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

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

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 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

Contents

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 Sosic 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