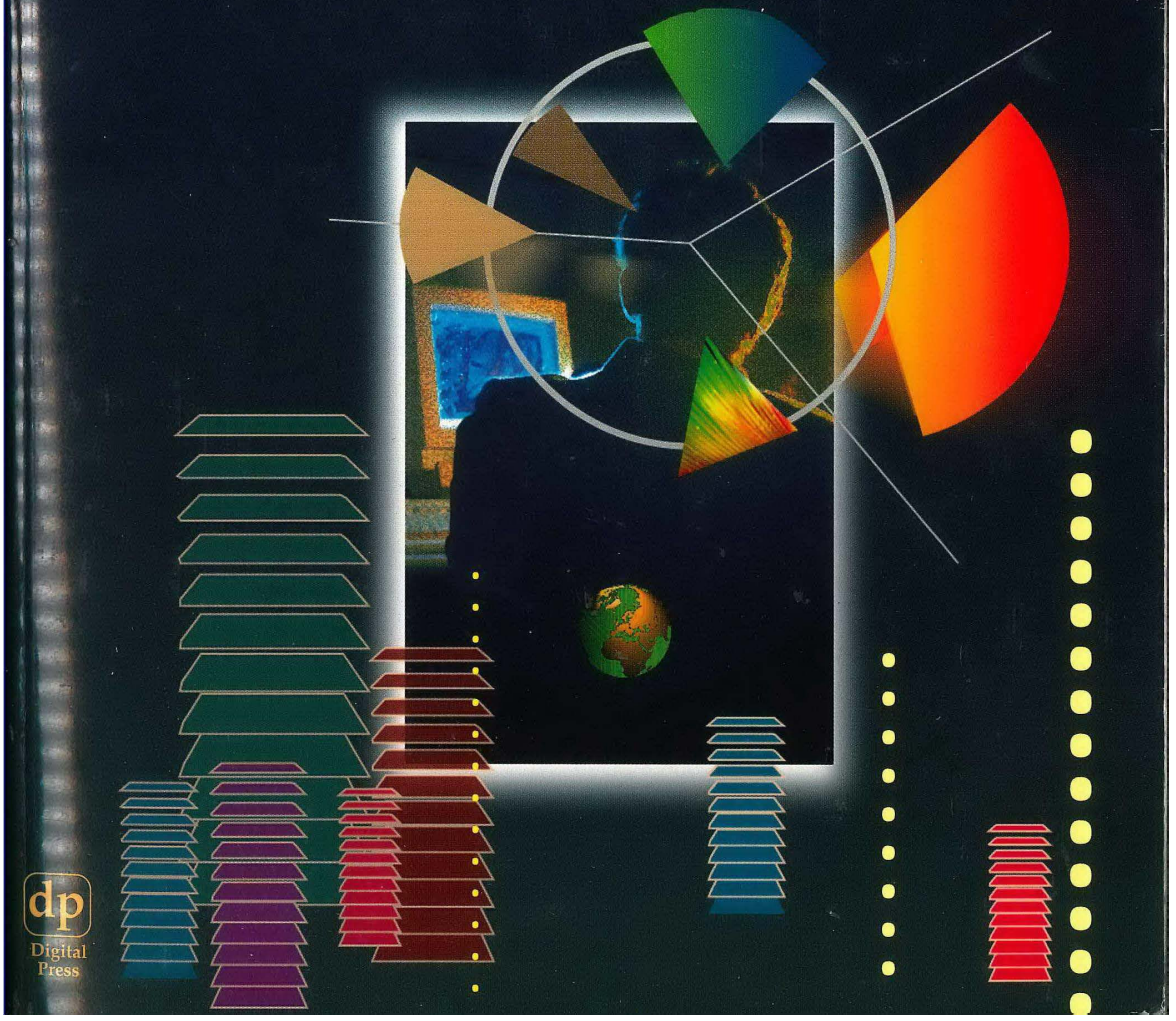
JESUS MENA

Data Mining Your Website



dp
Digital Press

Amazon/AlmondNet
Exhibit 1011
Page 1


MENA
Data Mining Your Website
dp
Digital Press

Catherine Guidry
4208 Lowell Dr.
Baltimore MD 21208-6027

Data Mining Your Website

Amazon/AlmondNet
Exhibit 1011
Page 3

Catherine Guidry
4208 Lowell Dr.
Baltimore, MD 21208



Data Mining Your Website

Jesus Mena



Boston • Oxford • Auckland • Johannesburg • Melbourne • New Delhi

Amazon/AlmondNet
Exhibit 1011
Page 4

Digital Press is an imprint of Butterworth–Heinemann.

Copyright © 1999 by Butterworth–Heinemann

A member of the Reed Elsevier group

All rights reserved.

All trademarks found herein are property of their respective owners.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

Recognizing the importance of preserving what has been written, Butterworth–Heinemann prints its books on acid-free paper whenever possible.

Butterworth–Heinemann supports the efforts of American Forests and the Global ReLeaf program in its campaign for the betterment of trees, forests, and our environment.

Library of Congress Cataloging-in-Publication Data

Mena, Jesus.
Data mining your website / Jesus Mena.
p. cm.
ISBN 1-55558-222-2 (alk. paper)
1. Information technology—Management. 2. Data mining.
3. Internet marketing. 4. Business enterprises—Computer networks.
5. Web sites. I. Title.
HD30.2.M455 1999
006.3—dc21 99-22694
CIP

British Library Cataloguing-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 Digital Press publications available, contact our World Wide Web home page at: http://www.bh.com/digitalpress

10 9 8 7 6 5 4 3 2

Printed in the United States of America

To Megs

Amazon/AlmondNet
Exhibit 1011
Page 6

Contents

CHAPTER 1 MINE2.COM **1**

The Idle Portal 1
Mining Data 4
Nothing New 7
Mining Online 8
How Data Mining Answers Key Business Questions 10
Classification/Prediction 10
Segmentation 11
Association 13
Clustering 15
Visualization 15
Optimization 17
Data Mining Drives Relational Marketing 19
Web Privacy 20
Benefits and Incentives 21
Ten Steps to Mining Your Web Data 22
A Bookseller Example 25
Look at the Data 26
Split the Data 28
Model the Data 30
Integrate the Solution 32
Compiling an Identity 33
The Architecture for Analysis 35
Personalized Future 36
Why Bother? 39

Amazon/AlmondNet
Exhibit 1011
Page 8

CHAPTER 2 WHAT IS DATA MINING? 41

It's about Predators and Patterns 41
A Definition 42
Drowning in Data 43
The Complexity of Loyalty 45
Traditional Data Mining Applications 46
New Data Mining Applications 48
Statistics 48
Data Mining vs. Statistics 50
Online Analytical Processing 51
Data Mining vs. OLAP 52
Website Analysis Tools 54
Website Analysis Tools vs. Data Mining Tools 61
Visual Data Mining 65
Visualization Packages 65
Link Analysis Systems 66
Quantitative Display 67
Data Mining Solutions 68
Why and Who, Not What 71
The Online Corner Store 72
Selling as a Service 73
Web Marketing 74
Online Coupons 76
Mining Customer Value 77

CHAPTER 3 NETWORKS AND ALGORITHMS 83

They're about Learning and Evolving 83
The AI War 84
The Top-Down Approach 84
The Bottom-Up Approach 85
Neural Networks 86

Amazon/AlmondNet
Exhibit 1011
Page 9

From Slugs to VISA 87
Back-Propagation Networks 88
Supervised Learning 90
Training a Network 91
Preparing Your Data 92
Webbing Your Features 93
Purchasing Propensity 94
Training Cycle 95
Self-Organizing Maps (SOM) 95
Calibrating Clusters 96
Neural Network Tools 97
Machine-Learning Algorithms 98
A Measure of Information 99
The Rules 100
The Trees 102
The Trees vs. the Rules 103
Automatic Segmentation 104
Machine Learning Your Site 105
Machine-Learning Tools 106
Genetic Algorithms 107
Code of Survival 107
Reproduction 108
Crossover 109
Mutation 110
A Step-by-Step Example Using a Binary-Coded GA 110
Data Mining with Genetic Algorithms 112
Genetic Algorithm Applications 113
The AI Cousins: Case-Based Reasoning and Fuzzy Logic 114
Case-Based Reasoning (CBR) 114
Fuzzy Logic 115
Conclusion 117

Amazon/AlmondNet
Exhibit 1011
Page 10

CHAPTER 4 TEN STEPS TO MINING YOUR DATA 119

Identify Your Business Objective 119
Baseline 121
Select Your Data 123
Creating Customer Data 124
Brand Sites 124
Banners 126
Buttons 127
Pop Ups 128
Push 129
E-mail 129
Chat 130
Agents 130
Dealing with Transactional Data 131
Prepare Your Data 132
Enhancing Your Data 134
Evaluate Your Data 134
Utility Tools for Preparing and Viewing Your Data 134
Format the Solution 135
Select the Tools 136
Scalability 137
Accuracy 137
Formats 138
Solutions 139
Preprocessing 139
Connectivity 140
Import/Export 140
Memory Management 141
Performance 141
Noise 142
Paradigms 142
Efficiency 143

Amazon/AlmondNet
Exhibit 1011
Page 11

Tool Selection 143
A Tool Shopping Tip: Kick the Tires before Buying 145
Construct the Models 145
Validate the Findings 147
Deliver the Findings 148
Integrate the Solutions 149
Integrating Your Website and Data Warehouse 149

CHAPTER 5 THE TOOLS 155

Criteria 155
Single Paradigm Tools 159
Database Mining Marksman 159
4Thought 161
KnowledgeSEEKER 163
NGO NeuroGenetic Optimizer 168
WizWhy 170
The Data Mining Suites 173
Clementine 174
DataEngine 178
Decisionhouse 181
Scorecard Builder 182
Other Tools 184
Final Tool Tip 188

CHAPTER 6 THE DATA COMPONENTS 193

Server Log Files 194
Transfer Log File 194
Host Field 195
Identification Field 195
Authuser Field 196
Time Stamp 196
HTTP Request 196

Amazon/AlmondNet
Exhibit 1011
Page 12

Status Code Field 196
Transfer Volume 197
Error Log 197
Referrer Log 198
Agent Log 198
Extended Log Format 198
Configuring Your Server 199
Server Configuration Tips 199
Apache and NCSA Server Configurations 200
CERN Server Configurations 201
Netscape Server Configurations 201
WebSite Server Configurations 202
The Multiplicity of a Visit 202
Cookie Files 203
A Cookie Future 206
Cookie Applications 207
Configuring Cookies 210
Cookies and CGI 211
Open Profiling System (OPS) 213
Forms 218
Common Gateway Interface 221
Assembling the Components 223
Components Considerations 225

CHAPTER 7 THE PROVIDERS 227

Marketing Data Providers 228
Applied Geographic Solutions 228
Acxiom 229
CACI (ACORN) 233
Claritas 238
Equifax 239
Experian 241

Amazon/AlmondNet
Exhibit 1011
Page 13

Polk 242
Trans Union 246
Web Services and Software Providers 247
Art Technology Group 248
Aptex 249
Autonomy 250
BroadVision 250
DoubleClick 252
Engage Technologies 253
Firefly 255
GuestTrack 256
Gustos 257
Internet Profiles Corporation (I/PRO) 257
IMGIS 258
InText 259
LikeMinds 259
MatchLogic 260
MicroMass 261
NetPerceptions 262
Vignette 263
Your Data Warehouse 264
Closing the Loop 265

CHAPTER 8 E-RETAILING 267

The Product Is the Message 269
Sale Segmentation 272
The Digital Marketplace 275
Product Customization 277
Consumer Profiling 279
Real-Time Buying 282
Organize Your Storefront 288
Retailing and Mining 290
Electronic Commerce 294

Amazon/AlmondNet
Exhibit 1011
Page 14

CHAPTER 9 E-MINING 299

Mining Your Site 299
Capturing the Look 301
A Sample Website Session 303
Interactive by Design 305
Solid State 305
The Data Set 308
Clustering (Association and Sequencing) 310
Segmentation 318
Classification and Prediction 325
Optimization 332
Visualization 335
It's an Integrated Iterative Process 339
Data Mining Future 345
The Personalization of Self 346
Right Turns, Hand Grenades, and Integrity 347

APPENDIX Privacy Consortiums, Standards, and Legislation 349

Glossary 355

Index 361

Amazon/AlmondNet
Exhibit 1011
Page 15

1

mine2.com

The Web is like an organism made up of millions of cells that are all connected, intertwined, and communicating. It is a massive worldwide creature that is continuously evolving, spawning, and changing; since its birth, it has already mutated into a new type of species, where millions of computers and consumers connect to every company in the world and their product inventory databases. In this organic environment of constant ferment, the consumers drive product availability, features, and pricing. Electronic retailing, which is supported and nurtured in this environment itself, mimics the mechanisms of a single-cell organism—adapting to the needs, tastes, habits, and preferences of its customers. Billions of business transactions flow and evolve, subtly transforming consumers and retailers in this dynamic marketplace and the relationships between them.

This book is about biological-based business practices—processes and tools that mimic biological functions. The mining of website transactional data with artificial intelligence-based tools (programs designed to mimic human functions) is an attempt to recognize, anticipate, and learn the buying habits and preferences of customers in this new evolutionary business environment. It is of paramount importance that the retailer in a networked economy such as this be adaptive and receptive to the needs of its customers. In this expansive, competitive, and volatile environment, data mining will be a critical process impacting a retailer's long-term success, where failure to quickly react, adapt, and evolve can translate into customer attrition in the click of a mouse.

The Idle Portal

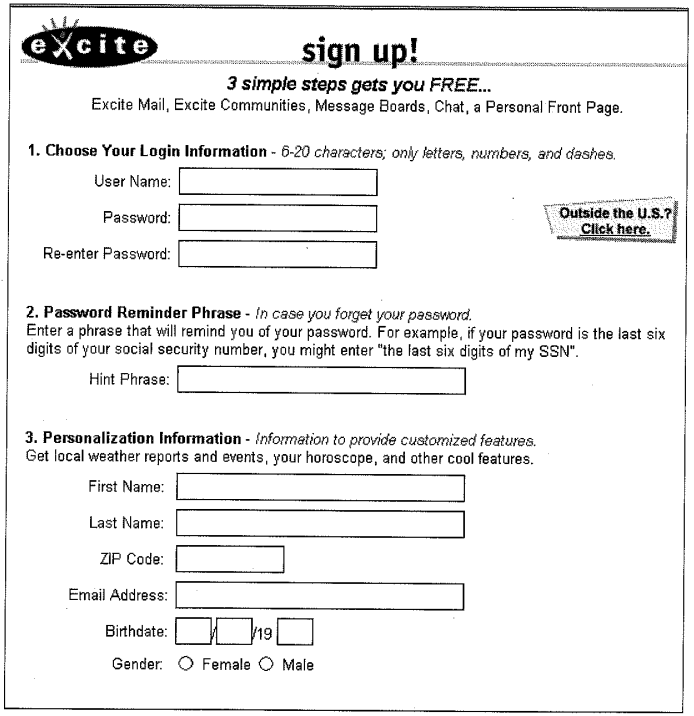
My decision to write this book stems in part from a series of e-mails I had with a marketing director from one of the major search engine companies. Like its competitors, this Internet company was expanding its services and products at a frantic pace. It was growing and

Amazon/AlmondNet
Exhibit 1011
Page 16

moving at "Internet time," as they say, through a dizzying array of acquisitions, deals, partnerships, alliances, and redesigns. However, in its rush to expand its share of the marketplace and stay ahead of its competitors, it was failing to leverage one of its most important weapons—its knowledge about its visitors. As our e-mail exchange progressed, the magnitude of the problems facing most websites, and how data mining could assist companies on the Internet in coping with the millions of customer interactions they were accumulating on a daily basis, became clear to me.

This portal, like dozens of its rivals, was making an effort to personalize the content it provided to its visitors through its "myportal.com" feature (see Figure 1–1). In doing so it was soliciting important personal information from them via registration forms, which was by itself generating valuable databases for mining. More important, this same database could provide the foundation for obtaining additional demographic and household information for detailed profiles about visitor preferences, lifestyles, and consuming habits.

Figure 1–1A *The personalization of portals.*



Amazon/AlmondNet
Exhibit 1011
Page 17

Figure 1–1B
The personalization of portals.

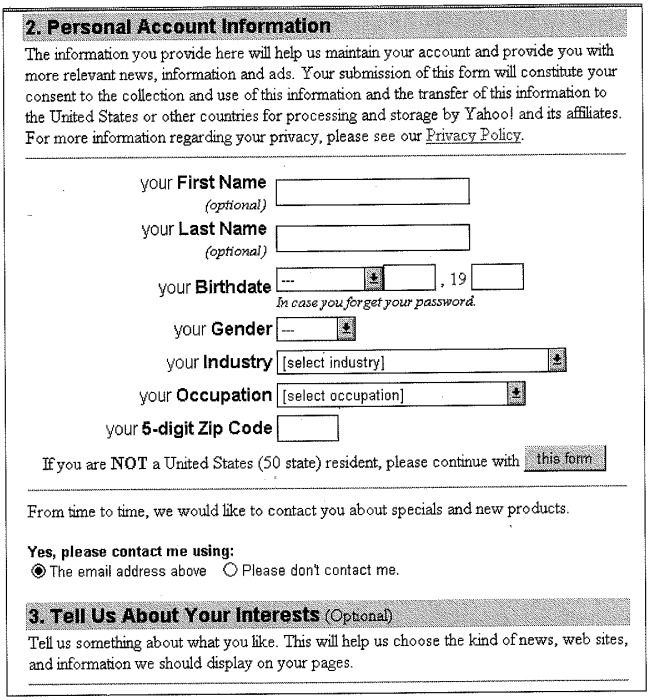
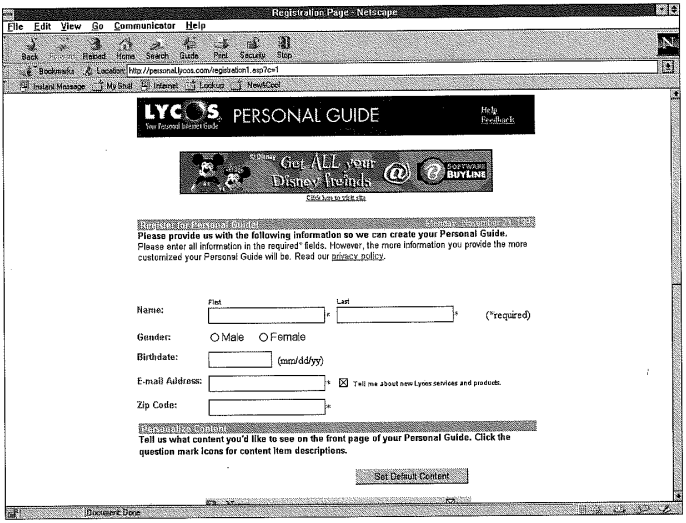


Figure 1–1C
The personalization of portals.



