

# Contents

	<i>Preface</i>	<i>page ix</i>
<b>1</b>	<b>Data Mining</b>	<b>1</b>
	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
<b>2</b>	<b>MapReduce and the New Software Stack</b>	<b>19</b>
	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
<b>3</b>	<b>Finding Similar Items</b>	<b>68</b>
	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
<b>4</b>	<b>Mining Data Streams</b>	<b>123</b>

vi	<b>Contents</b>	
	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
<b>5</b>	<b>Link Analysis</b>	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
<b>6</b>	<b>Frequent Itemsets</b>	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
<b>7</b>	<b>Clustering</b>	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
<b>8</b>	<b>Advertising on the Web</b>	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

	<b>Contents</b>	vii
8.6	Summary of Chapter 8	289
8.7	References for Chapter 8	290
<b>9</b>	<b>Recommendation Systems</b>	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
<b>10</b>	<b>Mining Social-Network Graphs</b>	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
<b>11</b>	<b>Dimensionality Reduction</b>	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
<b>12</b>	<b>Large-Scale Machine Learning</b>	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