Social Media Data Mining and Analytics

Gabor Szabo
Gungor Polatkan
Oscar Boykin
Antonios Chalkiopoulos

WILEY

Contents

Introduction **xvii**

Chapter 1 **Users: The *Who* of Social Media** **1**
Measuring Variations in User Behavior in Wikipedia 2
The Diversity of User Activities 3
The Origin of the User Activity Distribution 12
The Consequences of the Power Law 20
The Long Tail in Human Activities 25
Long Tails Everywhere: The 80/20 Rule (*p*/*q* Rule) 28
Online Behavior on Twitter 32
Retrieving Tweets for Users 33
Logarithmic Binning 36
User Activities on Twitter 37
Summary 39

Chapter 2 **Networks: The *How* of Social Media** **41**
Types and Properties of Social Networks 42
When Users Create the Connections: Explicit Networks 43
Directed Versus Undirected Graphs 45
Node and Edge Properties 45
Weighted Graphs 46
Creating Graphs from Activities: Implicit Networks 48
Visualizing Networks 51
Degrees: The Winner Takes All 55
Counting the Number of Connections 57
The Long Tail in User Connections 58
Beyond the Idealized Network Model 62

Capturing Correlations: Triangles, Clustering, and Assortativity 64
Local Triangles and Clustering 64
Assortativity 70
Summary 75

Chapter 3 Temporal Processes: The *When* of Social Media 77
What Traditional Models Tell You About Events in Time 77
When Events Happen Uniformly in Time 79
Inter-Event Times 81
Comparing to a Memoryless Process 86
Autocorrelations 89
Deviations from Memorylessness 91
Periodicities in Time in User Activities 93
Bursty Activities of Individuals 99
Correlations and Bursts 105
Reservoir Sampling 106
Forecasting Metrics in Time 110
Finding Trends 112
Finding Seasonality 115
Forecasting Time Series with ARIMA 117
The Autoregressive Part (“AR”) 118
The Moving Average Part (“MA”) 119
The Full ARIMA(p, d, q) Model 119
Summary 121

Chapter 4 Content: The *What* of Social Media 123
Defining Content: Focus on Text and Unstructured Data 123
Creating Features from Text: The Basics of Natural Language Processing 125
The Basic Statistics of Term Occurrences in Text 128
Using Content Features to Identify Topics 129
The Popularity of Topics 138
How Diverse Are Individual Users’ Interests? 141
Extracting Low-Dimensional Information from High-Dimensional Text 144
Topic Modeling 145
Unsupervised Topic Modeling 147
Supervised Topic Modeling 155
Relational Topic Modeling 162
Summary 169

Chapter 5 Processing Large Datasets 171
MapReduce: Structuring Parallel and Sequential Operations 172
Counting Words 174
Skew: The Curse of the Last Reducer 177
Multi-Stage MapReduce Flows 179
Fan-Out 180

Merging Data Streams 181
Joining Two Data Sources 183
Joining Against Small Datasets 186
Models of Large-Scale MapReduce 187
Patterns in MapReduce Programming 188
Static MapReduce Jobs 188
Iterative MapReduce Jobs 195
PageRank for Ranking in Graphs 195
k-means Clustering 199
Incremental MapReduce Jobs 203
Temporal MapReduce Jobs 204
Rollups and Data Cubing 205
Expanding Rollup Jobs 211
Challenges with Processing Long-Tailed Social Media Data 212
Sampling and Approximations: Getting Results with Less Computation 214
HyperLogLog 217
HyperLogLog Example 219
HyperLogLog on the Stack Exchange Dataset 221
Performance of HLL on Large Datasets 222
Bloom Filters 223
A Bloom Filter Example 226
Bloom Filter as Pre-Computed Membership Knowledge 228
Bloom Filters on Large Social Datasets 229
Count-Min Sketch 231
Count-Min Sketch—Heavy Hitters Example 233
Count-Min Sketch—Top Percentage Example 235
Aggregating Approximate Data Structures 235
Summary of Approximations 236
Executing on a Hadoop Cluster (Amazon EC2) 237
Installing a CDH Cluster on Amazon EC2 237
Providing IAM Access to Collaborators 241
Adding On-Demand Cluster Capabilities 242
Summary 243

Chapter 6 Learn, Map, and Recommend 245
Social Media Services Online 246
Search Engines 246
Content Engagement 246
Interactions with the Real World 248
Interactions with People 249
Problem Formulation 251
Learning and Mapping 253
Matrix Factorization 255
Learning, Training 257
Under- and Overfitting 257
Regularizing in Matrix Factorization 259
Non-Negative Matrix Factorization and Sparsity 260

Demonstration on Movie Ratings 261
Interpreting the Learned Stereotypes 265
Exploratory Analysis 269
Prediction and Recommendation 274
Evaluation 277
Overview of Methodologies 278
Nearest Neighbor-Based Approaches 278
Approaches Based on Supervised Learning 280
Predicting Movie Ratings with Logistic Regression 280
Common Issues with Features 288
Domain-Specific Applications 289
Summary 290

Chapter 7 Conclusions 293
The Surprising Stability of Human Interaction Patterns 293
Averages, Standard Deviations, and Sampling 296
Removing Outliers 303

Index 309