Amazon/AlmondNet
Exhibit 1011
Page 18

However, this search engine company really did not know who it was attracting or who was leaving its website. It did not know who its most valuable visitors were, or why. Like many major and small websites, this Internet company was discarding the valuable intelligence it was gathering every hour of every day—it was not mining the information it was accumulating from its millions of visitors. Although it was moving in the direction of data mining (due to pressure from advertisers), it did not know the characteristics and behavior of its visitors. It was placing ads in front of visitors based on the keywords they entered, rather than observed patterns modeled from data mining. It was a hardwired solution to an application ideally suited to modeling of observed human behavior.

Mining Data

Data mining, in a nutshell, is inductive data analysis. When data is too large and complex to be examined by humans, a summation in the form of a ratio or a formula can often reveal a pattern. The voluminous amounts of data generated from a website, for example, often hide patterns that reveal conditions when visitors are likely to make purchases or click on certain ads or banners.

IF	domain	AOL	
AND	keyword	broncos	
AND	gender	male	
AND	age	18–49	
THEN	ad	ESPN	68%
		AUTOTEL	18%
		Amazon	9%
		Microsoft	5%

Data mining software can reveal how the value of one field in a database is affected by the value of other fields. For example a field such as "total purchases" may be affected by other values or fields in your logs or forms database, such as the visitor's age, gender, referral engine, or keywords he or she used. In this example, "total purchases" is the field to be explained or predicted (that is, it is the *dependent variable* or *output* of your analysis) while the other fields are the inputs to your data mining model (the *conditions* or *independent variables*):

Amazon/AlmondNet
Exhibit 1011
Page 19

Value to Predict	Values or Conditions to Consider
Total Purchases (Output)	*Age, Gender, Engine, Keywords*
1	32, Female, Excite, children's software
2	43, Male, Yahoo, math software
3	53, Male, Infoseek, software

Data mining software enables us to relate the field to predict (in the above table, "total purchases") with the other fields in the database. The output from the data mining analysis can detail this relationship in the form of a decision tree diagram, IF/THEN rules, or a mathematical formula. The score obtained from this analysis can be used to grade new visitors based on their propensity to make purchases equal to the amount modeled via the data mining analysis.

Data mining, then, can be defined as the iterative process of detecting and extracting these patterns from large databases: it's *pattern-recognition.*

Data mining lets us identify "signatures" hidden in large databases, as well as learn from repeated examples. It lets us create models for the detection of, say, future potential customers likely to make a large number of total purchases. Data mining involves various solutions, techniques, and technologies, sometimes in combination, such as predictive modeling and customer profiling, arrived at via various techniques including association, classification, clustering, estimation, optimization, segmentation, sequencing, prediction, and visualization.

For the portal company, for example, data mining could be used to predict which visitors will click on what ads or banners by constructing a model based on observations of prior visitors who have behaved both positively and negatively to the advertisement banner. Another end product of data mining is an analysis of these clickthroughs that discovers and describes the factors (e.g., connect time) and the attributes of respondents (e.g., gender) who responded to an ad. This can provide you a profile of who is likely to respond to future ads and assist you in the design of future banners and campaigns.

Data mining is *not* query- or user-driven, nor is it a cumulative traffic report of hits to your site. Data mining is instead driven by the need to uncover the hidden undercurrents in the data, such as the "features" of the visitors generating the hits to your site. So far most analyses of websites data have involved "traffic reports," of which most are geared toward providing cumulative accounts of server TCP/IP-specific browser-to-server activity. Most server analysis

Amazon/AlmondNet
Exhibit 1011
Page 20

packages lack the ability to provide any true business insight about customer demographics and online behavior, although some are moving in that direction as electronic commerce heats up.

Current traffic analysis tools, like Accrued, Andromedia, HitList, NetIntellect, NetTracker, and WebTrends, are geared at providing high-level predefined reports about domain names, IP addresses, browsers, cookies, and other machine-to-machine activity. These software packages originated from the need to report on the activity of the *server* and not on the behavior of its *visitors*. These server activity reports simply do not provide the type of bottom-line analysis that retailers, service providers, marketers, and advertisers in the business world have come to demand.

Data mining obviously requires the strategic use of a special type of software: one that is based on artificial intelligence technologies, such as self-organizing or back propagating neural networks or genetic and machine-learning algorithms. It may involve using the output from these tools with graphs and maps or incorporating code into production systems or automated e-mail generators. However, data mining is much more than its software tools; it is a *process* that involves a set of methodologies and procedures for extracting and preparing the data and then incorporating the solutions into your business and website. Data mining, for example, can involve the strategic capturing of visitor information via registration forms, merging of this online data with additional demographics, and then analyzing it with powerful pattern recognition software for the creation of predictive models. The formats of data mining are as varied and diverse as the multiple approaches and technologies. What you get out of data mining will to a large extent depend on what you require from it: it may be executable C code, SQL syntax, a 3D graph or map, a decision trees, or a set of rules. The end product of data mining, however, is *insight* about the identity and preferences of your online customers.

Most often, data mining is about providing business and marketing solutions using a combination of resources and techniques—a combination driven by not just the data mining tools you use, but also your IT resources, the analysts' skills, the quality of the data available for the analysis, and the support of the marketing and web team members. Data mining often requires developing some strategic techniques for preparing your server data and enhancing its value. Most of the data in today's systems, such as web servers, were designed for transaction processing, not for analysis. In the mining of website databases you will be involved in working with structured data, such as logs files and databases created from registration or purchase

Amazon/AlmondNet
Exhibit 1011
Page 21

forms. You will most likely be working with either relational tables or flat text files, not with unstructured data like sound, video, paper, and digital feeds—the type of data with which the analysts in Langley, Virginia, must wrestle.

Nothing New

Today the goal for marketers is to know and serve every customer, one at a time, and to build long-term, mutually beneficial relationships. Data mining is the key to this customer knowledge and intimacy. For years marketers have been using databases to get a picture of their customers. Using data mining tools they can also compile a composite profile to predict which of their customers are most likely to buy a given product and service or respond to certain communication. Artificial intelligence (AI) technology in the form of data mining has been in use by:

- Cellular phone companies, to stop "churn" (customer attrition)
- Financial services firms, for portfolio and risk management
- Credit card companies, to detect fraud and set pricing
- Mail catalogers, to lift their response rates
- Retailers, for market basket analyses

These companies typically store their customer information in massive data warehouses where it is used for business intelligence, decision support, and relational marketing. To improve on the quality of their internal database, these firms have been merging their internal customer and transactional data with demographic and household information purchased from data resellers like Acxiom, Equifax, Metromail, Polk, and others. They do this in order to enhance their knowledge about their customers' lifestyles and to find out what kind of products and services they consume. Through the use of powerful pattern-recognition technologies incorporated in today's data mining suites, these firms have been attempting to anticipate their customers' behavior—will they respond, will they flee, will they pay, and (most often), will they *buy*, and if so, what, when, and where.

One of the world's largest data mining applications is that of Wal-Mart, which individually profiles every one of their 2,900 stores 52 weeks a year for product demands on over 700 million unique store/item combinations. Through data mining Wal-Mart is able to anticipate demand and thus reduce overhead, inventory costs, and stock. It is able to position the preferred type of mouthwash and dog

Amazon/AlmondNet
Exhibit 1011
Page 22

food in front of the right consumers. This physical retailer uses NeoVista data mining software to be able to anticipate their customers' demands very accurately, similar to the way electronic retailers like Amazon are beginning to want to know how to position the right products online and manage their inventory in the back-end more effectively.

Depending on the type and detail of customer information a company has, such as ZIP code, date of birth, prefix, etc., additional information can be appended from data resellers and modeling vendors. For example, industry-specific scores can be purchased from such companies as Fair, Isaac, or Experian, which assign risk or value to individuals. Banks, insurance, and credit card companies commonly use these scores in determining what product offers to make to their clients or potential prospects. Coupled with enhanced customer information, these firms are using data mining technologies and techniques to drive their decisions and knowledge about their customers. Of course there are other data mining applications in use, such as quality control, monitoring production processes, and law enforcement, but the bulk of the applications (as with the mining of websites) are customer-centric, in which client behavior and knowledge of their preferences is the golden nugget being sought.

Mining Online

Websites today find themselves competing for customer loyalty and dollars in markets where it costs little for customers to switch to competitors. The electronic commerce landscape is evolving into a fast, competitive marketplace where millions of online transactions are being generated from log files and registration forms every hour of every day, and online shoppers browse by electronic retailing sites with their finger poised on their mouse, ready to buy or click on should they not find what they are looking for—that is, should the content, wording, incentive, promotion, product, or service of a website not meet their preferences. In other words, it is a market where browsers are attracted and retained based on how well the retailer remembers their needs and whims.

In such a hyper-competitive marketplace, the strategic use of customer information is critical to survival. As such, AI in the form of data mining has become a mainstay to doing business in fast-moving crowded markets. In accelerated markets the margins and profits go to the quick and responsive players who are able to leverage predic-

Amazon/AlmondNet
Exhibit 1011
Page 23

tive models to anticipate customer behavior and preferences. Data mining of customer information is required in order to make decisions about which clients are the most profitable and desirable and what their characteristics are in order to find more customers just like them—the type of customer profiles and business knowledge electronic retailers and advertisers are beginning to expect from the Web after years of heavy investments and marginal ROIs (returns on investment).

The savviest retailers utilize sophisticated data mining software that is powered by genetic algorithm (GA) technology to optimize their inventory delivery systems based on consumer preferences and profiles. GAs are programs that replicate the process of evolution (selection, crossover, and mutation) in their search for an optimized solution from a given set of items. For example, GAs in combination with neural networks are being used by some retailers to not only control what inventory is placed in what store, but also design the stores themselves, leading to changes in everything from shelf heights to parking spaces. So if GAs can be used to optimize the design and layout of physical stores, such as a Wal-Mart or Sears, think what they could do to optimize the design of a large electronic retailing website like Amazon.

Several varied attempts are being made to personalize the browsing experience of website visitors, including collaborative filtering (Firefly) and the aggregate pooling of cookies through ad networks (DoubleClick). However, few companies involved in electronic retailing are looking at data mining technology for the analysis, modeling, and prediction of customer behavior. Some major portals are using software from Aptex, which is a spin-off from HNC, the world's largest neural network company, to position the appropriate banner and ad in front of "right" visitors based on Aptex's proprietary text and neural net technology. The cost and maintenance of this sophisticated type of software, however, is too steep for most commercial websites.

In the end, the mining and identification of your visitors' profiles can provide you an insight into what type of messages, banners, ads, offers, incentives, products, and services you want to place in front of them. Through the analysis of your website data—after it has been commingled with household and demographic information—your firm can begin to compile a profile of your potential future clients. The mining of your web data is also a strategic necessity for creating a lasting relation with your current customers and establishing a profitable online storefront. As previously mentioned, data mining often

Amazon/AlmondNet
Exhibit 1011
Page 24

involves a combination of technologies that are able to provide assorted solutions and formats.

How Data Mining Answers Key Business Questions

Classification/Prediction

Who will buy, what will they buy, and how much will they buy? These types of business questions can be answered via data mining. They involve classification and prediction (which is a form of classification into the future) and can best be accomplished through the use of neural networks and genetic algorithms (see Figure 1–2). This type of inductive data analysis can be used to predict who will buy or how much they are likely to buy.

Neural and polynomial networks can be used to construct predictive models in order to anticipate your customers' behavior and their propensity to respond to ads, banners, and offers and (of course) their propensity to make online purchases. These predictive models are constructed out of observed behavior of current visitors and customers and are designed to predict the behavior of future visitors. As with fraud-detection models created with neural networks, a web-based model is trained with a sampling of positive (Sale) and negative (No Sale) customer accounts. A back-propagation neural network will recycle through thousands of samples until it learns to recognize the features and patterns made by the visitors to your web-

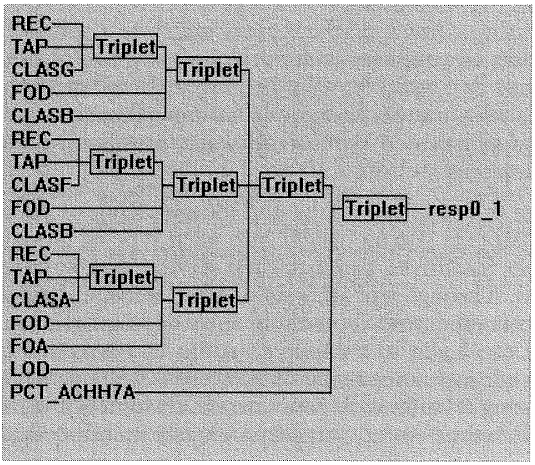


Figure 1–2 *View of a network with multiple inputs designed to predict a response.*

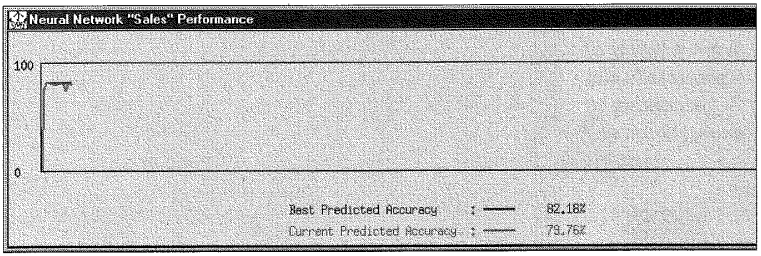


Figure 1–3 *Data mining instrument monitoring the accuracy to a model.*

site who made purchases or clicked through certain ads or banners. (See Figure 1–3.)

The end product of neural network analyses is usually C code, which can be used to target future visitors to your site via e-mail with product-specific offers. The C code itself is a set of weights, or a formula, that uses the inputs of the model to compute a projected output or score. This score is usually in the form of a binary output such as "0 = will not buy, or 1 = will buy." The output can also be a continuous value, such as the projected amount of purchases a visitor is likely to make. Here is an example of C code that can be extracted and used to predict propensity to purchase models:

```
double *purchase(double inarr[])
{
  /* visitor_age—Triplet */
  visitor_age  = 0 - 22.0566 - 40.4871*node5 - 18.3542*pow2(node5)
        + 0.174813*pow3(node5) - 68.848*node31
        - 156.303*node5*node31 - 83.2704*pow2(node5)*node31
        - 44.5781*pow2(node31) - 37.7005*node5*pow2(node31)
        - 0.0442751*pow3(node31) - 64.3699*node32
        - 54.3879*node5*node32 - 185.61*node31*node32
        - 156.467*node5*node31*node32 -
        - 0.109251*pow2(node32) - 0.272628*node31*pow2(node32)
        - 0.00897233*pow3(node32) ;
```

Segmentation

What are the different types of visitors to your website? Segmentation analysis using machine-learning algorithms can break down visitors to a website into unique groups with individual behavior. (See Figure 1–4.) These stratified groupings can then be used to make statistical projections, such as the potential amount of purchases they are likely to make.

Amazon/AlmondNet
Exhibit 1011
Page 26

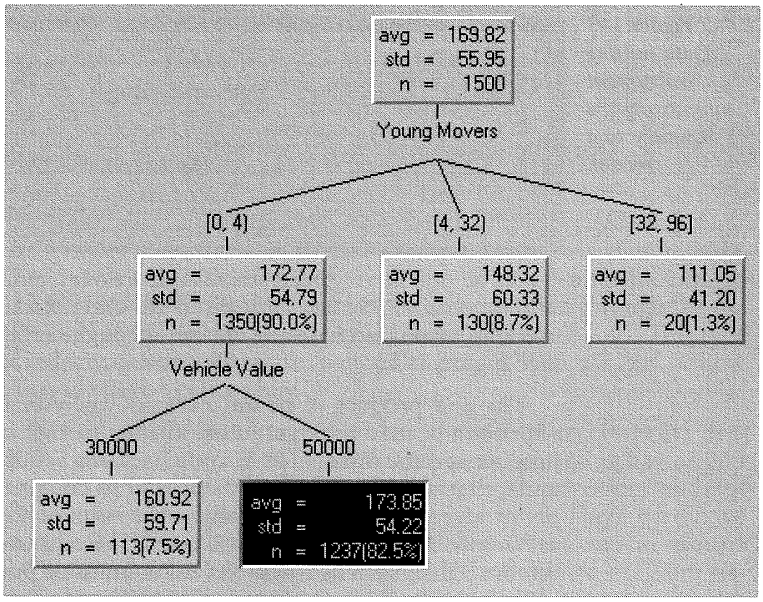


Figure 1–4 *Segmentation of projected sales according to demographics of visitors.*

Business rules can also be extracted directly from the web data during this type of segmentation analysis:

IF	search keyword is	*"math software"*
AND	age is	38–42
AND	gender	male
THEN	average projected sale amount is	$79.95

Segmentation analysis can take the form of a graphical decision tree (see Figure 1–5), where the average sales are lower to the left and higher to the right. The splits in the tree are determined by the transactional *conditions*, such as the connect time or domain, or by the visitors' demographics, such as their age or number of children, or on the basis of their consumer class according to their ZIP code.

