Contents

About the Authors vi

Credits v

Foreword xvii

Chapter 1 Introduction to Data Mining 1

What Is Data Mining 2

Business Problems for Data Mining 5

Data Mining Tasks 6

Classification 6

Clustering 6

Association 7

Regression 8

Forecasting 8

Sequence Analysis 9

Deviation Analysis 10

Data Mining Techniques 11

Data Flow 11

Data Mining Project Cycle 13

Step 1: Data Collection 13

Step 2: Data Cleaning and Transformation 13

Step 3: Model Building 15

Step 4: Model Assessment 16

Step 5: Reporting 16

Step 6: Prediction (Scoring) 16

Step 7: Application Integration 17

Step 8: Model Management 17

Data Mining and the Current Market 17
Data Mining Market Size 17
Major Vendors and Products 18
Current Issues and Challenges 19
Data Mining Standards 20
OLE DB for DM and XML for Analysis 21
SQL/Multimedia for Data Mining 21
Java Data Mining API 23
Predictive Model Markup Language 24
Crisp-DM 28
Common Warehouse Metadata 29
New Trends in Data Mining 31
Summary 33

Chapter 2 OLE DB for Data Mining 35
Introducing OLE DB 36
Why OLE DB for Data Mining? 38
Exploring the Basic Concepts in OLE DB for Data Mining 40
Case 40
The Case Key 41
The Nested Key 41
Case Tables and Nested Tables 42
Scalar Columns and Table Columns 42
The Data Mining Model 42
Model Creation 43
Model Training 43
Model Prediction 43
DMX 43
Three Steps of Data Mining 43
Step 1: Model Creation 45
Step 2: Model Training 49
Step 3: Model Prediction 51
Prediction Functions 54
Singleton Queries 63
Making Predictions Using Content Only 64
Drilling through the Model's Content 65
Content Query 65
Understanding Schema Rowsets 65
The Mining_Services Schema Rowset 66
The Service_Parameters Schema Rowset 68
The Mining_Models Schema Rowset 68
The Mining_Columns Schema Rowset 69
The Mining_Model_Content Schema Rowset 70
The Query_Content Schema Rowset 73
The Mining_Functions Schema Rowset 74
The Model_PMML Schema Rowset 75

Understanding Extensions for Mining Structures 76
The Mining Structure 76
DMX Extensions on Mining Structure 77
Mining Structure Schema Rowsets 78
Summary 79

Chapter 3 Using SQL Server Data Mining 81
Introducing the Business Intelligence Development Studio 82
Understanding the User Interface 82
Offline Mode and Immediate Mode 84
Immediate Mode 85
Getting Started in Immediate Mode 85
Offline Mode 86
Getting Started in Offline Mode 87
Switching Project Modes 89
Creating Data Mining Objects 89
Setting Up Your Data Sources 89
Data Source 89
Creating the MovieClick Data Source 91
Using the Data Source View 91
Creating the MovieClick Data Source View 92
Working with Named Calculations 93
Creating a Named Calculation on the Customers Table 95
Working with Named Queries 96
Creating a Named Query Based on the Customers Table 97
Organizing the DSV 98
Exploring Data 99
Creating and Editing Models 101
Structures and Models 101
Using the Data Mining Wizard 101
Creating the MovieClick Mining Structure and Model 107
Using the Data Mining Designer 108
Working with the Mining Structure Editor 108
Working with the Mining Models Editor 111
Creating and Modifying Additional Models 113
Processing 114
Processing the MovieClick Mining Structure 116
Using Your Models 116
Understanding the Model Viewers 116
Using the Mining Accuracy Chart 118
Creating a Lift Chart on MovieClick 122
Using the Mining Model Prediction Builder 122
Executing a Query on the MovieClick Model 123
Creating Data Mining Reports 124

Using SQL Server Management Studio 126
Understanding the Management Studio User Interface 127
Using the Object Explorer 128
Using the Query Editor 128
Summary 129

Chapter 4 Microsoft Naïve Bayes 131
Introducing the Naïve Bayes Algorithm 132
Understanding Naïve Bayes Principles 132
Naïve Bayes Parameters 135
Using the Naïve Bayes Algorithm 136
DMX 137
Understanding Naïve Bayes Content 138
Exploring a Naïve Bayes Model 140
Dependency Net 140
Attribute Profiles 141
Attribute Characteristics 142
Attribute Discrimination 143
Summary 144

