

Contents

1	Introduction — Key Subroutines of Multimedia Data Management ..	1
1.1	Overview	1
1.2	Feature Extraction	2
1.3	Similarity	3
1.4	Learning	4
1.5	Multimodal Fusion	5
1.6	Indexing	8
1.7	Scalability	9
1.8	Concluding Remarks	9
	References	10
2	Perceptual Feature Extraction	13
2.1	Introduction	13
2.2	DMD Algorithm	16
2.2.1	Model-Based Pipeline	16
2.2.2	Data-Driven Pipeline	22
2.3	Experiments	23
2.3.1	Dataset and Setup	24
2.3.2	Model-Based vs. Data-Driven	24
2.3.3	DMD vs. Individual Models	29
2.3.4	Regularization Tuning	31
2.3.5	Tough Categories	31
2.4	Related Reading	31
2.5	Concluding Remarks	33
	References	33
3	Query Concept Learning	37
3.1	Introduction	37
3.2	Support Vector Machines and Version Space	39
3.3	Active Learning and Batch Sampling Strategies	42
3.3.1	Theoretical Foundation	43

3.3.2	Sampling Strategies	45
3.4	Concept-Dependent Learning	49
3.4.1	Concept Complexity	50
3.4.2	Limitations of Active Learning	53
3.4.3	Concept-Dependent Active Learning Algorithms	55
3.5	Experiments and Discussion	58
3.5.1	Testbed and Setup	58
3.5.2	Active vs. Passive Learning	60
3.5.3	Against Traditional Relevance Feedback Schemes	61
3.5.4	Sampling Method Evaluation	62
3.5.5	Concept-Dependent Learning	63
3.5.6	Concept Diversity Evaluation	67
3.5.7	Evaluation Summary	67
3.6	Related Reading	68
3.6.1	Machine Learning	68
3.6.2	Relevance Feedback	69
3.7	Relation to Other Chapters	70
3.8	Concluding Remarks	70
	References	70
4	Similarity	75
4.1	Introduction	75
4.2	Mining Image Feature Set	77
4.2.1	Image Testbed Setup	77
4.2.2	Feature Extraction	78
4.2.3	Feature Selection	79
4.3	Discovering the Dynamic Partial Distance Function	80
4.3.1	Minkowski Metric and Its Limitations	80
4.3.2	Dynamic Partial Distance Function	84
4.3.3	Psychological Interpretation of Dynamic Partial Distance Function	85
4.4	Empirical Study	86
4.4.1	Image Retrieval	86
4.4.2	Video Shot-Transition Detection	91
4.4.3	Near Duplicated Articles	94
4.4.4	Weighted DPF vs. Weighted Euclidean	95
4.4.5	Observations	95
4.5	Related Reading	96
4.6	Concluding Remarks	97
	References	98
5	Formulating Distance Functions	101
5.1	Introduction	101
5.2	DFA Algorithm	105
5.2.1	Transformation Model	105

5.2.2	Distance Metric Learning	108
5.3	Experimental Evaluation	112
5.3.1	Evaluation on Contextual Information	114
5.3.2	Evaluation on Effectiveness	115
5.3.3	Observations	118
5.4	Related Reading	119
5.4.1	Metric Learning	119
5.4.2	Kernel Learning	121
5.5	Concluding Remarks	123
	References	123
6	Multimodal Fusion	125
6.1	Introduction	125
6.2	Related Reading	128
6.2.1	Modality Identification	129
6.2.2	Modality Fusion	130
6.3	Independent Modality Analysis	131
6.3.1	PCA	131
6.3.2	ICA	131
6.3.3	IMG	133
6.4	Super-Kernel Fusion	134
6.5	Experiments	137
6.5.1	Evaluation of Modality Analysis	139
6.5.2	Evaluation of Multimodal Kernel Fusion	140
6.5.3	Observations	142
6.6	Concluding Remarks	142
	References	143
7	Fusing Content and Context with Causality	145
7.1	Introduction	145
7.2	Related Reading	147
7.2.1	Photo Annotation	147
7.2.2	Probabilistic Graphical Models	149
7.3	Multimodal Metadata	149
7.3.1	Contextual Information	149
7.3.2	Perceptual Content	151
7.3.3	Semantic Ontology	151
7.4	Influence Diagrams	152
7.4.1	Structure Learning	153
7.4.2	Causal Strength	159
7.4.3	Case Study	160
7.4.4	Dealing with Missing Attributes	163
7.5	Experiments	163
7.5.1	Experiment on Learning Structure	165
7.5.2	Experiment on Causal Strength Inference	165