The IF-THEN rules extracted from segmentation analysis can also be concise statements describing specific market segments:

Amazon/AlmondNet
Exhibit 1011
Page 27

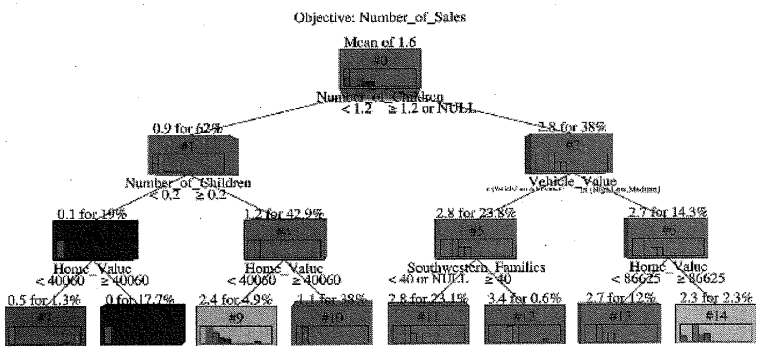


Figure 1–5
One common data mining format is a decision tree, which segments a database.

Condition	Attribute	
IF	Number of Children	>= 1.2
AND	Vehicle Value	Luxury
AND	Southwestern Families	>= 40%
THEN	Average Sale	3.4 (Node #12)

Or,

Condition	Attribute	
IF	Sub-Domain is	AOL
AND	Gender is	MALE
AND	Age is	37–42
AND	Is from ZIP Code	"High Rise Renter" 78–85%
THEN Will Purchase	Probability Score 81%	

Association

What relationships exist between your visitors and your products? Still another type of data mining analysis looks for hidden associations in your data, such as gender, age, domains, etc.

In the analysis shown in Figure 1–6, a relationship was found between the search keywords and product lines. Several data mining,

Amazon/AlmondNet
Exhibit 1011
Page 28

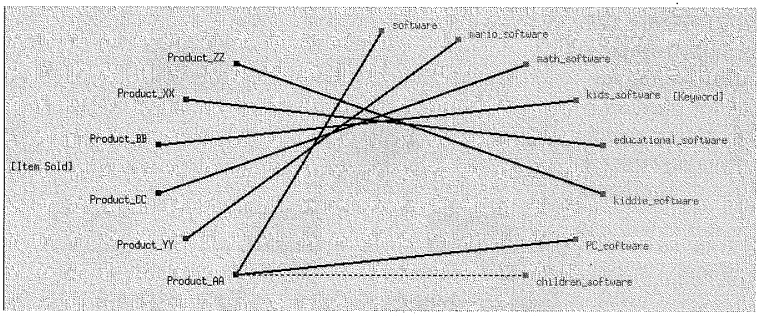


Figure 1–6
The stronger the association, the bolder the links.

```
STRONG LINKS:
 Item Sold = Product_AA, Keyword = software (573)
 Item Sold = Product_YY, Keyword = mario_software (250)
 Item Sold = Product_CC, Keyword = math_software (211)
 Item Sold = Product_BB, Keyword = kids_software (187)
 Item Sold = Product_XX, Keyword = educational_software (181)
 Item Sold = Product_ZZ, Keyword = kiddie_software (74)
 Item Sold = Product_AA, Keyword = PC_software (53)

MEDIUM LINKS:

WEAK LINKS:
 Item Sold = Product_AA, Keyword = children_software (3)
```

Figure 1–7
Summary report of links found at a website.

visualization, and web analysis software programs provide this type of association output. In this instance we are using a data mining tool that not only provides a visual representation of the strong links between keywords and particular product lines, but it also generates a narrative summary report (Figure 1–7).

Association rules and graphs associate a particular conclusion (e.g., the purchase of a particular product) with a set of conditions, such as the purchase of several other products. Association rule algorithms automatically find the associations and present them via visualization techniques. The advantage of this type of association analysis over segmentation decision tree analysis is that association can exist between any of the attributes. A decision tree algorithm will only build rules with a single conclusion

Amazon/AlmondNet
Exhibit 1011
Page 29

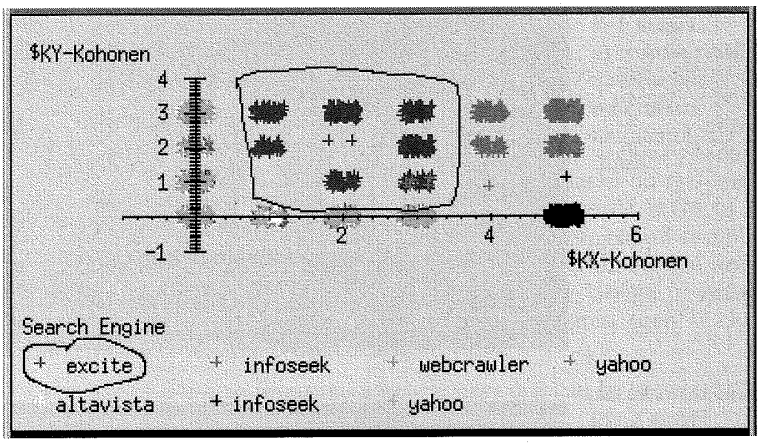


Figure 1–8
Clustering of visitors based on the search engines they used to arrive at this website.

whereas association algortihms attempt to find rules that may have different conclusions.

Clustering

What are the groupings hidden in your data? Yet another format of data mining is that of *clustering,* which is commonly used to search for unique groupings within a data set. (See Figure 1–8.) A special type of neural network, known as a Kohonen network, or a Self-Organizing Map (SOM), is used to perform this type of analysis. An SOM is a form of "unsupervised learning" that organizes itself around distinct clusters in a data set without input from the analyst. SOMs are commonly used by retailers in market basket analysis to search for groupings between products that tend to sell in combination. SOMs can also be used to search for clusters in web data in order to find both the cross-selling patterns of products, as well as distinct clusters, such as the groupings of unique visitors' traits.

Visualization

What are the distributions and patterns in my website data? The human eye is critical in the mining of your data. Visualization can take many forms. For example, in Figures 1–9, 1–10, and 1–11, we have three increasingly narrowed geographical views of a website's visitors as seen from the ZIP codes as submitted in its registration form.

Amazon/AlmondNet
Exhibit 1011
Page 30

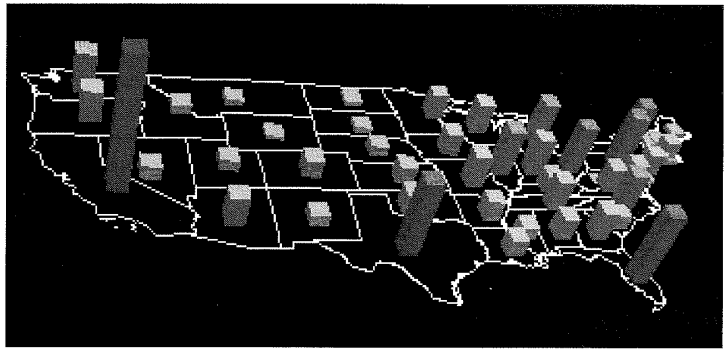


Figure 1–9
Most visitors to this website come from California and Texas.

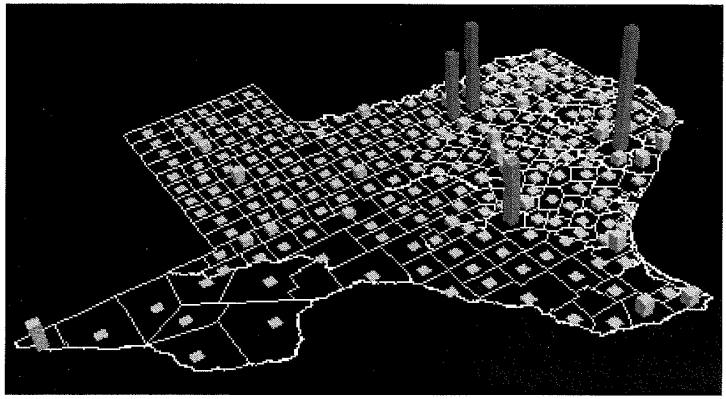


Figure 1–10
This data mining tool allows the user to drill down for further analysis.

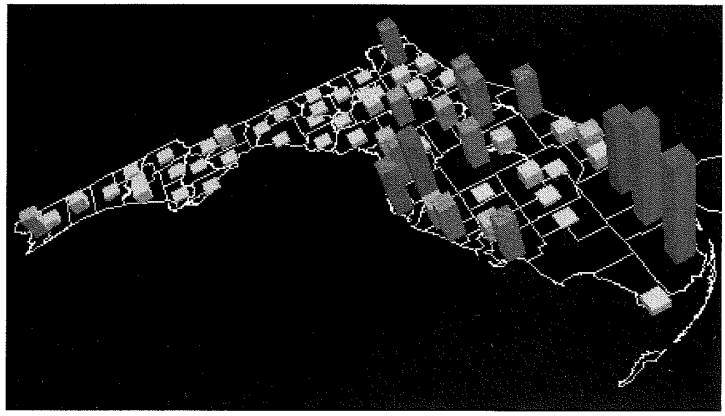


Figure 1–11
Color and height represent a higher than average amount of visitors.

Amazon/AlmondNet
Exhibit 1011
Page 31

Optimization

How can you maximize your online presence and sales? Lastly, genetic algorithms (GAs) can be used to optimize your website presence and online sales. (See Figure 1–12.) As previously noted, GAs are optimization techniques and programs that use processes such as genetic selection, crossover, and mutation for the solution based on the concept of natural evolution. (See Figure 1–13.) For example, a GA in a data mining tool can be used to try thousands of settings and combinations of inputs in the fine-tuning of a neural network. Neural networks and genetic algorithms are also being used to design physical storefronts and can also be used to optimize websites, such as the arrangements of a site's products and services.

The use of a GA to fine-tune a neural network can be viewed as a form of data compression. For example, a GA can prune a very long file containing hundreds of variables down to only a few inputs for predicting sales. This process of optimization is a form of data reduction in which the software will try and reduce its own error rate, while at the same time compressing the number of inputs to a neural network.

A data mining analysis will likely involve the use of multiple techniques and technologies. (See Figure 1–14.) For example, it is a common practice to use a segmentation tool, such as a decision tree program or an association technique tool, to identify important variables in a database with a large number of fields, and then use a

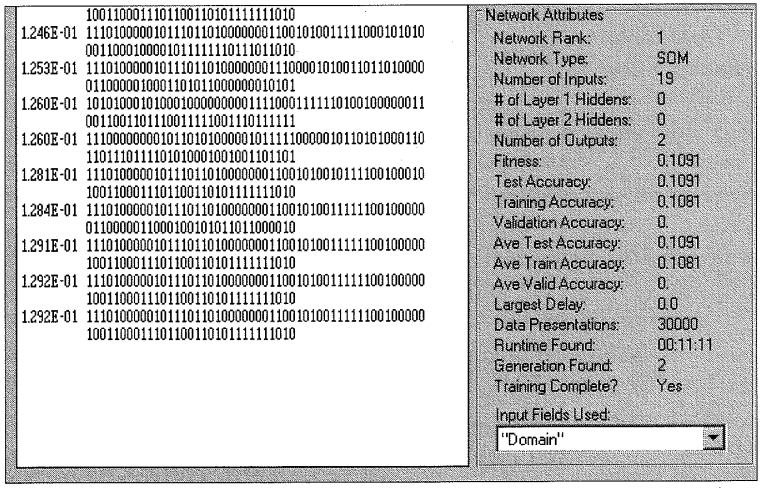


Figure 1–12
A GA can be used to optimize the performance of a neural network.

Amazon/AlmondNet
Exhibit 1011
Page 32

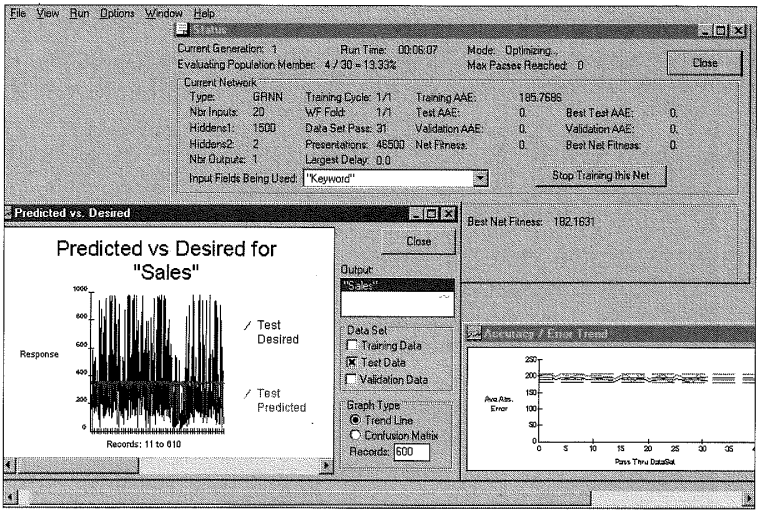


Figure 1–13 *GAs and neural networks can be used to predict online sales.*

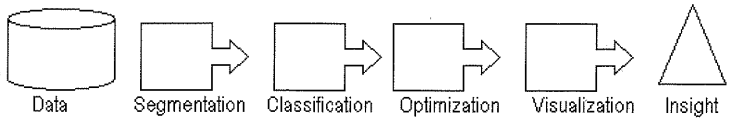


Figure 1–14 *The identification of hidden patterns in your data often involves several steps.*

neural network to construct a model using those selected fields as inputs. Next, a genetic algorithm may be used to further try to optimize the neural network so that even fewer variables are used to predict such outcomes as total sales or the propensity to make purchases by new visitors to your website. Lastly, a visual tool may be employed to display the results and perhaps enhance the insight gained from the entire data mining process.

While all of these formats provide different types of inputs and results, they are not mutually exclusive; instead, they complement each other and should be considered and used in tandem. For example, the visualization of your data can be done both during your initial analysis (to get an overview of your data) as well as after you have performed your segmentation and classification processes. What data mining formats you obtain from your analysis and what sequence you select is really driven by your business needs. In fact, you may select to perform but one single process—it is your option.

Amazon/AlmondNet
Exhibit 1011
Page 33

The formats you obtain from your analysis will affect the steps that you go through in the mining of your website. There are ten general steps that are covered in more detail later in the book. Your data mining format will impact each of these steps, since it will influence your decisions on your objectives, the selection and preparation of your server data, the tools that you use, and how you integrate your findings.

Data Mining Drives Relational Marketing

While early adopters of data mining technology have tended to be from data warehouses in information-intensive and ultra-competitive industries such as financial services, telecommunications, and retailing, today virtually every company has a website generating massive amounts of customer information. Similar to data warehouses, large commercial websites involved in electronic retailing are today dealing with huge amounts of data that data mining tools are especially apt at dealing with. Indeed, with the advent of the Web and electronic commerce, nearly every company large or small is now in an "information-intensive and ultra-competitive" mode and thus likely to benefit greatly from data mining. The Web is an ideal marketing environment where every transaction can be captured, stored, and subsequently mined for a strategic advantage. The data mining of your website can allow you to:

- build unique market segments identifying the attributes of high-value prospects
- identify the key attributes of your web customers for each of your products
- select promotional strategies that best reach your web customer segments
- analyze your online sales to improve targeting of high-value customers
- test and determine which marketing activities have the greatest impact
- identify customers most likely to be interested in your new products
- reduce costs and improve the value of customer relationships
- improve your site's product cross-selling and up-selling
- leverage the knowledge about your online customers
- identify the best online prospects for your services
- understand the reasons for brand switching

Amazon/AlmondNet
Exhibit 1011
Page 34

- manage your web customer relationships
- improve your web ads and sales process
- maximize online ad throughputs

Mining your website data lets you focus your efforts on your most profitable customers and prospects. It allows you to design marketing strategies to reach your clients incrementally—from mass, to market segment, to individuals. As you begin to do electronic retailing you can also improve your relational marketing efforts by mining the data that you daily accumulate from your website. Relational marketing is about marketing in a continuously narrower and targeted range. Ideally it involves marketing to individuals based on observed behavior patterns rather than just demographics or on the basis of keywords you glean from log files.

Today, with ready access to increasingly larger amounts of website visitor information and data mining technology, a website can truly begin to offer individualized treatment to its online customer. Unlike mass marketing, with relational marketing, every customer is his or her own market segment. Relational marketing focuses a site on its true source of revenue and profits: *its customers.* In a dynamic and highly competitive environment such as the Web, electronic retailers must recognize that they can gain a competitive advantage by better understanding and caring for their most frequent visitors and best customers.

Web Privacy

At this point it is important to remind you of the importance of privacy and your need to explicitly tell your website customers how you plan to protect the information they provide you, because if you don't you will have no data to mine or business on the Web. It is important that you look at your website customers' personal information as your company's most valuable asset. You must be prepared to protect it and ensure it is not disclosed. The privacy of your website visitors must be guaranteed, period. Make no mistake, recent surveys indicate people on the Web are not willing to share personal or financial information that can be sold to third parties. In Europe, failure to protect this type of personal information is already illegal, so take care.

You must convince your website visitors and customers that whatever information they provide you will be used to save them time and money—in fact you need to prominently guarantee it in writing. The protection of your customer information is not only good business sense, but it is also the most valuable asset you have in terms of com-

Amazon/AlmondNet
Exhibit 1011
Page 35

petitive intelligence. Think about it—what do you know about your online customers that your rival does not? If you are going to be data mining web-generated information for purposes of marketing and commerce, you need to create a trusting relationship with your visitors and customers. You need to adopt an explicit and prominent security and privacy policy in order to ensure an intimate dialog with your online clients. You need to state:

1. The information you need from them and why this will save them time and money.
2. The tangible benefits they will get in exchange for the information they disclose.
3. The actions you will take to protect and not disclose the information they provide.
4. The control they have over the information, such as their right to purge it or update it.
5. Your assurance that the personal information they provide you will solely be used to improve the service you provide them.

