Barry de Ville

dp
Digital
Press

Microsoft® Data Mining

Integrated Business Intelligence for e-Commerce and Knowledge Management

DATA MANAGEMENT

Microsoft® Data Mining

Related Titles From

Rhonda Delmater and Monte Hancock, *Data Mining Explained: A Manager's Guide to Customer-Centric Business Intelligence*, ISBN 1-55558-231-1, 352pp, 2001

Thomas C. Redman, *Data Quality: The Field Guide*, ISBN 1-55558-251-6, 240pp, 2001

Jesus Mena, *Data Mining Your Website*, ISBN 1-55558-222-2, 384pp, 1999

Lilian Hobbs and Susan Hillson, *Oracle8i Data Warehousing*, ISBN 1-55558-205-2, 400pp, 1999

Lilian Hobbs, *Oracle8 on Windows NT*, ISBN 1-55558-190-0, 384pp, 1998

Tony Redmond, *Microsoft® Exchange Server for Windows 2000: Planning, Design, and Implementation*, ISBN 1-55558-224-9, 1072pp, 2000

Jerry Cochran, *Mission-Critical Microsoft® Exchange 2000: Building Highly Available Messaging and Knowledge Management Systems*, ISBN 1-55558-233-8, 352pp, 2000

For more information or to order these and other Digital Press titles please visit our website at www.bhusa.com/digitalpress!

At www.bhusa.com/digitalpress you can:

- Join the Digital Press Email Service and have news about our books delivered right to your desktop
- Read the latest news on titles
- Sample chapters on featured titles for free
- Question our expert authors and editors
- Download free software to accompany select texts

Microsoft® Data Mining

Integrated Business Intelligence for e-Commerce and Knowledge Management

Barry de Ville

Digital Press
An imprint of Butterworth-Heinemann
Boston • Oxford • Auckland • Johannesburg • Melbourne • New Delhi

Library of Congress Cataloging-in-Publication Data

de Ville, Barry.
Microsoft® data mining : integrated business intelligence for e-commerce and knowledge management / by Barry de Ville.
p. cm.
Includes index.
ISBN 1-55558-242-7 (pbk. : alk. paper)
1. Data mining. 2. OLE (Computer file) 3. SQL server. I. Title.

QA76.9.D343 D43 2000
006.3--dc21

00-047514

British Library Cataloging-in-Publication Data

A catalogue record for this book is available from the British Library.

The publisher offers special discounts on bulk orders of this book.
For information, please contact:

Manager of Special Sales
Butterworth–Heinemann
225 Wildwood Avenue
Woburn, MA 01801-2041
Tel: 781-904-2500
Fax: 781-904-2620

For information on all Butterworth–Heinemann publications available, contact our World Wide Web home page at: http://www.bh.com.

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America

To Naomi and Gaetan

This Page Intentionally Left Blank

Contents

Foreword **xi**

Preface **xiii**

Acknowledgments **xix**

1 Introduction to Data Mining **1**

1.1 Something old, something new 3
1.2 Microsoft's approach to developing the right set of tools 7
1.3 Benefits of data mining 10
1.4 Microsoft's entry into data mining 18
1.5 Concept of operations 19

2 The Data Mining Process **23**

2.1 Best practices in knowledge discovery in databases 24
2.2 The scientific method and the paradigms that come with it 25
2.3 How to develop your paradigm 30
2.4 The data mining process methodology 37
2.5 Business understanding 39
2.6 Data understanding 41
2.7 Data preparation 44
2.8 Modeling 45
2.9 Evaluation 49
2.10 Deployment 51
2.11 Performance measurement 54
2.12 Collaborative data mining: the confluence of data mining and knowledge management 55

3 Data Mining Tools and Techniques **59**

3.1 Microsoft's entry into data mining 60
3.2 The Microsoft data mining perspective 60

3.3 Data mining and exploration (DMX) projects 64
3.4 OLE DB for data mining architecture 65
3.5 The Microsoft data warehousing framework and alliance 71
3.6 Data mining tasks supported by SQL Server 2000 Analysis Services 72
3.7 Other elements of the Microsoft data mining strategy 86