7.5.3	Experiment on Semantic Fusion	169
7.5.4	Experiment on Missing Features	171
7.6	Concluding Remarks	172
	References	173
8	Combinational Collaborative Filtering, Considering Personalization	175
8.1	Introduction	175
8.2	Related Reading	176
8.3	CCF: Combinational Collaborative Filtering	177
8.3.1	C-U and C-D Baseline Models	178
8.3.2	CCF Model	179
8.3.3	Gibbs & EM Hybrid Training	179
8.3.4	Parallelization	182
8.3.5	Inference	183
8.4	Experiments	185
8.4.1	Gibbs + EM vs. EM	185
8.4.2	The Orkut Dataset	187
8.4.3	Runtime Speedup	192
8.5	Concluding Remarks	194
	References	195
9	Imbalanced Data Learning	197
9.1	Introduction	197
9.2	Related Reading	200
9.3	Kernel Boundary Alignment	202
9.3.1	Conformally Transforming Kernel K	203
9.3.2	Modifying Kernel Matrix K	205
9.4	Experimental Results	211
9.4.1	Vector-Space Evaluation	212
9.4.2	Non-Vector-Space Evaluation	215
9.5	Concluding Remarks	215
	References	216
10	PSVM: Parallelizing Support Vector Machines on Distributed Computers	219
10.1	Introduction	219
10.2	Interior Point Method with Incomplete Cholesky Factorization	221
10.3	PSVM Algorithm	223
10.3.1	Parallel ICF	225
10.3.2	Parallel IPM	229
10.3.3	Computing Parameter b and Writing Back	230
10.4	Experiments	231
10.4.1	Class-Prediction Accuracy	231
10.4.2	Scalability	232
10.4.3	Overheads	233

10.5 Concluding Remarks	235
References	235
11 Approximate High-Dimensional Indexing with Kernel	237
11.1 Introduction	238
11.2 Related Reading	239
11.3 Algorithm SphereDex	240
11.3.1 Create — Building the Index	241
11.3.2 Search — Querying the Index	244
11.3.3 Update — Insertion and Deletion	249
11.4 Experiments	253
11.4.1 Setup	254
11.4.2 Performance with Disk IOs	256
11.4.3 Choice of Parameter g	259
11.4.4 Impact of Insertions	260
11.4.5 Sequential vs. Random	261
11.4.6 Percentage of Data Processed	261
11.4.7 Summary	263
11.5 Concluding Remarks	263
11.5.1 Range Queries	263
11.5.2 Farthest Neighbor Queries	264
References	264
12 Speeding Up Latent Dirichlet Allocation with Parallelization and Pipeline Strategies	267
12.1 Introduction	267
12.2 Related Reading	269
12.3 AD-LDA: Approximate Distributed LDA	271
12.3.1 Parallel Gibbs Sampling and AllReduce	271
12.3.2 MPI Implementation of AD-LDA	272
12.4 PLDA+	274
12.4.1 Reduce Bottleneck of AD-LDA	274
12.4.2 Framework of PLDA+	275
12.4.3 Algorithm for P_w Processors	277
12.4.4 Algorithm for P_d Processors	279
12.4.5 Straggler Handling	283
12.4.6 Parameters and Complexity	284
12.5 Experimental Results	285
12.5.1 Datasets and Experiment Environment	286
12.5.2 Perplexity	286
12.5.3 Speedups and Scalability	287
12.6 Large-Scale Applications	290
12.6.1 Mining Social-Network User Latent Behavior	291
12.6.2 Question Labeling (QL)	292
12.7 Concluding Remarks	293

References	294
Index	299