Benefits and Incentives

Website visitors and customers will provide information about themselves if they see some tangible benefits for doing so. For example, the Web continues to expand on a daily basis with the number of websites in the tens of millions and new ones going up every hour of every day. If you can assure your visitors that by completing a registration form you will be able to facilitate the location of important products or services they specifically want, then you are on your way to establishing a productive relationship with them. You need to assure them and prove to them that the more you know about them, the better you will get at serving them in a timely, accurate, and individualized manner. Personalized customer information must be mined with the full intent of improving your individual service to them.

The mining of this data will only ensure you are more efficient in how you respond to your customers' likes and preferences. Data mining is about leveraging pattern recognition technology toward a business objective, which in this case is customer service. This translates into knowing what your customers want before they know it themselves. Ultimately they will want to stay in the client relationship, since training another vendor or competitor will cost them time and effort.

One method of obtaining valuable customer information is by providing a free service in exchange (this is the business model used

Amazon/AlmondNet
Exhibit 1011
Page 36

to provide free e-mail by HotMail). Another method of securing valuable information from your website visitors and customers is by providing them a special discount on your products and services. This may be in the form of a one-time offer, a coupon, a reduced rate, a free ticket, a free upgrade, technical analysis or support, a white paper, etc. Then there is the straightforward appeal. Phrase your requests in this manner: How can I best serve you? How can I provide you with local information, such as a weather report, if I don't have your ZIP code?

An established practice for your website should be that information collected for one purpose should not be used for another without your visitor's or customer's permission. For example, if your customer's e-mail address is being secured solely to contact them to rectify a billing error or to confirm an order, that should be made clear. Your website should be very straightforward about why you are asking for the information and how you plan to use it. Website visitors and customers should be given prior notice regarding the collection and usage of personal information or other type of tracking data. Finally, you should offer your customers the option to "opt in or out"—offer them various levels of privacy protection. The more information they provide you the more service they are likely to receive from you since you will be better able to deliver to them what they want, when they want it, in a manner unique to their needs.

Ten Steps to Mining Your Web Data

Before you start to mine your data you need to consider what your objective is and what kind of information you will need to capture to achieve your objective. For example, you may need to issue identification cookies during the completion of registration forms at your website. This will enable you to match the information captured from your forms, such as the visitor's ZIP code, with the transaction information generated from your cookies. It will also allow you to merge your cookie information, which will detail the locations where your visitors go to while in your website, with the specific attributes like age and gender from your forms. Additionally, a ZIP code or visitor address will allow you to match your cookie and form data with demographics and household information matched from third-party data resellers.

The data from your website will likely need to be scrubbed and prepared prior to any sort of data mining analysis. Log files, for example, can be fairly redundant since a single "hit" generates a record of not only that HTML but also of every graphic on that page.

Amazon/AlmondNet
Exhibit 1011
Page 37

However, once a template, script, or procedure has been developed for generating the proper recording of a single visit, the data can be input into a database format from which additional manipulations and refinements can take place. If you are using a site traffic analyzer tool, this data may already be format-ready for additional mining analysis. Keep in mind that several steps may be required prior to undertaking your analysis, including the following ones, which will be discussed in detail later in the book:

1. Identify your objective — *Profile your visitors?*
2. Select your data — *Form database?*
3. Prepare the data — *Append demographic information?*
4. Evaluate the data — *Visualization?*
5. Format the solution — *Segmentation, prediction?*
6. Select the tools — *Single or suite?*
7. Construct the models — *Train and test?*
8. Validate the findings — *Share with teams?*
9. Deliver the findings — *Provide report, code?*
10. Integrate the solutions — *Marketing campaign?*

Clearly not all of these steps are required, but you should consider them prior to starting any in-depth analysis. They certainly do not always follow this exact sequence, but personal experience bears out the fact that in most assignments these were the issues that needed to be resolved prior to completion of most data mining projects. Most of my prior data mining projects involved working with customer information files, datamarts, and data warehouses from retailers, banks, insurers, phone, and credit card companies, but they typically dealt with the same client-centered issues or questions, mainly: Who are the customers? What are their features? And how are they likely to behave? Electronic retailers face the same questions today.

Most likely you will need to do your data mining on a separate server dedicated to analysis. After your analysis you will need to validate your results through some sort of production system such as a marketing test e-mail campaign. Note that the costs involved with e-mail versus physical mail or phone calls allows for a very rapid assessment of your data mining and marketing efforts. It is certainly a very economical way to evaluate your data mining project: it only costs about five cents to e-mail a potential customer, compared with as much as five dollars for direct mail and eight to twelve dollars for a phone sales call. Planning and executing a traditional marketing campaign used to take months; today on the Web an e-mail campaign can

Amazon/AlmondNet
Exhibit 1011
Page 38

take hours. The Web has precipitated the trend toward one-to-one marketing and the validation of data mining results by allowing the rapid evaluation of predictive models.

It is not difficult to assess the benefits of data mining and its return-on-investment (ROI). Simply consider the quantitative counts of clickthroughs of ads or banners prior and after your data mining analysis. Consider the percentage of sales or requests for product information, as well as the amounts of purchases made as a result of a data mining analysis. Consider the rates prior to your data mining efforts and afterwards. If you initiate a marketing e-mail campaign on the basis of your data mining analysis, consider the rate of responses by splitting your e-mails on those individuals targeted via your analysis and those excluded from the targeting. Measure the improved rate of responses and sales from those targeted via the data mining analysis and those without it.

The dynamics of your industry and marketplace will dictate how often you should mine your website data. The intervals for mining your data will depend on how often the attributes of your customers change. For example, a bank may have a cross-selling model for its call site that can be quite effective for months. The intervals in which the bank model are created may take place on a quarterly or monthly basis and still be relevant to the business questions they are trying to answer, such as cross-selling opportunities of their financial products like CDs, bankcards, loans, etc. For a portal, such as a search engine, models may need to be refreshed on a weekly basis, because the dynamics of the content, their visitors, and their features change more quickly than, say, with a bank's customers. The end products they are trying to predict are also subject to change more frequently, for a bank it is a loan, for a portal it is a banner or ad.

For an Internet company which exists completely on the Web, the data mining process represents a biofeedback system to its entire supply chain. Data mining can identify for electronic retailers key market segments, which can impact directly on its overall website design and inventory control systems. As with physical retailers, the leveraging of data mining pays off in the positioning of the right message, product, and service in front of the right customers at the right time in the right format.

Data mining is not an isolated process carried in a vacuum; it must be integrated into the entire electronic retailing and marketing processes. This is especially true with virtual storefronts because everything—selections, transactions, orders, customer communications—is accelerated to "Internet time." For a website entirely sup-

Amazon/AlmondNet
Exhibit 1011
Page 39

ported by advertising, data mining is even more critical since it can quickly discover and measure the effectiveness of a multitude of banners and ads on its continuous stream of visitors.

A Bookseller Example

Data mining can provide multiple solutions to website designers, marketers, and merchants, such as an online bookseller. For example, it can provide to a book site an insight about its electronic commerce activities and its clientele. Data mining analysis can help it resolve such questions as:

- What books are being sold to which visitors?
- Who and where are these visitors located?
- Which visitors are the most profitable?
- What factors impact its online sales?
- Who is likely to buy what books?

To answer these types of questions, the bookseller needs to start with an analysis of its customer information. The customer information can come from various sources, including of course the clients themselves, as captured by the registration forms located at its website. It may also be information maintained in its data warehouse or purchased externally from demographers or household data providers (see Figure 1–15).

On the basis of its visitors' ZIP code, ZIP+4, or full address, data resellers can provide the bookseller information about their incomes, the type of automobiles they drive, whether they own or rent, the presence of children and other lifestyle information. Once the book

Book Category	Income	State	Sales	Children	Gender	Age	LastSale	Top 1%	Wealthy Seaboard	Empty Nesters
Computers	51928	MA	366	0	M	55-59	19980215	0	0	0
Literature	25337	MA	272	2	M	40-44	19970829	0	0	0
Investing	25339	MA	153	1	M	40-44	19970829	0	0	0
Home	45575	MA	132	3	M	50-54	19970913	0	0	0
Internet	25340	MA	270	1	M	30-34	19970829	0	0	51
History	45576	MA	144	1	M	30-34	19970912	0	0	51
Horror	25342	MA	132	4	M	30-34	19970829	0	0	4
Garden	25343	MA	292	1	M	44-49	19970829	0	0	4
Home	25344	MA	144	2	M	40-44	19970829	0	0	4
History	25345	MA	226	2	M	50-54	19970829	0	0	0

Figure 1–15 *A bookseller data set combines transactional and customer information.*

Amazon/AlmondNet
Exhibit 1011
Page 40

site's data has been enhanced, several data mining techniques and technologies can be used to visualize, cluster, segment, classify, and profile its customers—including the modeling and prediction of their online behavior. For starters, this bookseller may want to:

1. *Visualize* the data to look for significant trends or hidden associations
2. *Split* the data to look for unique clusters or different groupings
3. *Model* the data to anticipate the behavior of its visitors

Look at the Data

This bookseller can graph and plot its customers' attributes in order to gain an overview of who they are. Such a visual inspection can reveal certain demographic features about who is buying and who is not, and provide the bookseller's website designers and marketers some direction on what type of campaigns, incentives, awards, and books to offer, and to whom. For example, at this book site (see Figure 1–16), fiction is the most popular category.

Relationships can also be discovered through visual data mining. Figure 1–17 depicts the links between book types and visitor age groups. The stronger the relationship the bolder the links appear on the web diagram. (See Figure 1–18.)

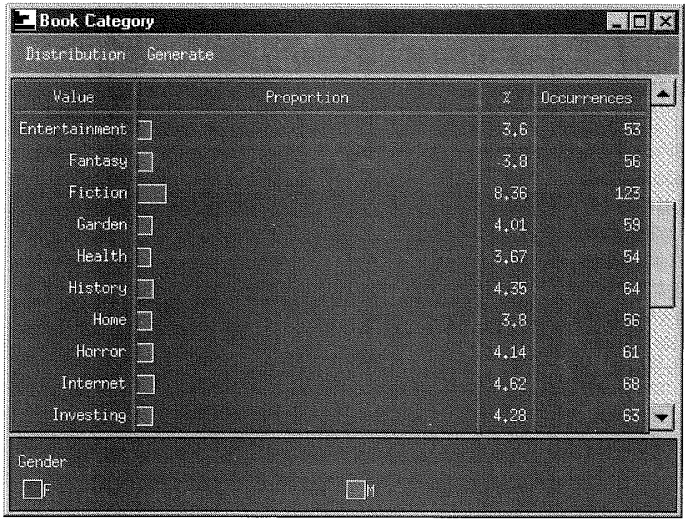


Figure 1–16 *Most customers are male and buy fiction and Internet books.*

Amazon/AlmondNet
Exhibit 1011
Page 41

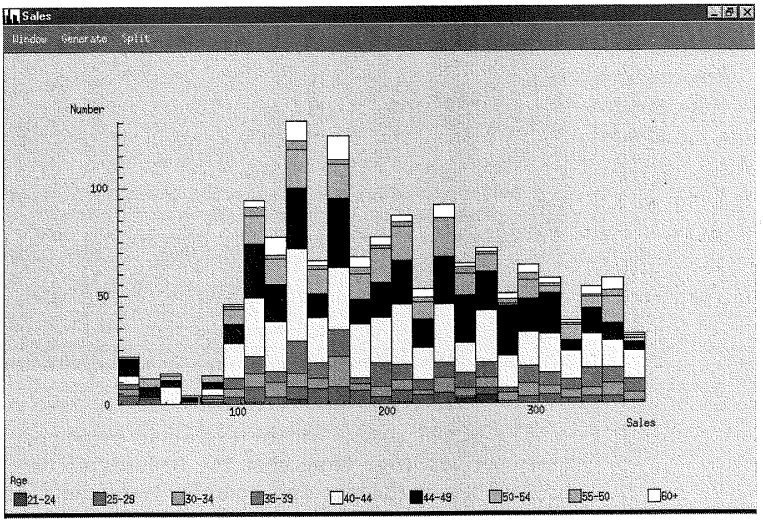


Figure 1–17 *Most of this bookseller's customers are between 40 and 54 years old.*

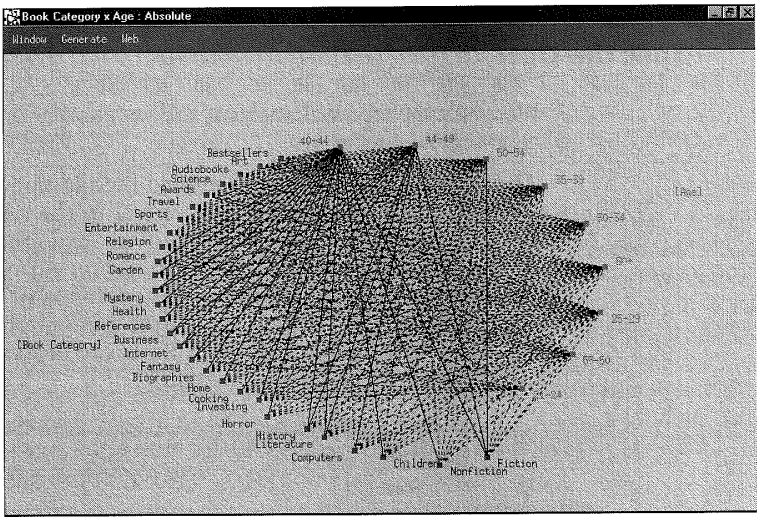


Figure 1–18 *Link analysis showing the associations between book categories and age groups.*

The associations in the diagram can also be summarized in tabular format with the actual counts of occurrences provided in a table, as in Figure 1–19.

Amazon/AlmondNet
Exhibit 1011
Page 42

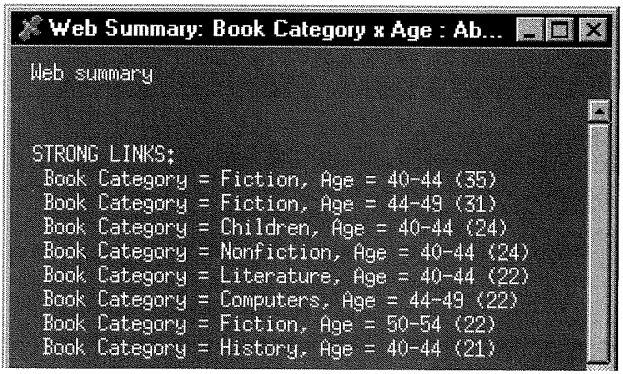


Figure 1–19 *Prevalent links between book types and age groups.*

The actual counts can be examined between products sold and customer features. Note that the benefits of data mining grow in importance as the size and complexity of the data increase. In this example, we are dealing with a very small sample size. An analysis of several hundred thousand transactions is quite common for large retailing websites.

Split the Data

A different type of visual analysis, and yet another way of splitting the data, is that of *clustering.*

Groups of clusters can be examined for additional information, as was the case with this analysis of region2 in the right side of Figure 1–20. Clustering offers the opportunity to find grouping in the data through a process known as *self-organization*—driven by the nature of data, rather than the intuition or hypothesis of the analyst. This sector can further be *"drilled"* by a data mining tool in order to discover some of its unique attributes. Rules can be extracted describing the features of the clusters in region2:

IF	Age	40–44
AND	State	NY
THEN	region2	Cluster for 56 visitors 98.2% Confidence

Or,

Amazon/AlmondNet
Exhibit 1011
Page 43

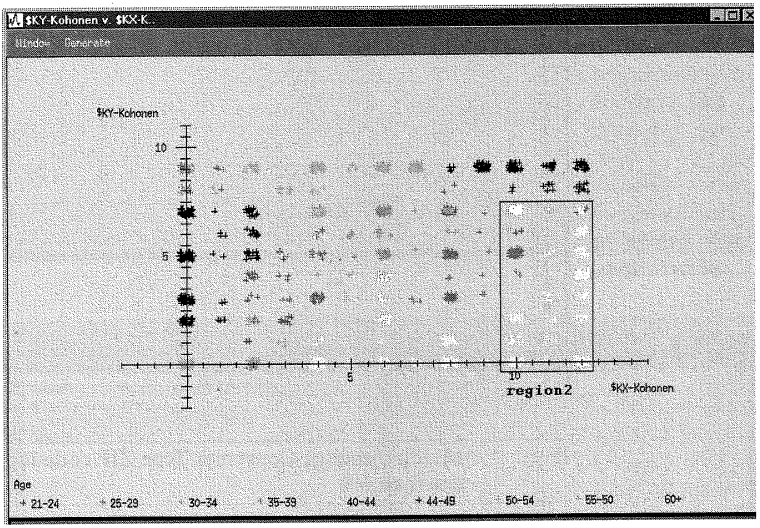


Figure 1–20 *Clusters of data as grouped according to visitor age groups.*

IF	Age	40–44
AND	Income	> $26,976
AND	State	NY
THEN	region2	Cluster for 37 visitors 97.3% Confidence

Segmentation analysis can discover important grouping in this bookseller's website data. The key difference between this and clustering analysis is that with the former you start with a clear objective, such as splitting the data according a specific value or output (i.e., "projected total sales"). In clustering, there is no predefined objective; it is a more exploratory type of analysis in which the data is allowed to organize itself. In segmentation, the bookseller starts the analysis by asking the questions, "who are my most profitable customers and how do I recognize them," which is in turn answered by the decision tree in Figure 1–21.