4 Managing the Data Mining Project 93

4.1 The mining mart 94
4.2 Unit of analysis 95
4.3 Defining the level of aggregation 97
4.4 Defining metadata 98
4.5 Calculations 99
4.6 Standardized values 102
4.7 Transformations for discrete values 103
4.8 Aggregates 103
4.9 Enrichments 111
4.10 Example process (target marketing) 112
4.11 The data mart 115

***5* Modeling Data 117**

5.1 The database 118
5.2 Problem scenario 118
5.3 Setting up analysis services 120
5.4 Defining the OLAP cube 124
5.5 Adding to the dimensional representation 132
5.6 Building the analysis view for data mining 135
5.7 Setting up the data mining analysis 137
5.8 Predictive modeling (classification) tasks 139
5.9 Creating the mining model 141
5.10 The tree navigator 147
5.11 Clustering (creating segments) with cluster analysis 151
5.12 Confirming the model through validation 158
5.13 Summary 159

6 Deploying the Results 163

6.1 Deployments for predictive tasks (classification) 164
6.2 Lift charts 172
6.3 Backing up and restoring databases 175

7 The Discovery and Delivery of Knowledge for Effective Enterprise Outcomes: Knowledge Management 177

7.1 The role of implicit and explicit knowledge 179
7.2 A primer on knowledge management 180
7.3 The Microsoft technology-enabling framework 199
7.4 Summary 208

Appendix A: Glossary 213

Appendix B: References 219

Appendix C: Web Sites 223

Appendix D: Data Mining and Knowledge Discovery Data Sets in the Public Domain 229

Appendix E: Microsoft Solution Providers 255

Appendix F: Summary of Knowledge Management Case Studies and Web Locations 289

Index 301

This Page Intentionally Left Blank

Foreword

The year 1989 seems so long ago! Back in those heady days of the software industry, a chap named Barry de Ville approached me with a view to having my organization license a rudimentary software tool for data mining. At the time, I worked in a large multinational software firm and was responsible for the business side of several mission-critical R&D projects aimed at changing the paradigm of software tools for knowledge workers. Each project involved data modeling and data mining. In the end, after spending many millions of dollars, these projects were either dropped or significantly altered. However, one piece that survived this purge was the software licensed from de Ville. In fact, it went on to become part of the product that changed the company and established a market for desktop analytics.

The business decision to terminate or severely curtail what were once corporate priorities had its roots in the realization that the marketplace, and in particular the high-end business customer in large corporations, was not yet ready for large-scale data mining. Two reasons for this dominated, and both related to the past and not the present. First, there were no generally accepted standards to link nascent mining tools to various data models, and there certainly were no widely used data mining frameworks. Second, there was a general lack of know-how and a poor understanding of analytics in the target user community.

Today, the advent of de facto standards such as OLAP databases and tools such as OLE DB for DM, along with the emergence of data mining frameworks, have firmly established data mining as a viable and important use of computing in business. For example, this capability has been honed into powerful applications such as customer relationship management. This application domain is becoming all the more important with the advent of large-scale databases underpinning e-commerce and e-business.

The second reason for the earlier failure had much more to do with the receptor capacity of the marketplace than with the vendor community's

ability to deliver appropriate tools. With the vast majority of organizations seeing the database only in terms of a relational model, the concept of applying multidimensional analytics to corporate data was little more than a dream. Consequently, the second key to opening the data mining market has been the spread of know-how. In the workplace this know-how is primarily supplied through widely available information in the trade press and commercial computer-related publications.

The decision by Microsoft Corporation, as early as 1998, to become a major player in the data mining arena set the stage for things to come. Today's coupling of the latest data mining capabilities with SQL Server 2000 has created a clear and present need to capture and consolidate in one place the principles of data mining and multidimensional analytics with a practical description of the Microsoft data mining architecture and tool set. This book does just that.