Chapter 5 Microsoft Decision Trees 145
Introducing Decision Trees 145
Decision Tree Principles 147
Basic Concepts of Tree Growth 147
Working with Many States in a Variable 149
Avoiding Overtraining 150
Incorporating Prior Knowledge 151
Feature Selection 151
Using Continuous Inputs 152
Regression 152
Association Analysis with Microsoft Decision Trees 153
Understanding the Algorithm Parameters 155
Using Decision Trees 157
DMX Queries 157
Classification Model 157
Regression Model 159
Association Model 161
Model Content 162
Interpreting the Model 163
Summary 167

Chapter 6 Microsoft Time Series 169
Introducing the Microsoft Time Series Algorithm 170
Introducing the Principles of The Microsoft Time Series Algorithm 171
Autoregression 171
Using Multiple Time Series 173

Autoregression Trees 173
Seasonality 174
Making Historical Predictions 175
Caching Predictions 176
Understanding the Algorithm Parameters 176
Using Microsoft Time Series 177
DMX Queries 178
Model Content 182
Interpreting the Model 182
Summary 185

Chapter 7 Microsoft Clustering 187
Introducing the Microsoft Clustering Algorithm 188
Introducing the Principles of Clustering 190
Hard versus Soft Clustering 191
Discrete Clustering 192
Scalable Clustering 193
Clustering Prediction 194
Introducing the Clustering Parameters 195
Using Clustering Models 198
Clustering as an Analytical Step 199
DMX 199
Cluster 200
ClusterProbability 200
PredictHistogram 201
CaseLikelihood 201
Model Content 202
Understanding Your Cluster Models 203
Get a High-Level Overview 204
Pick a Cluster and Determine How It Is Different 205
Determine How a Cluster Is Different from Nearby Clusters 206
Verify That Your Assertions Are True 207
Label the Cluster 207
Summary 207

Chapter 8 Microsoft Sequence Clustering 209
Introducing the Microsoft Sequence Clustering Algorithm 210
Microsoft Sequence Clustering Algorithm Principles 210
What Is a Markov Chain? 210
Order of a Markov Chain 211
State Transition Matrix 212
Clustering with a Markov Chain 213
Cluster Decomposition 215
Algorithm Parameters 215

Using the Sequence Clustering Algorithm 216
DMX Queries 217
Model Content 222
Interpreting the Model 222
Summary 227

Chapter 9 Microsoft Association Rules 229
Introducing Microsoft Association Rules 230
Association Algorithm Principles 230
Understanding Basic Association Algorithm Concepts 231
Itemset 231
Support 232
Probability (Confidence) 232
Importance 233
Finding Frequent Itemsets 234
Generating Association Rules 237
Prediction 238
Algorithm Parameters 239
Using the Association Algorithm 240
DMX Queries 241
Model Content 243
Interpreting the Model 244
Summary 246

Chapter 10 Microsoft Neural Network 247
Introducing the Principles of the Microsoft Neural Network Algorithm 247
What Is Neural Network? 248
Combination and Activation 250
Backpropagation, Error Function, and Conjugate Gradient 252
A Simple Example of Processing a Neural Network 254
Normalization and Mapping 255
Topology of the Network 257
Training the Ending Condition 258
Introducing the Algorithm Parameters 258
DMX Queries 259
Model Content 261
Interpreting the Model 262
Summary 264

Chapter 11 Mining OLAP Cubes 265
Introducing OLAP 266
Understanding Star and Snowflake Schema 267
Understanding Dimension and Hierarchy 268
Understanding Measures and Measure Groups 269
Understanding Cube Processing and Storage 270
Using Proactive Caching 271
Querying a Cube 272

Performing Calculations 273
Browsing a Cube 274
Understanding Unified Dimension Modeling 275
Understanding the Relationship Between OLAP and Data Mining 278
Data Mining Benefits of OLAP for Aggregated Data 279
OLAP Needs Data Mining for Pattern Discovery 280
OLAP Mining versus Relational Mining 281
Building OLAP Mining Models Using Wizards and Editors 282
Using the Data Mining Wizard 282
Building the Customer Segmentation Model 283
Creating a Market Basket Model 285
Creating a Sales Forecast Model 288
Using the Data Mining Editor 293
Understanding Data Mining Dimensions 294
Using MDX inside DMX Queries 296
Using Analysis Management Objects for the OLAP Mining Model 297
Summary 301