The results of segmentation analysis can also be interpreted as rules. For example, in Figure 1–21 the highest projected sales is $244.77, which is described in the following rule:

Amazon/AlmondNet
Exhibit 1011
Page 44

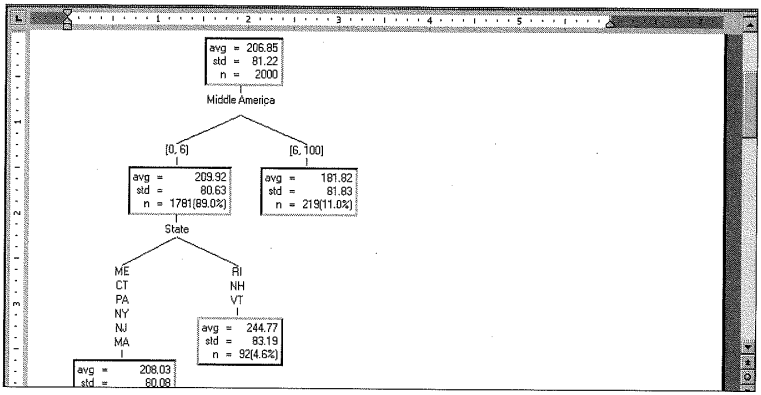


Figure 1–21 *Segmentation analysis decision tree, in which* avg = *average sales,* std = *standard deviation, and* n = *number of observations.*

IF	Middle America Consumer Type ZIP code is greater than 0 but less than 6%
AND	visitor comes from Rhode Island, New Hampshire, or Vermont
THEN	projected online sales is $244.77, with a standard deviation of $83.19

NOTE: The average total customer sales for this site is $206.85.

Model the Data

One of the most important types of analysis a bookseller is likely to make is one involving the modeling of its website visitors' behavior. This type of analysis is known as *classification* or *prediction* and involves the discovery and learning of visitors' features and their online behavior. Like segmentation, classification involves examining visitors' features and their propensity to click through an ad or banner or make an online purchase. Modeling and predicting website visitor behavior can be accomplished using data mining tools incorporating neural networks. (See Figure 1–22.) A neural network is software designed to train itself on the patterns in large sets of samples, which in this case involve the online activity of its website visitors.

Modeling involves uncovering subtle customer and transactional nuances that can be coded in the form of "weights" in a formula for scoring the behavior of future new visitors. (See Figure 1–23.) On the basis of these scores, the bookseller can mount multiple targeted

Amazon/AlmondNet
Exhibit 1011
Page 45

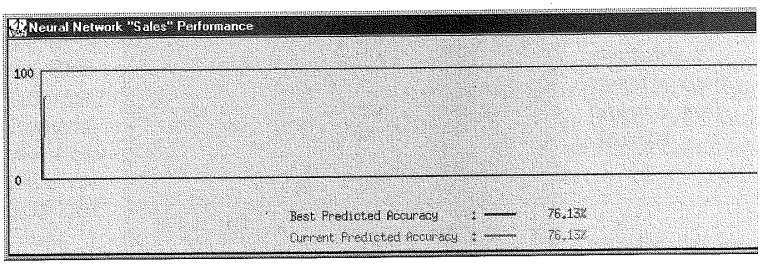


Figure 1–22
A network will iterate through thousands of observations to discover a pattern.

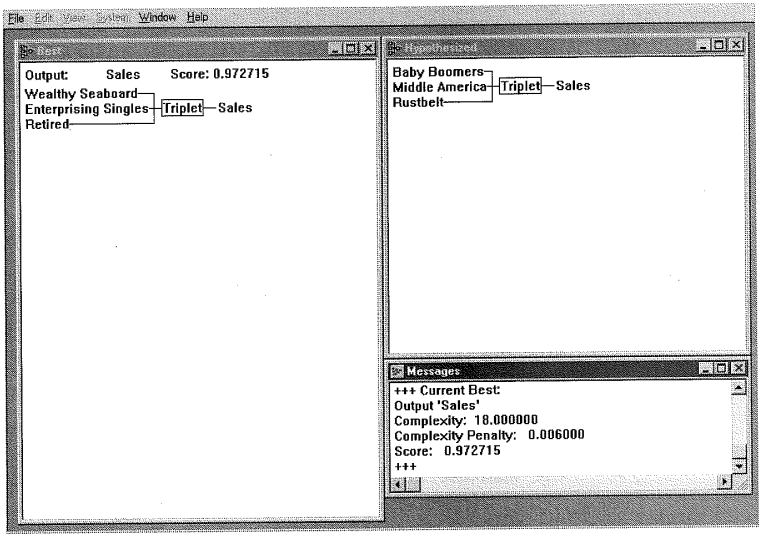


Figure 1–23
A network will try several combinations of inputs in an effort to predict an output.

e-mail campaigns offering specific types of books only to those individuals likely to be interested and have a high propensity to want to make a purchase. A recent study found that unique visitors to one book site never go to another online bookseller. Once visitors learn how to use one site and once that site learns their preferences and buying habits, they're likely to remain with that bookseller.

Our bookseller, via data mining, can now offer enough value to lock its customers into an electronic relationship to improve its sales and to retain them as clients. A marketing law says it costs ten times more to acquire a new customer than it does to sell to an existing one. By using knowledge it already has about its clients, the

Amazon/AlmondNet
Exhibit 1011
Page 46

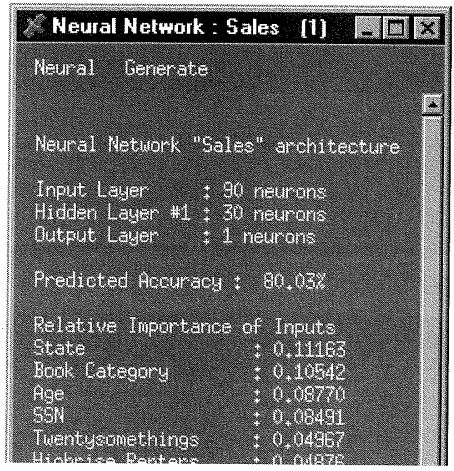


Figure 1–24 *A sensitivity analysis report from a neural network.*

bookseller can increase the efficiency of its marketing efforts. An additional value is that through this profiling effort it can now target potential new visitors armed with the knowledge of who its existing customers are. It can now furnish an image, lifestyle, and attitude to these potential new visitors who match the composite of its existing clients.

A *sensitivity report* from a neural network can tell the bookseller what factors are impacting his online sales, which in this analysis is the visitors' home state, book category, age, and so forth. In Figure 1–24, the sensitivity report coincides with the findings of the visual, association, and segmentation analyses, which found that customers who purchase fiction and Internet books (Books Category), live in Rhode Island, New Hampshire, and Vermont (State), and are between 40 and 44 (Age), to be the most profitable and possess the greatest loyalty. The visitor's gender was excluded as an input from the neural network model, since we already knew that males are a dominant characteristic.

Integrate the Solution

Data mining should not be considered as an isolated independent, one-time process. It should instead be an integrated process of analysis impacting every aspect of a retailing site. On the basis of data mining analyses, a retailer can project and anticipate the allo-

Amazon/AlmondNet
Exhibit 1011
Page 47

cation of shipping resources, which can affect the speed by which it is able to deliver the goods, purchased at its website. Shipping on the Web involves not only the physical transportation of goods, but includes the electronic delivery of information. On the basis of this analysis, seasonal and product-specific trends can be discovered and monitored. Patterns can be uncovered that can alert the retailer on what resources it may need to ensure prompt deliveries of its products and services.

Through the mining of its sales data, an electronic retailer can formulate the marketing strategy to deploy at its site. Analysis of its visitors' demographics and the products being purchased by them can impart valuable insight. Product-specific patterns may be uncovered, which can be used to evaluate marketing campaigns. Mining can reveal what domains, keywords, gender, age, income levels, and other visitor attributes are associated with each sale at a retailing site. As such, marketing strategies can be keyed to specific products, involving targeted ads and e-mail promotions.

The analysis of retail sales data, as with physical stores, can directly impact a website's everyday operations, including that of its inventory controls. Through the analysis of its retailing data a website can adequately plan and anticipate product demand. The benefits of such analysis may take the form of reduced overhead and the prompt delivery of its products and services. Seasonal trends can be anticipated and incorporated into how a site stocks its inventory, whether it be books, CDs, or software. Analysis of the demographics of its customers can provide the retailer vital clues about who is buying what, when, and why.

Because of the rapid nature of electronic commerce, retailers must be prepared to move quickly to meet the needs of its online customers. The flexibility of online retailing allows merchants to monitor sales and quickly adjust pricing and availability of products and services. Data mining of retail information, whether in the form of bar code data or server databases, can reveal important trends and patterns that can affect all aspects of production including shipping, marketing, inventory, and sales.

Compiling an Identity

The key to knowing who will buy, what is selling, and how to sell resides in the transactional and customer data generated from a retailing site,

Amazon/AlmondNet
Exhibit 1011
Page 48

and the key to compiling and capturing this shopper information starts with the establishment of a unique identifier: a visitor ID number, or *key*. This key is needed to associate all of these components into a single record for each visitor. Once linked, this single record string can be used to link to additional customer information from your data warehouse, or appended with external household and demographic information from commercial data reseller. As this visitor profile is built multiple analyses can be performed.

A proven strategy for assigning keys to your visitors is to have them register initially at your site by providing them a special service or incentive or by offering access to a special section in your site. Another strategy is to hold periodic contests or door prizes for which you ask visitors to "register to win." When they register, you may set a cookie that can be used as the unique ID number. From that point on, the unique key can enable you to track every interaction with that visitor. The key will also allow you to link the information from your server log files, cookies, and forms with your data warehouse and other demographic and household information.

Server log files provide domains, time of access, keywords, search engine used:



13:04.9, "furby", excite, id344,384 etc.

id344,384

Cookies dispensed from the server track browser visits and pages viewed:



id344,384, visit_3, prod22.htm, etc.

id344,384

Forms capture important visitor personal information, such as gender, age, or ZIP code:



id344,384, male, 35, 94551, etc.

id344,384

Demographic and household data can be appended, such as total sales or annual income:

Amazon/AlmondNet
Exhibit 1011
Page 49



id344,384, excite, prod22.htm, male, 35, 94551, total_sales_$350, $75k_income, SUV, 2_children, etc.

id344,384

The above string can take the form of an Oracle table or a flat-file, comma-delimited text file, which can be linked or imported into your data mining tool of choice.

The Architecture for Analysis

The analysis you perform can serve as a feedback system to your retailing site, which can impact your site design, sales, inventory, and marketing teams. Data mining of your customer information can address such issues as:

- *Who is buying:* Customer profiles, who purchases what items and at what rates.
- *What is selling:* Adjust inventory, plan your orders, shipping, and inventory.
- *How to sell:* What incentives, offers, and ads work, and what site design and style are most successful.

Data mining reports take various formats, including the delivery of reports, statistical models, graphics, maps, SQL syntax, and/or C code containing the weights for prediction of online visitor behavior. These analyses will be iterative in nature due in part to the dynamic nature of the website and the traffic patterns it is designed to model.

Access to data for analysis can be made either in the format of a flat ASCII file or by invoking ODBC drivers. Depending on the type of tool you are using, data can be extracted directly from existing tables on your server (or servers). Most high-end data mining tools support access to databases such as Informix, Ingress, Oracle, and Sybase. If a flat-file format is chosen, a fixed-length or a delimited character file must be specified. Comma-delimited files are the most compact and should be used when space is a concern. Most tools will convert the data they import into a proprietary structure of their own, which is then used for the actual views and analyses. In most instances, your data mining analysis will need to be performed in a dedicated server or data mart.

Amazon/AlmondNet
Exhibit 1011
Page 50

Personalized Future

Today you can personalize every major "portal" you use to access the Web. This means that "the Web" is now "my Web," allowing you to tailor all types of information streams into your own customized news, weather, market, and stock reports. However, in exchange for your own unique view of the world, most of these portals (such as major search engines, online directories, and large content providers) require some personal information from you. They typically require you to complete a short form—which in turn provides some useful data to the portal that can be mined, allowing for a more targeted and proactive dialog between the website and you. This information, coupled with the topics you select and the "keywords" that you use over time, can be mined for purposes of creating a *profile* of you. This profile can be used, among other things, to suggest new sites; provide timely recommendations on products and services; and favorably position certain ads, banners, and buttons that are more likely to get a response from you. Gradually, users are becoming accustomed to getting timely and accurate product recommendations based on their preferences and personalized features—features that can quickly be discovered by mining the data from these portal registration forms.

The Web provides an extensive and prevalent choice of options, which has spawned an audience of very savvy and knowledgeable customers. Today, it is easy to accumulate a vast amount of product information and comparative data. In a customer-centric environment such as the Web, the customer is not only always right, she is also very well informed. To stay competitive in such an environment your website must establish a relationship with your customers. The data that you gather from them must not only be protected, it must be strategically used. Even inactive data must be converted into actionable intelligence via data mining.

The increased availability of choice has also eroded brand loyalty. For example, if a portal does not provide a user with the type of content that interests him, he will, with a click of the mouse, switch to a competitor who will. To succeed in such an environment you need to personalize your site, and in doing so you can benefit from better customer loyalty, lower marketing costs, and a competitive advantage. The advantage comes through the relationships established between your site and your online customers.

You must attend to your web customers' needs by making it convenient for them to gain a substantial benefit for doing business with you by customizing your products and services to meet their unique

Amazon/AlmondNet
Exhibit 1011
Page 51

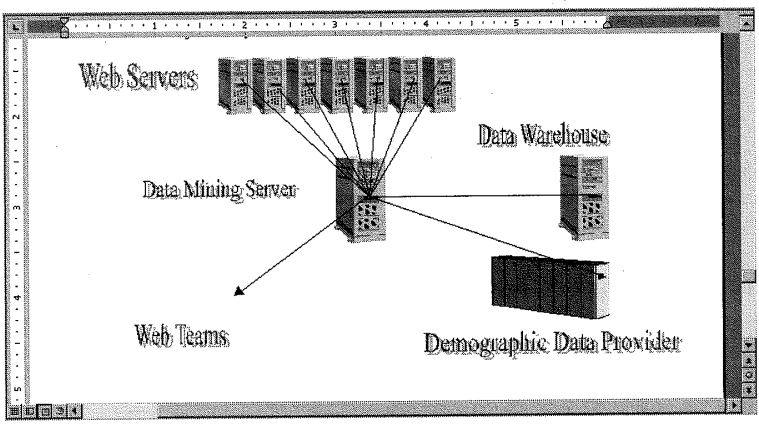


Figure 1–25
Data mining can involve many different enterprise professionals, processes, and sources.

needs. When a customer takes the time to complete a form about them, they have just invested time and effort into building a business relation with your website. To gain the most benefit from that customer information you must mine the data.

As previously discussed, one way of maximizing the benefit of that information is by the integration of your website with your data warehouse of other customer information file. (See Figure 1–25.) This allows you to mine interactions on an individual basis and to construct customer profiles. Your website and its various components, such as forums, chat, and e-mail, allow you to learn from your customers and to engage in an interactive dialog, which in turn leads to the personalization of your site.

The personalization of your customers' websites requires that you gather information from your customers and store it in a database for mining. This increases your chances of becoming indispensable to your clients, and it also encourages customer retention with each transaction. Personalization requires that your website customers and visitors have the ability to create, modify, control, and delete their profiles. The benefits of this interaction, however, is that over time (1) it will cost you less to market to these web customers, and (2) their purchases will surpass the initial expenses of personalizing your website. You will be able to calculate their value based on the duration of their relationship with your site and the amount of money they spend with you or the number of transactions they have with you.

Personalization empowers web visitors and customers to dictate what information they want and in what format and at what intervals.

Amazon/AlmondNet
Exhibit 1011
Page 52

It allows them to construct their own profiles and to control the duration of their relation with your company. The Web has accelerated the marketplace by increasing the availability of options to the consumer. To deliver personalized services at your website you need to mine your data to discover customer segments and then specific individual profiles. To what extent you personalize your site via data mining is really dependent on your budget and resources.

General	Segmented	Relational
Mass Communication	Demographic	One-to-One E-mail
No Targeting	Content Targeting	Specific Targeting
No Date Mining	Mining by Decision Tree	Mining by Neural Network
No Web Data Analysis	Log and Form Analysis	Cookie and Profiling Analysis

Nonintrusive, yet responsive personalized service is or will become the norm on the Web. Customers will appreciate being able to search and find the products, service, and information they want when they need it at the lowest possible price. Commercial websites will have to integrate themselves with their company inventory databases in order to come up with the proper product requests unique to that customer's profile. This integration will also allow for additional cross-selling recommendations based on data mining analysis. For example, if a data mining analysis using a clustering network discovers a propensity for certain individuals with matching demographics to like a specific product, you may want to make a product recommendation to that customer based on that match.

Analysis of customer transactions and personal information should be scheduled and performed on a regular basis. Keep in mind that personal customer information is cumulative and should not be requested all in one session. It should be requested when you are going to use it to personalize your service. For example, don't ask for an e-mail address unless you plan to use it immediately for some sort of notification. Gradually, as the customer returns to your site to make more purchases, you should collect more data regarding each transaction so that your interactions are more targeted and based on accumulated data you have mined from prior visits. Personalization is about using information accumulated through your relationship with customers and mining the data in order to deliver a better shopping experience for them.

Amazon/AlmondNet
Exhibit 1011
Page 53

The current explosion in web databases has resulted in a growing glut of customer data. Your company's ability to strategically deal with these terabytes of web data will depend on your effective use of data mining technology, which can offer new insights and improve your firm's ability to deal with the personalization of the Web. To stay competitive, you need to actively monitor your online transactional trends within your web data in order to target and nurture your most loyal and profitable customers. If you are successful in obtaining customer information you need to immediately put it to use in order for them to return to your website again and again. After all, the mining of your website is simply about being able to see, talk, and relate with your online customers.