Recognizing the receptor problem and the power and ease of use of the new Microsoft data mining solution has afforded Barry de Ville with the opportunity to help redress receptor capacity by writing this practical guidebook, which contains illustrative and illuminating examples from business, science, and society. Moreover, he has taken an approach that compartmentalizes concepts and relationships so that the reader can more readily assimilate the content in terms of his or her own general knowledge and work experience, rather than dig through the more classical formalisms of an academic treatise.

Peter K. MacKinnon
Managing Director
Synergy Technology Management
e-mail: petemac@istar.ca
telephone: (613) 241-1264

Preface

Data mining exploits the knowledge that is held in the enterprise data store by examining the data to reveal patterns that suggest better ways to produce profit, savings, higher-quality products, and greater customer satisfaction. Just as the lines on our faces reveal a history of laughter and frowns, the patterns embedded in data reveal a history of, for example, profits and losses. The retrieval of these patterns from data and the implementation of the lessons learned from the patterns are what data mining and knowledge discovery are all about.

This book will appeal to people who have come to depend upon Microsoft to provide a high-performance and economical point of entry for an ever-increasing range of computer applications and who sense the potential value of pursuing data mining approaches to support business intelligence initiatives in their enterprises. Traditional producers and consumers of business intelligence products and processes, especially OLAP (On-Line Analytical Processing), will also be attracted by this information. Most business intelligence vendors, especially Microsoft, recognize that business intelligence and data mining are different facets of the same process of turning data into knowledge. SQL Server 7, released late in 1998, introduced SQL Server 7 OLAP services, thus providing a built-in OLAP reporting facility for the database. In the same manner, SQL Server 2000 provides built-in data mining services as a fundamental part of the database. Now, both these important forms of business reporting will be available as core components of the database functionality; further, by providing both sets of facilities in a common interface and platform, Microsoft has taken the first step in providing a seamless integration of the various methods and metaphors of business reporting so that one simple, unified interface to the knowledge contained in data is provided. Whether that knowledge was delivered on the basis of an OLAP technique or data mining technique is irrelevant to most users, and now it will be irrelevant in a unified SQL 2000 framework.

This book will emphasize the data mining aspects of business intelligence in order to explain and illustrate data mining techniques and best practices, particularly with respect to the data mining functionality that is available in the new generation of Microsoft business intelligence tools: the new OLE DB for DM (data mining) and SQL Server functions. Both OLAP and data mining are complex technologies. OLAP, however, is intuitively easier to grasp, since the reporting dimensions are almost always business terms and concepts and are organized as such. Data mining is more flexible than OLAP, however, and the patterns that are sought out in data through data mining are often counter-intuitive from a business standpoint. So, initially, it can be more difficult to conceptualize data mining. A core goal of this book is to help all users to move through this conceptualizational task in order to reap the benefits of an integrated OLAP and data mining framework.

Discovering successful patterns that are contained in data, but that are normally hidden, can be a formidable challenge. For example, take gross margins in a retail sales data store. Here we see that the margins fluctuate over the course of a year. A plot of the values held in the gross margin field in the data store might reveal a 10 percent increase in gross margin between summer and fall. We might be tempted to conclude that sales margins increase as we move from summer to fall. In this case we would say that the increase in gross margin depends upon the season.

But there are many other potential dependencies, which could influence gross margin, that are locked in the data store. Along with the field *season* are other fields of data—for example, quantity sold, discount rate, commission paid, customer location, other purchases made, length of time as a customer, and so on. What if the discount rate is greater in the summer than in the fall? Then, possibly, the increase in gross margin that we see in the fall is simply a result of a lower discount rate. In this case gross margin does not vary by season at all—it varies according to the discount rate! In this case the apparent relationship, or dependency, that we observed between season and discount rate is a spurious one. If we adjust our view of gross margins to remove the effect of discount rate, then maybe we would find that, actually, gross margins would be higher in the summer. So, in order to do a thorough job of data mining and knowledge discovery it is essential to look at all potential explanatory factors and associated data elements to ensure that the very best pattern is retrieved from the data and that no spurious, and potentially misleading, effects are introduced into the patterns that we select.

