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.
## Contents

*Preface*

1 Data Mining
   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 16
   1.6 References for Chapter 1 17

2 Large-Scale File Systems and Map-Reduce
   2.1 Distributed File Systems 18
   2.2 Map-Reduce 21
   2.3 Algorithms Using Map-Reduce 26
   2.4 Extensions to Map-Reduce 37
   2.5 Efficiency of Cluster-Computing Algorithms 42
   2.6 Summary of Chapter 2 49
   2.7 References for Chapter 2 51

3 Finding Similar Items
   3.1 Applications of Near-Neighbor Search 53
   3.2 Shingling of Documents 57
   3.3 Similarity-Preserving Summaries of Sets 60
   3.4 Locality-Sensitive Hashing for Documents 67
   3.5 Distance Measures 71
   3.6 The Theory of Locality-Sensitive Functions 77
   3.7 LSH Families for Other Distance Measures 83
   3.8 Applications of Locality-Sensitive Hashing 88
   3.9 Methods for High Degrees of Similarity 96
   3.10 Summary of Chapter 3 104
   3.11 References for Chapter 3 106

4 Mining Data Streams
   4.1 The Stream Data Model 108
## Contents

4.2 Sampling Data in a Stream 112  
4.3 Filtering Streams 115  
4.4 Counting Distinct Elements in a Stream 118  
4.5 Estimating Moments 122  
4.6 Counting Ones in a Window 127  
4.7 Decaying Windows 133  
4.8 Summary of Chapter 4 136  
4.9 References for Chapter 4 137  

5 Link Analysis  
5.1 PageRank 139  
5.2 Efficient Computation of PageRank 153  
5.3 Topic-Sensitive PageRank 159  
5.4 Link Spam 163  
5.5 Hubs and Authorities 167  
5.6 Summary of Chapter 5 172  
5.7 References for Chapter 5 175  

6 Frequent Itemsets 176  
6.1 The Market-Basket Model 176  
6.2 Market Baskets and the A-Priori Algorithm 183  
6.3 Handling Larger Datasets in Main Memory 192  
6.4 Limited-Pass Algorithms 199  
6.5 Counting Frequent Items in a Stream 205  
6.6 Summary of Chapter 6 209  
6.7 References for Chapter 6 211  

7 Clustering 213  
7.1 Introduction to Clustering Techniques 213  
7.2 Hierarchical Clustering 217  
7.3 K-means Algorithms 226  
7.4 The CURE Algorithm 234  
7.5 Clustering in Non-Euclidean Spaces 237  
7.6 Clustering for Streams and Parallelism 241  
7.7 Summary of Chapter 7 247  
7.8 References for Chapter 7 250  

8 Advertising on the Web 252  
8.1 Issues in On-Line Advertising 252  
8.2 On-Line Algorithms 255  
8.3 The Matching Problem 258  
8.4 The Adwords Problem 261  
8.5 Adwords Implementation 270  
8.6 Summary of Chapter 8 273
## Contents

| 8.7 | References for Chapter 8 | 275 |
| 9   | **Recommendation Systems** | 277 |
| 9.1 | A Model for Recommendation Systems | 277 |
| 9.2 | Content-Based Recommendations | 281 |
| 9.3 | Collaborative Filtering | 291 |
| 9.4 | Dimensionality Reduction | 297 |
| 9.5 | The NetFlix Challenge | 305 |
| 9.6 | Summary of Chapter 9 | 306 |
| 9.7 | References for Chapter 9 | 308 |
|     | **Index** | 310 |
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.
Preface

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, Philipps 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