Why Bother?

Data mining, whether in the physical or virtual marketplace, has always been much more than the mere use of software tools. It is a business *process* designed to extract actionable intelligence for product and financial analysis sales, marketing support, and customer segmentation. It is about knowledge discovery for a strategic, tangible, competitive business advantage. Now more than ever this actionable insight comes from the highly dynamic data a firm daily gathers from its online customer information file: its website.

Through the mining of its website data and the discovery of the distinguishing features of its visitors, a company can provide them with the type of information, product, and services they are most likely to want. The mining of website data can also assist in the overall design and direction that site will evolve into over time. The process of mining, modification, marketing, and evolution of a website as it changes to meet the demands of its visitors and customers is almost *organic* in nature. As with some of the data mining tools rooted in the evolutionary concepts of computing, the purpose for data mining your website data is to recognize and learn from your visitors, so that you may serve them better than your competitors.

As websites increasingly become the first point of contact between a company and its current and future customers, the mining of website data is becoming a critical method by which to attract and retain visitors. This is especially true in a marketplace with millions of websites—all competing for the attentions and dollars of millions of browsers and shoppers. As with the physical retailers, telcos, banks, and credit card companies, the mining of web server data is driven by the competition in the marketplace and the need to know or predict

Amazon/AlmondNet
Exhibit 1011
Page 54

the behavior of consumers. More than anything the Web has precipitated the trend of one-to-one marketing of which data mining is an important component.

At around the time I was having my e-mail dialog with the search engine company, I had also been involved in preliminary discussions with an insurance company about the mining of their call site data. It occurred to me then that as more and more individuals had started to use the Web on a daily basis, their interactions with this insurance firm, like that with that of other retailers and service providers, were changing from phoning their call site to directly interacting with their website. The transactional data that I commonly mine for these companies is today rapidly migrating to their websites, where it is often faster to place or check on the status of an order, make a reservation, or get the information a customer traditionally deals with through intermediates, whether independent agents or their call site employees.

Clearly the future of electronic retailing pivots around positioning the "right" web content, products, and services to attract and retain customers; an electronic commerce future in which the number of businesses conducting online transactions is expected to exceed hundreds of thousands; a marketplace where the number of consumers buying over the Web will exceed 100 million and the projected number of sales will go from a few to hundreds of billions; where in the ebb and flow of these online transactions gradual patterns will evolve from which electronic retailers will be able to mine and discover their customer tendencies, preferences, features, and behavior. The future of electronic retailing hinges on the strategic use of data mining technology for providing prompt personalized services and products—where the sales clerk will likely be an AI.

Amazon/AlmondNet
Exhibit 1011
Page 55

2

What Is Data Mining?

It's about Predators and Patterns

Data mining is a "hot" new technology about one of the oldest processes of human endeavor: *pattern recognition*. Our hairy ancestors relied on their ability to recognize the patterns of predators, paths, prey, and the seasons to survive. Today, companies inundated with data that is daily generated with every customer transaction, web hit, bar scan, credit card swipe, and phone call are faced with the same challenge of recognizing the patterns of opportunity and threat to their survival. Throw in the real-time, hyper-competitive, networked business environment that is the Web, and you will see why the ability to decipher these patterns is fast becoming a critical factor to a company.

For example, let's say you are a retailer of hot pepper sauces on the Web. A data mining tool can discover the following rule—which is actually an important ratio—or insight about your customers' identity and purchasing patterns:

IF	Visitor Sub-Domain is AOL
AND	Visitor Gender is MALE
AND	Visitor Age is 37–42
AND	Visitor is from ZIP Code "High Rise Renter" 78–85%
THEN	Visitor Will Purchase Score = 81%

You can benefit from data mining even if you don't sell habaneros online. Whether you provide a service or a product, data mining can assist you in discovering previously unknown patterns or features about your website visitor, your online customers, and your relationship with them. By recognizing who your customers are you

Amazon/AlmondNet
Exhibit 1011
Page 56

can begin to establish a stronger relationship with them, which in the long run will enhance your chances of retaining them over their consumer life.

Data mining differs from other data analysis methods in a fundamental way: it discovers hidden structures, ratios, patterns, and signatures. Data mining is uniquely dynamic in that you don't need to prepare queries or set out to resolve a particular problem before you mine. Data mining retrieves relationships and enables "drill-up" analysis based on them. For example, your discovery in the hot sauce sample may lead you to do subsequent queries for website visitors with the same consumer class "High Rise Renters" or age bracket.

Just as a forensic investigator is able to approximate the weight and height of a perpetrator from a single footprint on a beach, so can your company calibrate the value and unique features of your most loyal and profitable website visitors and customers. How? Through a careful analysis of the clues in the patterns evolving from your daily website interactions. Customer profiles and prospect composites are provided to you every day with every visit to your site. The clues to identifying profitable visitors and potential new clients are contained in the data generated from your log, cookie, and registration forms. The tools and techniques for uncovering them are in data mining technology.

A Definition

Data mining is the process of discovering actionable and meaningful patterns, profiles, and trends by sifting through your (website's) data using pattern recognition technologies such as neural networks and machine-learning and genetic algorithms. It is an iterative process of extracting patterns from online business transactions—everything from clickthroughs of your banner ad to requests for e-mail information and form registrations—for purposes of improving your website company's bottom line. Data mining is the automatic discovery of usable knowledge from your stored server data and leveraging pattern recognition technology toward answering such business questions as:

- Who are my most profitable online customers?
- How do I increase my online market share?
- How do I optimize my online inventory?
- Who are my website visitors?

Amazon/AlmondNet
Exhibit 1011
Page 57

Drowning in Data

Data storage is getting cheaper while at the same time the amount of data is exploding. The United Nations 1996 Population Report estimated the world population growth to be about 3.1 percent per year. In contrast, the estimated disk storage is growing at about 130 percent per year, as reported by the information technology research group the International Data Corporation (IDC). All of these figures and projections are of course exclusive of the Web and the new data it is generating every second of the day with thousands of new websites going online.

Information flows into your company's website every day—information about your visitors and potential future customers. So much information is coming in that the average business uses only about 7 percent of the data they keep. Still another marketing research firm, the Gartner Group, estimates that only 2 percent of the existing online data is currently being analyzed, a percentage which is rapidly falling toward zero as storage more than doubles each year. Meanwhile the growth of the Internet, websites, and electronic commerce is also expected to continue its growth over the next few years. Further, Figure 2–1 shows Emerging Technologies Research Group estimates that Internet usage from all types of methods of access will continue to increase into the next century. Forrester Research, Inc. estimates e-commerce will reach $3.2 trillion in 2003.

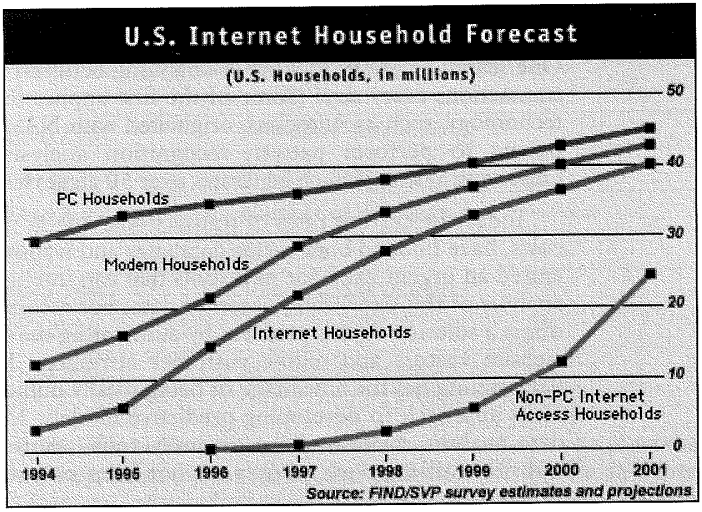


Figure 2–1 *The growth of the Web continues to increase. Courtesy of Emerging Technologies Research Group.*

Amazon/AlmondNet
Exhibit 1011
Page 58

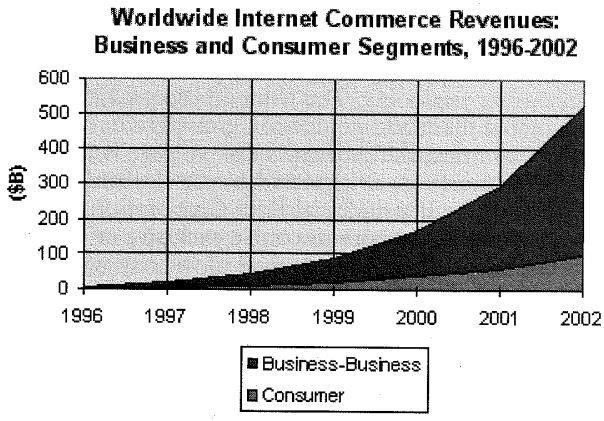


Figure 2–2 *Projected growth for EC. Courtesy of International Data Corporation.*

This growth is expected to be even more dramatic as electronic commerce increases in business-to-business transactions. (See Figure 2–2.) Given such a scenario, the only way to address this gross underutilization of stored data is through advanced methods of analysis and pattern recognition. These new approaches are the foundation of the technologies incorporated in today's data mining tools.

During the last decade, the capabilities to both generate and collect data have grown enormously. For example, a retailer such as Wal-Mart needs to make daily updates of over a million transactions to its data warehouse for inventory analysis. With the advent of the Web, millions of new transactions are now occurring around the clock worldwide. One search engine is looking at analyzing between 25 and 75 million transactions every *day!* (Some of the first applications of data mining technology, such as Autoclass, originated with NASA. Why? Consider having to perform pattern recognition analysis on their Earth Observing System, which generates over 50 gigabytes of data per hour.)

Today the current rates of transactions, due in part to low storage costs, have flooded both forms databases and website servers and generated an urgent need for new tools that can intelligently assist in the analysis of this expanding ocean of data. For a company, data mining offers a solution to this problem by automating the search for relevant website visitors and online customer attributes. More importantly, data mining has the advantage of being totally unbiased and nonlinear in its approach to developing predictive models. Manual methods of data analysis, such as those of query tools, statistical programs, or website traffic reports, simply cannot keep pace with the explosive

Amazon/AlmondNet
Exhibit 1011
Page 59

growth of online data. Hypothesis-based methods of data analyses are very time-consuming, while current website analysis tools are very limited in discovering the features of site visitors and customers.

The Complexity of Loyalty

The most valuable assets a website has are its customers—especially loyal customers, who return again and again over a period of time. The value of customers can be measured according to their lifetime value (LTV), i.e., the duration of their relationship with a site and the amount of money they spend during that time. As electronic retailing has evolved, sites have begun to find it profitable to group their clients into LTV classes with unique marketing needs. The retailer can then tailor its marketing activities according to different customer objectives and their potential worth to their site. The more loyal, and profitable, a client is, the more perks and incentives he/she is provided.

Distilling the value of customers involves in-depth market research and analysis. In a forms database with thousands of records and hundreds of attributes, how do you determine the demographic and psychographic attributes that distinguish valuable visitors from less profitable ones? At what point do you differentiate between "high" and "premium" customers? How do you find hidden groups of new users, brand switchers, dormants, upgraders, recent win-backs, and all the other clusters of customer classes? In short, how do you measure loyalty? The answer is in data mining.

A company's customer database can have hundreds of fields recording different types of interactions both offline and online culled from various contact points including a firm's website:

Data Field	Customer Attribute
Field 1	Account start date
Field 2	Number of purchases
Field 3	ZIP code
Field 798	Total number of visits
Field 799	Average length of visits
Field 800	Monthly access rate

In order to construct a customer loyalty model, you must discover two key items in the data: the significant *attributes* and *intervals*,

Amazon/AlmondNet
Exhibit 1011
Page 60

which identify a pattern or signature of customers who have been retained by the company. This can involve combining attributes of existing customers and new website visitors who have recently signed on with a website. Through this process a "signature" of potentially highly loyal and profitable new clients can be discovered. These new customers may be targeted for high incentives and special offers.

The first goal of this analysis is to discover the attributes that are the most important signals in identifying those clients likely to be the most profitable: Is it the ZIP code they reside in, or maybe their establishment account date? Is it their monthly access rate, or their age or gender? Or is it a combination of factors, and if it is, which attributes?

The second question that needs to be answered from the data is that of ranges or intervals. For example, after what establishment account date is a customer likely to not defect? Or what range of ZIP codes? Or what website access rate signals a "highly loyal" or profitable customer? Now, theoretically you could go about discovering these attributes and ranges manually; however, it would take you weeks of trial and error, rather than hours to discover the answers that can be achieved using a data mining tool incorporating a machine-learning algorithm or a neural network. Again, as the numbers of data attributes increase, so does the complexity of the problem, making data mining the clear alternative to fast and effective analysis for an electronic retailer.

Measuring loyalty, however, is difficult, especially on the Web, where the complexity of data analysis is compounded by the dynamic of time. If you are in a market where it costs little for a customer to switch products or services, such as financial services, telecommunications, or retail, then you know you are in a hyper-competitive market. Likewise it is so on the Internet, where it costs little for a visitor or customer to click away from your website and over to your competitor. Data mining provides an electronic retailer many benefits in such a competitive online environment. Firms in highly saturated markets like retailing and financial services are dependent on using their data to drive their business decisions. They depend on their data to retain their customers, expand their market share, and to differentiate themselves from their rivals. Likewise this is so on the Web, where competition is fierce and loyalty is ephemeral.

Traditional Data Mining Applications

A recent META Group survey revealed that Fortune 500 firms were using data mining for three general purposes: 64 percent for strate-

Amazon/AlmondNet
Exhibit 1011
Page 61

gic planning, 49 percent for competitive intelligence, and 46 percent to increase their market share. A study by this research group entitled *Data Mining: Trends, Technology, and Implementation Imperatives* (1998) forecasted a growth in the area of data mining driven by database marketing. It cites such applications as customer retention and churn management; cross-selling and up-selling; campaign management; market, channel and pricing analysis; and customer segmentation analysis. The study warns, "The competitive advantage enjoyed by the technology elite that deployed data mining early on is certain to dissipate through the Year 2000. Driven by the need for customer-centric marketing, data mining technology is rapidly becoming mainstream. Information technology and marketing departments that don't stay on top of this trend will wake up to discover that their competitors have beaten them to the punch." The report goes on to state that data mining "provides companies extremely valuable insight that has, to date, been closely guarded as corporate stealth technology to protect competitive advantage." So far, the following have been the traditional applications of data mining technology by industry sectors:

Retailers	Database marketing, advertising effectiveness, inventory and category management.
Banking and Insurance	Financial modeling, fraud detection, market and industry profiling, database marketing, customer segmentation.
Brokerage/Stock Exchange	Financial modeling, fraud detection.
Telecommunication	Call detail record analysis, optimal use of capital equipment, targeted marketing.
Government	Collections, workload selection, fraud detection, logistics, and intelligence gathering.
Manufacturing	Production quality control, supply-chain and inventory control.

The study fails to examine the potential growth of data mining in the area of electronic retailing. Customer-centric marketing requires sophisticated data mining applications because it seeks to better understand customers in their totality, including their buying patterns, demographic, and psychographic trends, and other types of customer knowledge. Such marketing will clearly accelerate with the advent of electronic commerce and the dynamic and explosive growth of Internet-generated databases.

Amazon/AlmondNet
Exhibit 1011
Page 62

New Data Mining Applications

In a web-centric networked environment, data mining is able to deliver new decision-support applications in two areas. First, data mining can be used to perform analysis on Internet-generated data from a website, which is what this book is about. And second, the technology can be used to monitor and detect the signature of network anomalies and potential problems before they happen.

One way of defining what data mining is about is to contrast it with more traditional methods of data analysis, such as statistics, Online Analytical Processing (OLAP), visualization, and web traffic analysis tools. We will now examine these various methods of data analysis and show that they are not exclusive of each other but are instead complementary, each contributing to the insight of your website visitors and customers and the effectiveness of your online presence.

Statistics

For centuries, man has been looking at the world and gathering data in an attempt to explain natural phenomena. People have been manually analyzing data in their search for explanations and patterns. The science of statistics originated as a way of making sense out of these observations. *Traditional statistics* have developed over time to help scientists, engineers, psychologists, and business analysts make sense of the data they generate and collect. *Descriptive statistics,* for example, provide general information about these observations, such as the average and median values, the observed errors, and the distribution of values. Another form of statistics is *regression analysis,* which is a technique used to interpolate and extrapolate these observations in an effort to make some intelligent predictions about the future.

Insurance and financial services companies were some of the first to try and predict customer behavior and to do risk analysis. Banks and insurers have typically used regression models to rank their customers' overall value or risk, all of which are attempts to fit a line to an observed phenomenon. SAS has built an empire by providing a host of enterprise-wide statistical applications to the financial services market sector. Fair Isaac, a modeling company, provides a wide assortment of services using statistics in the areas of behavior and their FICO credit scoring for financial services firms and other companies that do not do in-house analysis. Claritas provides demographic data to financial service firms designed to help them estimate sales potential of customers and segment markets by age, income,

Amazon/AlmondNet
Exhibit 1011
Page 63

etc.; it sells its P$YCLE demographic data to financial companies, which allows them to differentiate households in terms of expected financial behavior.

This kind of statistical data analysis involves analysts who formulate a theory about a possible relation in a database and then convert this hypothesis into a query; it is a manual, user-driven, top-down approach to data analysis. In statistics, you usually begin the process of data analysis with a hypothesis about the relation in the data, which is the reverse of how data mining works. Some of the most popular statistical tools include the following:

Data Desk uses animation to help you see patterns you might miss in a static display. For example, you can easily link a sliding control to any part of an equation and see the effects of sliding the value as display update. Data Desk automatically makes sliders to help you find optimal transformations of variables, to learn about the sensitivity of analyses to small shifts in variables, and to assess the sensitivity of nonlinear regressions. You can easily build your own animations.