What if the data store could be manipulated so that all of the dependencies that affect the questions we are looking at (e.g., gross margin) could be

considered together? What if we could search through all the combinations of dependencies and find a unique combination, or pattern, that isolates a particular combination of events that maximizes the gross margin? Then, instead of simply showing the effect of one condition, say season, on gross margin, we could show the combined effect of a pattern, say a particular time, location, and discount rate, that produces the maximum gross margin. Once we have isolated this optimal pattern, we have a particular gem of wisdom, since, if we can reproduce that pattern more often in the future, we can establish a strategy that will systematically increase our gross margin and associated profitability over time.

There is no lack of data in the modern enterprise. So the raw material for data mining and knowledge discovery is abundantly available. The data store contains records that have the potential to reveal patterns of dependencies that can enrich a wide variety of enterprise goals, missions, and objectives. Retail sales can benefit from the examination of sales records to reveal highly profitable retail sales patterns. Financial analysts can examine the records of financial transactions to reveal patterns of successful transactions. An engineering enterprise can search through its records surrounding the engineering process—manufacturing time, lot size, assembly parameters, and operator number—to determine the combination of data conditions that relate to the quality measure of the device coming off the assembly line. Marketing analysts can look at the marketing data store to detect patterns that are associated with market growth or customer responsiveness.

The data are freely available and the pay-offs are enormous: the ability to decrease inventory, increase customer buying propensity, drive product defects detection closer to the assembly line, and so on by as little as 1 percent represents a truly staggering, Midas-like fortune in the billion-dollar-a-day industries of finance, manufacturing, retail services, and high technology. The key to reaping the rewards of data mining is to have a cost-effective set of tools and body of knowledge to undertake the knowledge discovery.

Until recently the tools that were available to accomplish this task were relatively rare and relatively expensive. Business intelligence OLAP facilities have become much more commonplace but, as demonstrated above, business intelligence OLAP tools may not find all the patterns and dependencies that lie in data. For this, a data mining tool is required.

Microsoft recognized this requirement after the release of SQL Server 7 and began a development program to migrate data mining and knowledge discovery capabilities into the SQL Server 2000 release. This release, and

the associated data mining and knowledge discovery tools, techniques, concepts, and best practices, are reviewed here. The primary task will be to explain data mining and the Microsoft data mining framework. The chapters are as follows:

1. Introduction to Data Mining: its relevance and utility to 2000-era enterprises and the role of Microsoft architecture and technologies. This chapter provides a big-picture view of data mining: what it is, why it is useful, and how it works. What are the barriers to the adoption of data mining and what is Microsoft doing about these barriers? This covers the Microsoft Socrates project and the directions that Microsoft will pursue in data mining in the future.

2. The Data Mining Process: This chapter discusses the process of using data to model and reflect real-world events and activity: the interoperation of measurement, data and business models, and conceptual paradigms to reflect real-world phenomena. Testing and refining the models—patterns, structure, relationships, explanation, and prediction—are also discussed. Best practices in executing the data mining mission, such as business goal, ROI outcome identification, the conceptual model, operational measures, data elements, data transformation, data exploration, model development, model exploration, model verification, and performance measurement, are addressed in depth.

   Chapter 2 also discusses the following topics:

   - ROI and the choice of an appropriate business objective
   - Creating a seamless business process for data mining
   - Closed-loop processes

   You can't manage what you can't measure—the role of performance measurement and campaign management for continuous improvement in data mining is explained in this chapter.

3. Data Mining Tools and Techniques (and the associated Microsoft data mining architecture): revealing structure in data—profiling and segmentation approaches, and predictive modeling—applications and their lifetime value optimization through profitable customer acquisition. Data mining query languages and the integration with OLAP, OLE DB for DM, and scaling to large databases are explained in detail. Leveraging the Microsoft architecture—how developers and users can leverage the Microsoft