Mining of Massive Datasets

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. The final chapters cover two applications: recommendation systems and Web advertising, each vital in e-commerce.

Written by two authorities in database and web technologies, this book will be essential for students and practitioners alike.

Cambridge University Press
978-1-107-01535-7 - Mining of Massive Datasets
Anand Rajaraman and Jeffrey David Ullman
Frontmatter
More information

Mining of Massive Datasets

ANAND RAJARAMAN
@WalmartLabs

JEFFREY DAVID ULLMAN
Stanford University

CAMBRIDGE
UNIVERSITY PRESS

cambridge university press
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9781107015357

First published 2012

Printed in the United Kingdom at the University Press, Cambridge

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

ISBN 978-1-107-01535-7 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.

Cambridge University Press
978-1-107-01535-7 - Mining of Massive Datasets
Anand Rajaraman and Jeffrey David Ullman
Frontmatter
More information

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.

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.

Prerequisites

CS345A, although its number indicates an advanced graduate course, has been found accessible by advanced undergraduates and beginning masters students. In the future, it is likely that the course will be given a mezzanine-level number. The prerequisites for CS345A are:

(1) The first course in database systems, covering application programming in SQL and other database-related languages such as XQuery.
(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://infolab.stanford.edu/~ullman/mining/mining.html`

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

Acknowledgements

Cover art is by Scott Ullman. We would like to thank Foto Afrati and Arun Marathe for critical readings of the draft of this manuscript. Errors were also reported by Leland Chen, Shrey Gupta, Xie Ke, Haewoon Kwak, Brad Penoff, Philips Kokoh Prasetyo, Mark Storus, Tim Triche Jr., and Roshan Sumbaly. The remaining errors are ours, of course.

A. R.
J. D. U.
Palo Alto, CA
June, 2011