MATLAB is a powerful integrated technical computing environment that combines numeric computation, advanced graphics and visualization, and a high-level programming language.

SAS is a modular, integrated, hardware-independent statistical and visualization system of software for enterprise-wide information delivery. SAS recently recognized the benefits of data mining and is now offering *Enterprise Miner* a data mining add-on module to its base system.

S-Plus is the commercial version of "S," an interactive, object-oriented programming language for data analysis developed by AT&T Bell Labs. It is supported and marketed by MathSoft. *S/S-Plus* has proved to be a useful platform for both general-purpose data analysis including clustering, classification, summarization, visualization, regression, and CART.

SPSS is a powerful, easy-to-use, easy-to-learn statistical package for business or research. It features most of the standard statistics, along with high-resolution graphics with reporting and distributing capabilities. SPSS also recognized the importance of data mining when it purchased Clementine, a data mining tool from ISL (Integrated Solutions Limited).

STATlab is an exploratory data analysis software for drilling down into data and performing a multitude of analyses. STATlab can import data from common formats such as relational databases, ASCII files, popular spreadsheets, and most of the popular statistical data systems.

Amazon/AlmondNet
Exhibit 1011
Page 64

Data Mining vs. Statistics

The deciding distinction between statistics and data mining is the direction of the query: In data mining, the interrogation of the data is done by the machine-learning algorithm or neural network, rather than by the statistician or business analyst. In other words, data mining is data-driven, rather than user-driven or verification-driven, as it is with most statistical analyses. Statistical manual factorial and multivariate analyses of variance may be performed in order to identify the relationships of factors influencing the outcome of product sales using such tools as SPSS or SAS. Pearson's correlation may be generated for every field in a database to measure the strength and direction of their relationship to some dependent variable, like total sales.

A skilled SAS statistician conversant with that system's PROC syntax can perform that type of analysis rather quickly. However, one of the problems with this approach, aside from the fact that it is very resource-intensive, is that the techniques tend to focus on tasks in which all the attributes have continuous or ordinal values. Many of them are also parametric; for instance, a linear classifier assumes that class can be expressed as a linear combination of the attribute values. Statistical methodology assumes a bell-shaped normal distribution of data—which in the real world of business and Internet databases simply is nonexistent and too costly to accommodate. However, these statistical tool vendors are well aware of these shortcomings; as both SPSS and SAS are now making available new data mining modules and add-ons to their main products.

Data mining also has major advantages over statistics when the scale of databases increase in size, simply because manual approaches to data analysis are rendered impractical. For example, suppose there are 100 attributes in a database to choose from, of which you don't know which are significant. With even this small problem there are $100 \times 99 = 9{,}900$ combinations of attributes to consider. If there are three classes, such as high, medium, and low, there are now $100 \times 99 \times 98 = 970{,}200$ possible combinations. If there are 800 attributes, such as in our large website bookseller customer database ... well, you get the picture. Consider analyzing millions of transactions on a daily basis, as is the case with a large electronic retailing site, and it quickly becomes apparent that the manual approach to pattern-recognition simply does not scale to the task. Data mining, rather than hindering the traditional statistical approach to data analysis and knowledge discovery, extends it by allowing the automated examination of large numbers of hypotheses and the segmentation of very large databases.

Amazon/AlmondNet
Exhibit 1011
Page 65

Online Analytical Processing

OLAP tools are descendants of query generation packages, which are in turn descendants of mainframe batch report programs. They, like their ancestors, are designed to answer top-down queries from the data or draw "what if" scenarios for business analysts. Recently, OLAP tools have grown very popular as the primary methods of accessing company database, datamarts, and data warehouses.

OLAP tools were designed to get data analysts out of the custom-report-writing business and into the "cube construction" business. The OLAP data structure is similar to a Rubik's Cube of data that an analyst can twist and twirl in different ways to work through multiple reports and "what-would-happen" scenarios. OLAP tools primarily provide multidimensional data analysis—that is, they allow data to be broken down and summarized by product line and marketing region. The basic difference between OLAP and data mining is that OLAP is about aggregates, while data mining is about ratios. OLAP is addition while data mining is division.

OLAP deals with facts or *dimensions* typically containing transactional data relating to a firm's products, locations, and times. Each dimension also can contain some hierarchy. For example, the time dimension may drill down from year, to quarter, to month, and even to weeks and days. A geographical dimension may drill up from city, to state, to region, to country, and even to hemisphere, if necessary. For example, Widget A, by Western Region, by Month of November, which can be further "drilled" by Blue Widgets A, by San Jose, by November 10, and so on. The data in these dimensions, called *measures,* is generally aggregated (for example, total or average sales in dollars or units, or budget dollars or sales-forecast numbers).

Many organizations have been in existence for some time and have accumulated considerable quantities of data that could be useful for business planning. Historical trends and future projections could be used to analyze business alternatives and make more informed decisions that could gain or maintain a competitive advantage. During the last decade, data warehouses have become common in large corporations, many of which use OLAP tools for reports and decision support. These OLAP applications span a variety of organizational functions. Finance departments use OLAP for applications such as budgeting, activity-based costing (allocations), financial performance analysis, and financial modeling. Sales analysis and forecasting are two of the OLAP applications found in sales departments. Among other applications,

Amazon/AlmondNet
Exhibit 1011
Page 66

marketing departments use OLAP for market research analysis, sales forecasting, promotion analysis, customer analysis, and market/customer segmentation. Typical manufacturing OLAP applications include production planning and defect analysis. The following are some of the current OLAP tools in the market:

Current OLAP Tools

ActiveOLAP, Analyzer, Aperio, Acuity/ES, Acumate ES, Advance for Windows, amis, Arbor Essbase Analysis Server, BrioQuery, Business Objects, Commander OLAP, Decision, Prism, Control, CrossTarget, Crystal Info for Essbase, Cube-It, FICS Group, datadriller, OpenAir Software, Inc., Dataman, DataTracker, dbProbe, DecisionSuite, DecisionView, Delta Solutions ALEA, Demon for Windows, DSS Agent, DynamicCube.OCX, EKS/Empower, Essbase/400, Express Server, Objects Fiscal, Fusion, FYI Planner, Gentia, Harry Cube, Helm, Holos, Hyperion OLAP, InfoBeacon, Informer, inSight, Intelligent Decision Server, IQ/Vision, Khalix, Lightship, Matryx, Media, Metacube, MIK-Solution, MineShare, MIT/400, MUSE, NetCube, NGS-IQ, OpenOLAP, Pablo, ParaScope, PowerPlay, Rapid OLAP, Sagent Datamart Solution, SAS System, SpaceOLAP, StarTrieve The Ant Colony, TM/1, Toto for Excel, Track Objects, Visualizer Plans for OS/2, VSFLEX

Data Mining vs. OLAP

With the proliferation of data warehouses, new data mining tools are flooding the market offering to discover hidden gold in your data. Many traditional report and query tools and statistical analysis systems are using the term "data mining" in their product descriptions. In the midst of this marketing activity it is sometimes difficult to know what a data mining tool is and how it works. Perhaps a definition would help in clearing the confusion.

The methodology of data mining involves the extraction of hidden predictive information from large databases. However, with such a broad definition as this, an online analytical processing OLAP product could be said to qualify as a data mining tool. That is where the technology comes in, because for true knowledge discovery to take place, a data mining tool should arrive at this hidden information *automatically.*

Still another difference between OLAP and data mining is how they *operate* on the data. Similar to the direction of statistics, OLAP is a top-down approach to data analysis. OLAP tools are powerful

Amazon/AlmondNet
Exhibit 1011
Page 67

and fast tools for *reporting* on data, in contrast to data mining tools that focus on *finding patterns* in data. For example, OLAP involves the summation of multiple databases into highly complex tables; OLAP tools deal with aggregates and are basically concerned with addition and the summation of numeric values, such as dollar amounts or cash. Manual OLAP may be based on need-to-know facts, such as regional sales reports stratified by type of businesses, while automatic data mining is based on the need to *discover* what factors are influencing these sales.

An OLAP tool is not a data mining tool since the query originates with the user. Neural networks, machine-learning, and genetic algorithms, on the other hand, do qualify as true automatic data mining tools because they autonomously interrogate the data for patterns. This is known as *supervised learning,* while another less common form of data mining is called "clustering" or *unsupervised learning*. In both instances, however, the bottom-up approach to data analysis distinguishes data mining from OLAP.

The current crop of OLAP tools have tremendous capabilities for performing sophisticated user-driven queries, but they are limited in their capability to discover hidden trends and patterns in a database. Statistical tools can provide excellent features for describing and visualizing large chunks of data, as well as performing verification-driven data analysis. Autonomous data mining tools, however, based on artificial intelligence (AI) technologies, are the only tools designed to automate the process of knowledge discovery—and to optimize this with their "built-for-business" features:

- Robustness for handling any kind of data: noisy, missing, mixed
- Exportable solutions in the form of practical business rules
- Easy to understand results via graphical decision trees
- Accurate results, especially with real-world data sets

Data mining by this definition is thus data-driven, not user-driven or verification-driven.

Traditionally, the goal of identifying and utilizing information hidden in data has been achieved through the use of query generators and data interpretation systems. This involves a user forming a theory about a possible relation in a database and converting this hypothesis into a query. For example, a user might have a hypothesis about the relationship between the sales of color printers to business customers. A query would be generated against the user's data warehouse or website inventory database and segmented into a report by client accounts using Standard Industry Codes and quarterly sales.

Amazon/AlmondNet
Exhibit 1011
Page 68

Certainly the information generated will provide a good overview on the relationship, if any, between types of businesses and sales of printers. However, this verification type of data mining is limited in several ways. First, it is based on a hunch—that there is a relationship between the industry a company is in and the number of printers it buys or leases. Second, the quality of the extracted information is based on the user's interpretation of the results and is thus subject to error.

In another example, an OLAP tool can tell a bookseller about the total number of books it sold for a region during a quarter. Data mining, on the other hand, can tell it about *the factors influencing* the sales of the books. A typical OLAP report is an aggregate count from several tables:

Category	Amount	Percentage Variance from Budget
Product Sales		
Fiction	1,847,743.00	2.13%
Nonfiction	606,735.00	2.03%
Periodicals	807,987.00	0.19%
Accessories	532,585.00	1.77%
Total Sales	3,262,465.00	1.76%
Margin	2,189,687.00	1.64%
Total Expenses	1,262,312.00	2.04%
Profit	927,375.00	1.07%

Thus, OLAP can tell you about the total number of books sold for this region during this quarter. Data mining can tell you about the patterns of book sales, while statistics can provide another dimension about these sales. Online analytical processing and statistics provide top-down, query-driven analysis, while data mining provides bottom-up, discovery-driven analysis. Data mining requires no assumptions. Rather, it identifies facts or conclusions based on patterns discovered.

Website Analysis Tools

With a firm's website increasingly becoming the first point of contact between a potential client and that company, its importance has increased. Today websites demand full-time design and maintenance

Amazon/AlmondNet
Exhibit 1011
Page 69

teams and multimillion-dollar budgets. Management must know who's visiting and, more important, who is buying. In addition, companies that place ads on a site want to gauge their effectiveness. Websites today must earn their keep as business investments and, just like any other marketing effort, they must justify themselves. That's where some of the current website analysis tools come in.

Every time you visit a site, the web server enters a valuable record of that transaction in a log file. Every time you visit an electronic commerce site, a cookie is issued to you for tracking what your interests are and what products or services you are purchasing; every time you complete a form at a site, that information is written to a file.

Although these server log files and form-generated databases are rich in information, the data is itself usually abbreviated and cryptic in plain text format with comma delimiters, making it difficult and time-consuming to mine them. The volume of information is also overwhelming: A one-megabyte log file typically contains 4,000 to 5,000 page requests. Website analysis tools typically import the log-file data into a built-in database, which in turn transforms the data into aggregate reports or graphs.

This information can be fine-tuned to meet the needs of different individuals in your firm. For example, a web administrator may want to know about the clicks leading to documents and images, files, scripts, and applets. A designer will want to know how visitors navigate the site and whether there are paths or points from which many visitors jump to another site. The marketing team will want to know the effectiveness of certain promotions. Advertisers and partners may be interested in the number of clickthroughs your website has generated to their sites. Most website analysis tools provide answers to such questions as:

- What are the most common paths to the most important pages on your site?
- What keywords bring the most traffic to your site from search engines?
- How long does it take visitors from Texas to view your homepage?
- How many pages do visitors typically view on your website?
- How may visitors are you getting from Europe or Asia?
- How much time do visitors spend in your website?
- How many new users visit your site every month?

Most tools provide graphical views of your website traffic via different criteria and, like statistical and OLAP tools, provide the same

Amazon/AlmondNet
Exhibit 1011
Page 70

type of aggregate views of web-generated data. Figure 2–3 is a report showing how many visits on a website by the hour in a day. A "visit" is a series of clicks by the same user. Figure 2–4 is a website traffic report showing the organizations or search engines that sent visitors to this site most frequently. There must be referral information in the log files for this report to be generated.

Obviously the more you know about how many people are visiting your site, who they are, and what parts of your site they are using, the better off your site will be. Using this information to improve your site and relating it to other traditional marketing programs will help you refine your overall website design. The average businesses today spends anywhere from $25,000 a year for a standard intranet site to $1.25 million for a full-featured website that supports electronic commerce, according to recent estimates from International

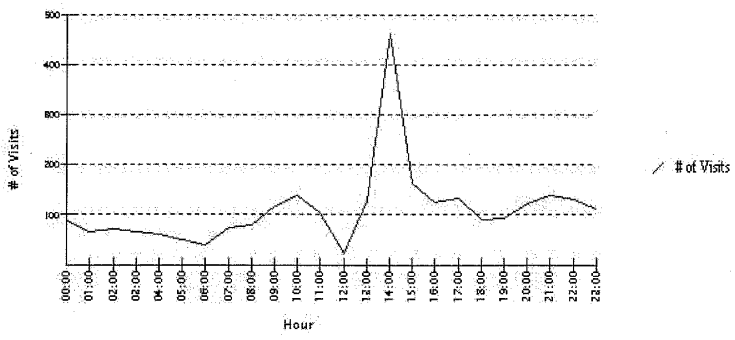


Figure 2–3
A graphical report of website daily "hits."

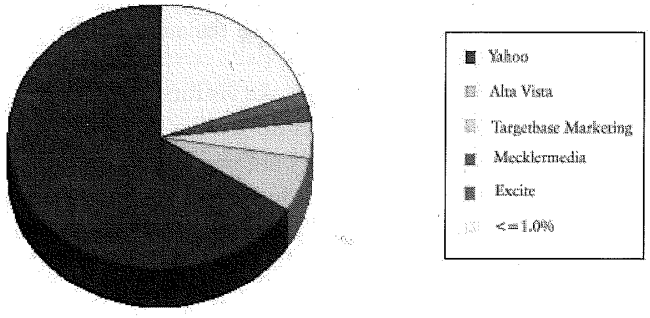


Figure 2–4
A graphical report of referring websites and search engines.

Amazon/AlmondNet
Exhibit 1011
Page 71

Data Corporation. With costs such as these it is common for a justification of these expenditures to be made. One way of protecting these huge investments is through the monitoring of web traffic.

The best—and sometimes only—way to glean that information in a usable format is with traffic analysis and monitoring software. The software runs the gamut in functionality and price, as tools range from free to upward of $100,000. On the low end, basic packages and downloadable utilities provide entry-level analysis for small- to medium-volume sites and intranets. These packages can tell you such information as how many visitors came to the site, the most- and least-requested pages, where visitors entered and exited, and other notable path analysis and historical trend data. High-volume websites involved in electronic retailing most likely will opt for more sophisticated high-dollar software that provides comprehensive analysis of traffic flow as well as detailed information about each visitor and their trek through the site. Knowing who enters—and who buys your product—can be critical to the success or failure of your electronic business.

"Log file analyzers" do as their name suggests—they take information recorded by your server and summarize the data in an intelligible format. Most analyzers compile data into a database from which they can run reports; their speed is measured in terms of both loading data into the database and compiling the report. Certain programs load in real time by "sniffing" every packet and stuffing the database. Others load the log file into the database at off-peak times specified by the administrator. Complex commerce sites may also require analysis tools with the capabilities to conduct qualitative and quantitative analysis of marketing and advertising data. If a company runs an ad on three different websites, a tool should be able to compare how many people were referred from each site, or which links were most commonly selected.

Lastly, for sites with search capabilities, some web analysis tools can report on which keywords visitors used, which sites have links to you and provide the majority of users, and which search engines provided the most valuable visitors. All of this information is readily available from log files. Since the first action for most visitors to a website is a search, it is sometimes very important to know what they're looking for. If you know that information along with where they're coming from, where they're going, and how long they will linger, then your site can truly become an effective marketing engine. The following is a partial listing of some of these web analysis tools, along with a discussion of some of their features and limitations:

Amazon/AlmondNet
Exhibit 1011
Page 72

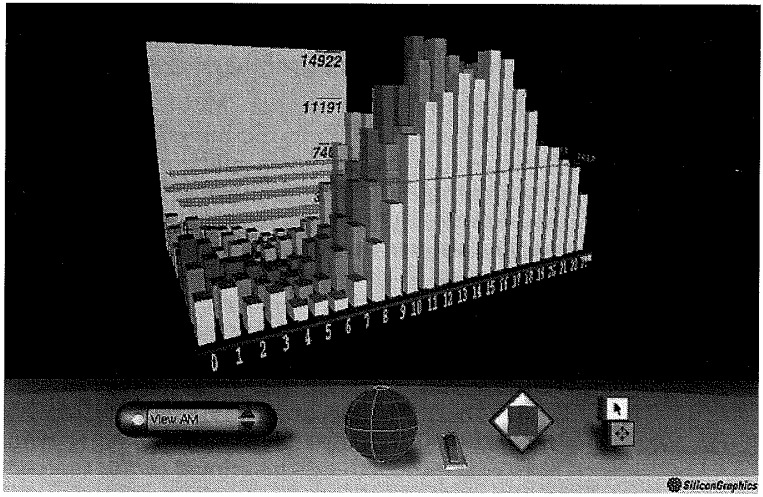


Figure 2–5
This is a three-dimensional view of website traffic for a single month.