Chapter 12 Data Mining with SQL Server Integration Services 303
Introducing SSIS 304
Understanding SSIS Packages 304
Task Flow 305
Standard Tasks in SSIS 306
Containers 307
Debugging 307
Exploring a Control Flow Example 307
Data Flow 308
Transforms 309
Viewers 310
Exploring a Data Flow Example 310
Date Mining in SSIS Environment 310
Data Mining Tasks 312
The Data Mining Query Task 312
Analysis Services Processing Task 314
Analysis Services Execute DDL Task 315
An Example of a Control Flow Using Data Mining 316
Data Mining Transforms 316
Data Mining Model Training Transform 316
Data Mining Query Transform 319
Example Data Flows 321
Term Extraction Transform 322
Term Lookup Transform 324
Example of Text Mining Project 326
Summary 327

Chapter 13 SQL Server Data Mining Architecture 329
Introducing Analysis Services Architecture 329
XML for Analysis 330
XMLA APIs 331
Discover 332
Execute 334
XMLA and Analysis Services 335
Processing Architecture 336
Data Mining Administration 337
Server Configuration 337
Data Mining Security 339
Summary 341

Chapter 14 Programming SQL Server Data Mining 343
Data Mining APIs 344
ADO 345
ADO.NET 345
ADOMD.NET 346
Server ADOMD 346
AMO 347
Using Analysis Services APIs 347
Using Microsoft.AnalysisServices to Create and Manage Mining Models 348
AMO Basics 348
AMO Applications and Security 350
Object Creation 351
Creating Data Access Objects 352
Creating the Mining Structure 355
Creating the Mining Models 356
Processing Mining Models 358
Deploying Mining Models 359
Setting Mining Permissions 361
Browsing and Querying Mining Models 362
Predicting Using ADOMD.NET 362
Browsing Models 365
Stored Procedures 368
Writing Stored Procedures 369
Stored Procedures and Prepare 369
A Stored Procedure Example 371
Executing Queries inside Stored Procedures 372
Deploying and Debugging Stored Procedure Assemblies 373
Summary 374

Chapter 15 Implementing a Web Cross-Selling Application 375
Source Data Description 376
Building Your Model 376
Identifying the Data Mining Task 377

Using Decision Trees for Association 377
Using the Association Rules Algorthim 379
Comparing the Two Models 381
Making Predictions 382
Making Batch Prediction Queries 382
Using Singleton Prediction Queries 384
Integrating Predictions with Web Applications 384
Understanding Web Application Architecture 385
Setting the Permissions 386
Examining Sample Code for the Web Recommendation Application 387
Summary 390

Chapter 16 Advanced Forecasting Using Microsoft Excel 391
Configuring Analysis Services for Session Models 392
Using the Advanced Forecasting Tool 392
ExcelTimeSeries Add-In Architecture 394
Building the Input Data Set 395
Creating the XMLA Rowset 396
Converting from Excel to XMLA 396
Building the XMLA Rowset 397
Creating and Training the Mining Model 398
Connecting to the Data Mining Engine 398
Creation and Training 399
ExcelTimeSeriesMining.CreateModel Implementation 399
Forecasting the Series 401
Bringing It All Together 402
Summary 405

Chapter 17 Extending SQL Server Data Mining 407
Understanding Plug-in Algorithms 408
Plug-in Algorithm Framework 408
Plug-in Algorithm Concepts 409
Model Creation and Processing 411
Prediction 412
Content Navigation 413
Managed Plug-Ins 414
Installing Plug-in Algorithms 414
Using Data Mining Viewers 414
Summary 416

Chapter 18 Conclusion and Additional Resources 417
Recapping the Highlights of SQL Server 2005 Data Mining 417
State-of-the-Art Algorithms 418
Easy-to-Use Tools 419
Simple Yet Powerful API 419
Integration with Sibling BI technologies 419
Exploring New Data Mining Frontiers and Opportunities 420

Further Readings 420
Microsoft Data Mining Resources 421
More on General Data Mining 421
Popular Data Mining Web Site 422
Popular Data Mining Conference 422

Appendix A Importing Datasets 423
Datasets 423
MovieClick Dataset 423
Voting Records Dataset 425
FoodMart 2000 Dataset 426
College Plans Dataset 426
Importing Datasets 427

Appendix B Supported VBA and Excel Functions 431

Index 435