1 A Model for Recommendation Systems	292
	9.2 Content-Based Recommendations	296
	9.3 Collaborative Filtering	306
	9.4 Dimensionality Reduction	312
	9.5 The NetFlix Challenge	321
	9.6 Summary of Chapter 9	322
	9.7 References for Chapter 9	323
10	Mining Social-Network Graphs	325
	10.1 Social Networks as Graphs	325
	10.2 Clustering of Social-Network Graphs	330
	10.3 Direct Discovery of Communities	338
	10.4 Partitioning of Graphs	343
	10.5 Finding Overlapping Communities	350
	10.6 Simrank	357
	10.7 Counting Triangles	361
	10.8 Neighborhood Properties of Graphs	367
	10.9 Summary of Chapter 10	378
	10.10 References for Chapter 10	381
11	Dimensionality Reduction	384
	11.1 Eigenvalues and Eigenvectors	384
	11.2 Principal-Component Analysis	391
	11.3 Singular-Value Decomposition	397
	11.4 CUR Decomposition	406
	11.5 Summary of Chapter 11	412
	11.6 References for Chapter 11	414
12	Large-Scale Machine Learning	415
	12.1 The Machine-Learning Model	416
	12.2 Perceptrons	422
	12.3 Support-Vector Machines	436
	12.4 Learning from Nearest Neighbors	447
	12.5 Comparison of Learning Methods	455
	12.6 Summary of Chapter 12	456
	12.7 References for Chapter 12	457
	<i>Index</i>	459

Cambridge University Press
978-1-107-07723-2 - Mining of Massive Datasets: Second Edition
Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman
Frontmatter
[More information](#)

Preface

This book evolved from material developed over several years by Anand Rajaraman and Jeff Ullman for a one-quarter course at Stanford. The course CS345A, titled “Web Mining,” was designed as an advanced graduate course, although it has become accessible and interesting to advanced undergraduates. When Jure Leskovec joined the Stanford faculty, we reorganized the material considerably. He introduced a new course CS224W on network analysis and added material to CS345A, which was renumbered CS246. The three authors also introduced a large-scale data-mining project course, CS341. The book now contains material taught in all three courses.

What the Book Is About

At the highest level of description, this book is about data mining. However, it focuses on data mining of very large amounts of data, that is, data so large it does not fit in main memory. Because of the emphasis on size, many of our examples are about the Web or data derived from the Web. Further, the book takes an algorithmic point of view: data mining is about applying algorithms to data, rather than using data to “train” a machine-learning engine of some sort. The principal topics covered are:

- (1) Distributed file systems and map-reduce as a tool for creating parallel algorithms that succeed on very large amounts of data.
- (2) Similarity search, including the key techniques of minhashing and locality-sensitive hashing.
- (3) Data-stream processing and specialized algorithms for dealing with data that arrives so fast it must be processed immediately or lost.
- (4) The technology of search engines, including Google’s PageRank, link-spam detection, and the hubs-and-authorities approach.
- (5) Frequent-itemset mining, including association rules, market-baskets, the A-Priori Algorithm and its improvements.
- (6) Algorithms for clustering very large, high-dimensional datasets.
- (7) Two key problems for Web applications: managing advertising and recommendation systems.

- (8) Algorithms for analyzing and mining the structure of very large graphs, especially social-network graphs.
- (9) Techniques for obtaining the important properties of a large dataset by dimensionality reduction, including singular-value decomposition and latent semantic indexing.
- (10) Machine-learning algorithms that can be applied to very large data, such as perceptrons, support-vector machines, and gradient descent.

Prerequisites

To appreciate fully the material in this book, we recommend the following prerequisites:

- (1) An introduction to database systems, covering SQL and related programming systems.
- (2) A sophomore-level course in data structures, algorithms, and discrete math.
- (3) A sophomore-level course in software systems, software engineering, and programming languages.

Exercises

The book contains extensive exercises, with some for almost every section. We indicate harder exercises or parts of exercises with an exclamation point. The hardest exercises have a double exclamation point.

Support on the Web

You can find materials from past offerings of CS345A at:

<http://i.stanford.edu/~ullman/mining/mining.html>

There, you will find slides, homework assignments, project requirements, and in some cases, exams.

Gradiance Automated Homework

There are automated exercises based on this book, using the Gradiance root-question technology, available at www.gradiance.com/services. Students may enter a public class by creating an account at that site and entering the class with code 1EDD8A1D. Instructors may use the site by making an account there and then emailing `support` at `gradiance dot com` with their login name, the name of their school, and a request to use the MMDS materials.

Acknowledgements

Cover art is by Scott Ullman.

We would like to thank Foto Afrati, Arun Marathe, and Rok Sosis for critical readings of a draft of this manuscript.

Errors were also reported by Apoorv Agarwal, Aris Anagnostopoulos, Atilla Soner Balkir, Robin Bennett, Susan Biancani, Amitabh Chaudhary, Leland Chen, Anastasios Gounaris, Shrey Gupta, Waleed Hameid, Ed Knorr, Haewoon Kwak, Ellis Lau, Ethan Lozano, Michael Mahoney, Justin Meyer, Brad Penoff, Philips Kokoh Prasetyo, Qi Ge, Angad Singh, Sandeep Sripada, Dennis Sidharta, Krzysztof Stencel, Mark Storus, Roshan Sumbaly, Zack Taylor, Tim Triche Jr., Wang Bin, Weng Zhen-Bin, Robert West, Oscar Wu, Xie Ke, Nicolas Zhao, and Zhou Jingbo. The remaining errors are ours, of course.

J. L.
A. R.
J. D. U.
Palo Alto, CA
March, 2014