3Dstats is an Access Statistics Generator. It analyzes web server log files and creates a Virtual Reality Markup Language (VRML) model with the average loads by day and hour. With a VRML viewer like CosmoPlayer, the user can "walk" through the statistical "scene" and examine the bar chart from different points of view for daily, weekly, or monthly reports. See Figure 2–5.

Accrue Insight analyzes packets of information in real time (also known as *packet sniffing*). Insight gauges performance according to how fast a server responds to requests. It can collect and store data on more than a million hits per hour, and the resulting data is stored in the most granular form possible to allow for the widest variety of sorting, analysis, and report-generation possibilities. The product is especially good at tracking click streams through a large commercial multi-server site.

Aquas Bazaar Analyzer is based entirely on Java, which means reports can be viewed with any browser. With this analyzer, the log file needn't be transferred to the viewer's machine to be massaged. Reports are generated in your browser window, with graphs appearing as customizable Java applets.

Aria uses agents to analyze real-time traffic right on the server and generates real-time reports with plenty of information. The Aria architecture consists of three parts: the monitor, the recorder, and the reporter. The monitor takes the form of server agents that intercept traffic from the web server; the recorder receives

Amazon/AlmondNet
Exhibit 1011
Page 73

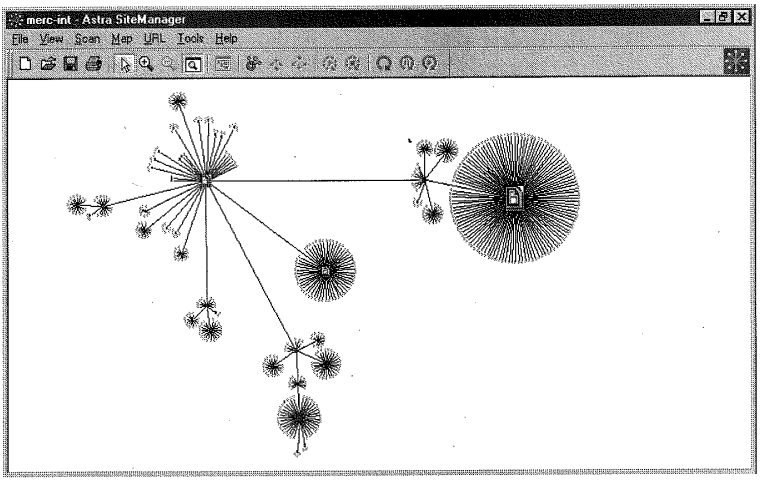


Figure 2–6 *Mercury Interactive's Astra SiteManger dynamically link maps your web-site activity.*

and processes the data; and the reporter presents the information in charts. The monitor module records information about the activity between the web client and server, including request data, form-field posted data, server-response data, cookie data, and user-profile data. It stores the information in an object-oriented database that comes with the software. No other database is needed.

Astra checks and validates all of your site's links and shows you any access problems you may be having. Your site is turned into a visual tree-diagram map, complete with links to CGI scripts and Java applets. Bad links are displayed as red points on the map for quick access and repairs. You can use Astra SiteManager as a log analysis tool: It reads your server's log file and superimposes usage patterns directly over your site map, showing you exactly which pages are getting the most clicks. (See Figure 2–6.) Astra SiteManager scans your entire site, highlighting functional areas with color-coded links and URLs to give you a complete visual map of your website.

Hit List Standard and *Hit List Pro* from Marketwave offers a wide variety of 27 preconfigured reports that help you track your web traffic. Hit List creates its own relational database of your server logs instead of working directly from the logs themselves. All reports can be redesigned through the design feature. Hit List Standard tracks user paths to and through a site and tallies download attempts. Hit List Pro can tell you the average number of

Amazon/AlmondNet
Exhibit 1011
Page 74

pages that users coming from a certain site tend to view, and it can parse search strings from other sites, such as Yahoo!, to determine exactly why and how a user comes to your site. It also offers extensive querying capabilities for clickthrough rates of specific online advertisements.

Intersé offers tracking software with several standard report types detailing hits, visitors, browser types, and traffic patterns. This tool creates its reports from log files off-line, which is good if your provider does not want your software on its network and if monthly reporting is all you need. You can cross-reference data from multiple websites or expand your usage analysis by integrating data from additional sources such as demographic and psychographic registration forms.

net.Analysis Pro, by net.Genesis, offers marketing and technical reports. It provides extensive path analysis and entry and exit page reports, as well as some demographic information. net.Analysis Pro not only reports top domains accessing the site and the most popular geographical areas, but it correlates them into a single report. This information is limited in value by large Internet service providers (ISPs), like EarthLink and FlashNet, that serve users throughout the country.

NetIntellect is another log-file analysis tool. Through compression, the NetIntellect database is as much as 95 percent smaller than your original server logs. Compression is essential for reports spanning weeks, months, or even years, because without it the combined files get too large for processing. NetIntellect lets you process more than one log file at a time, and it provides a feature for comparing log files as well.

NetStats Pro has a query builder with which you can construct your own customized log reports. You can filter whatever fields you wish either into or out of the report (complete with multiple Boolean expressions and time/date ranges). The program can analyze local log files or download files from your server machine through FTP. The standard report includes details on the top three domain extensions from which your visitors come, the files they access, and the activity across the day and the week.

NetTracker is another Java-written analyzer. Users do not need to download log files to their machines, nor do they need to run any software besides their browsers.

PressVu is an log analysis tool for small peer websites running over dedicated dial-up connections using Microsoft PWS on Windows 95 or NT Workstation. PressVu Desktop imports plain text source

Amazon/AlmondNet
Exhibit 1011
Page 75

logs into an xBase file format database. The PressVu main interface supports easy access to the InterNIC WHOIS database and DNS hostname lookup.

SurfReport is a unique web traffic analysis tool in that it can generate reports via e-mail or your browser. SurfReport runs on virtually any platform and any operating system and is completely server-based, allowing the generation of reports from anywhere on the Web.

Webtrends lets you scan your server statistics almost immediately, working in the background to compile the information it needs. It doesn't require lengthy initial importing of log files into its database format, but instead creates its first reports on the fly as the log files are being read in. WebTrends works primarily on server log files, but it also allows you to capture cookie information, referring sites, browser identity, and user details.

Website Analysis Tools vs. Data Mining Tools

Modern web analysis tools provide the same type of views of data as statistical and OLAP tools. They emphasize aggregate counts and spatial views of website traffic over time, which like statistical and OLAP tools, are not easily able to discover hidden patterns without extensive interactive queries.

Because of the massive amount of data that even the simplest log file contains, the process of pattern recognition can be extremely difficult if not impossible with this generation of tools. As with statistical and OLAP tools, web analysis tools are verification-driven. They operate from the top down rather than self-organize themselves around the data via the use of a machine-learning algorithm or a neural network. Like OLAP tools, web analysis tools are good at generating aggregate reports via several dimensions, which are important to management, web administrators, and designers, who are limited to measuring browser, not shopper activity.

Daily Totals for http://www.webminer.com/

Date	*Clients*	*Hits*	*Pages*
Mon Oct 28 1999	196	1291	407
Tue Oct 29 1999	165	1092	330

Amazon/AlmondNet
Exhibit 1011
Page 76

Period Totals for http://www.webminer.com/

Individual Clients	*Total Hits*	*Page Hits*
353	2383	737

Daily Averages for http://www.webminer.com/

Clients	*Hits*	*Page Hits*
181	1192	369

However, almost all web analysis tools lack the capability to identify important market segments from web-generated data. With the exception of Aptex, which is really a text-analysis software product incorporating neural network and proprietary modeling technology, no web analysis tool does pattern recognition. (Aptex is developing neural network predictive models to do "intelligent advertising" in such sites as Netscape, Excite, and InfoSeek.)

Almost all of the current web analysis tools do very good and innovative data reporting via tables, charts, graphs, and 3D VRML. Several vendors, like Astra, are beginning to explore advanced forms of visualization such as link analysis. Some, like NetTracker and SurfReport, are taking advantage of the Web to deliver their results via thin client browser by writing their reports tools in Java. Still other firms are offering network services by which your website traffic data is transferred to their servers, like I/PRO's *NetLine* for third party analysis and auditing.

Today's advertising-funded websites rely on usage statistics like TV networks rely on Nielsen rating points. Hit traffic counts, the traditional measure of a website's effectiveness, is all but useless to advertisers. Instead of counting the number of times a page is viewed, they are more interested in tabulating every inline graphic, Common Gateway Interface (CGI) script, and Java class resulted in a click-through or sale. Advertisers need to know not only how many pages are viewed, but also how many people visit their website. These marketers need to know the demographics of their visitors, so they can sell ads. For starters they want to know who is accessing the website and where these visitors are coming from. Some of this information is available from log files, like referral information, but some is not. It is at this juncture that data mining tools and techniques can assist the web marketer and electronic retailer.

A data mining tool does not replace a web analysis tool, but it does give the web administrator a lot of additional possibilities for answer-

Amazon/AlmondNet
Exhibit 1011
Page 77

ing some of the marketing and business questions. For example, imagine trying to formulate answers to questions such as:

What is an optimal segmentation of my website visitors?

Who is likely to purchase my new online products and services?

What are the most important trends in my website visitors' behavior?

What are the characteristics or features of my most loyal online clients?

Theoretically, these questions could be answered with a web analysis tool. For example, a web administrator could try to define criteria for a customer profile and query the data to see whether they work or not. In a process of trial and error, a marketer could gradually develop enough intuitions about the distinguishing features of its predominant website customers, such as their gender, age, location, income levels, etc. However, in a dynamic environment such as the Web this type of analysis is very time-consuming and subject to bias and error.

Manual processing in this fashion could take days or months to find the optimal customer profile for this dynamic market segment from a large retailing site. On the other hand, a data mining tool (such as a decision tree generator) that incorporates machine-learning technology could find a better answer automatically, in a much shorter time—typically within minutes. More importantly, this type of autonomous segmentation is unbiased and driven by the data, not the analyst's intuition.

For example, using a data mining tool, a log file can be segmented into statistically significant clusters very quickly. The decision tree in Figure 2–7 was generated by a data mining tool incorporating several symbolic classifiers, including ID3 (Interactive Dichotomizer, a machine-learning algorithm) and two statistical algorithms, CHAID (CHI-squared Automatic Interaction Detection) and CART (Classification and Regression Tree). A unique feature of the tool is that it runs CART and CHAID concurrently, which means both continuous and discrete dependent and independent variables can be modeled without preprocessing or scaling. The tool generates Multivariate Decision Trees, IF-THEN rules, and SQL syntax.

In Figure 2–8 a data mining tool was able to generate a decision tree identifying two important features about its website visitors. It identified key *attributes,* the type of computer products being sold by this website and *intervals,* the income dollar

Amazon/AlmondNet
Exhibit 1011
Page 78

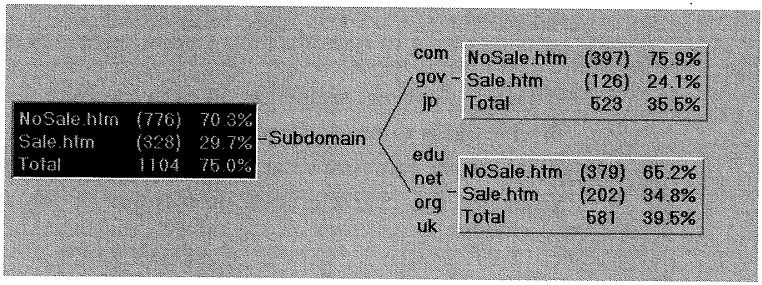


Figure 2–7
Data mining a website can reveal hidden ratios for online sales.

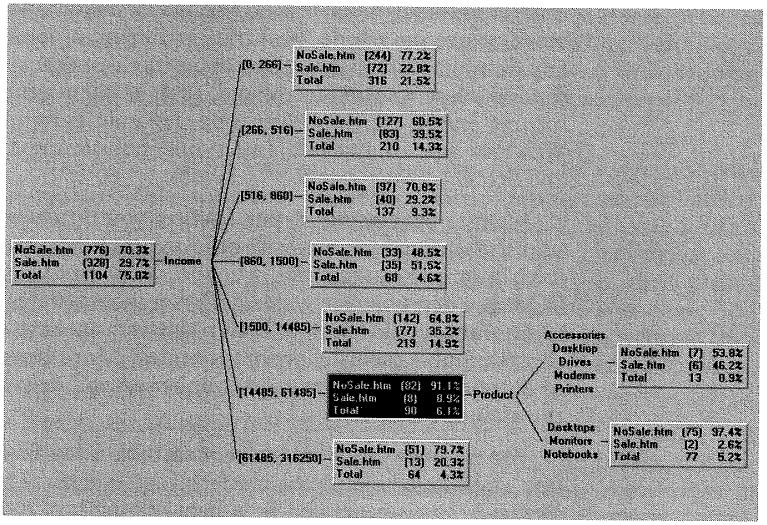


Figure 2–8
The leaves of a decision tree represent market segments of online customers.

ranges—differentiating between customers who made online purchases and those who did not.

What is important to note is that these clusters, which represent market segments, were discovered by a machine-learning algorithm in a matter of minutes rather than by a web administrator's manual analysis of the reports. Data mining does not replace data analysis; it simply automates it. Equally important to note is that while a data mining tool may discover a cluster or a potentially profitable market segment, a web analysis tool can provide the needed analysis to perform the projected profits from each of these market segments. A web analysis statistical tool can provide the charts to visually represent the findings to management about the significance of these newly discov-

Amazon/AlmondNet
Exhibit 1011
Page 79

ered market segments. In conclusion, you should use all of the tools available to you. As in the construction of a building, several types of tools are required, each designed for specific tasks and yet each contributing to the completion of the total project.

Visual Data Mining

An optional method by which to perform data mining is through the use of visualization tools. The value of these tools is gained when used in conjunction with other tools such as neural networks and machine-learning algorithms. Some of the features of these tools are being incorporated in current web analysis tools such as Mercury Interactive's Astra SiteManager, which uses a form of link analysis.

The sections that follow list several of these systems and discuss their unique capabilities, including features, functions, user interfaces, volume of information, customization, integration issues, and typical uses. Of interest is of course how they can be used to supplement your insight about your website patterns in conjunction with other data mining tools. These visualization systems fall into three general classes: visualization packages, link analysis systems, and quantitative displays.

Visualization Packages

Information visualization uses abstract representations in interactive, immersive, 3D, virtual environments to display large quantities of data. It is one of the best methods for proactively exposing trends within a data warehouse (and website data), which is usually accomplished by navigating data landscapes and visually orienting the data to reveal hidden biases. Information visualization systems are also geared toward support of real-time applications, since parametric values can be displayed as an animated or simulated dimension of data. The following tools are representative of these visualization packages: *SpotFire, Visual-ID (Lucent Technologies), Visible Decisions* (VDI), and *Metaphor Mixer* (Maxus Corporation).

Figure 2–9 is a view of customers of a department store that utilizes a visualization tool to design a direct mailing campaign based on several client features, such as ZIP code, gender, and age. A tool such as this can be used to examine the demographics of your website visitors in order to explore possible pockets of geographic interest by ZIP codes or other criteria. This tool provides an interactive feature that allows the user to move the sliders on the right and view the

Amazon/AlmondNet
Exhibit 1011
Page 80

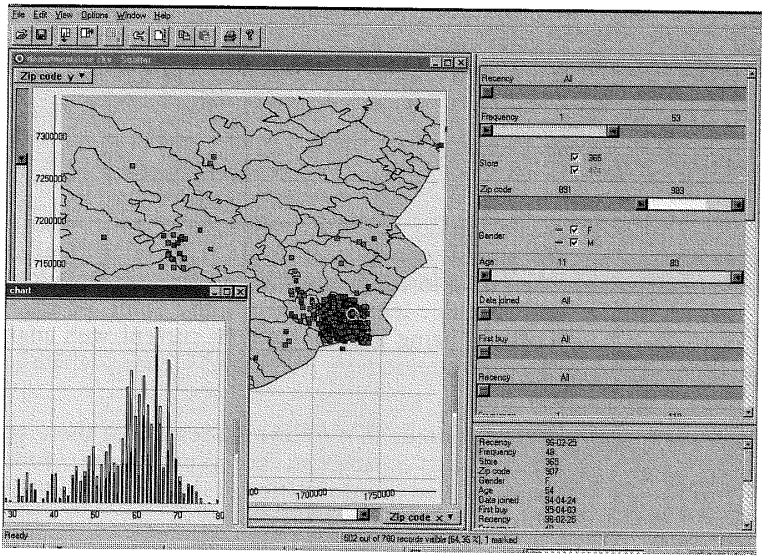


Figure 2–9 *Variables, like gender, can be moved on the right and viewed on the chart and map.*

changes between multiple variables, such as frequency of purchase and receipt average, both on the map and on the chart on the left.

Link Analysis Systems

Link analysis is based on a branch of mathematics called *graph theory,* and uses a combination of binary relationships established between object representations to generate networks of interactions and associations from defined data sets. Link analysis is not applicable to all types of data nor can it solve all types of problems. The aggregate number of data records presented using link analysis is somewhat limited and the analyses tend to focus on verifying sets of related information. Different layouts, filter assessments, and presentation formats provide link analysis systems with the candid ability to quickly identify patterns that occur from unusual relationships, emerging groups, and important connections.

Some of the areas where link analysis has yielded good results are in analyzing telephone call patterns and combining leads of information sources. A cellular phone company may use a link system in order to analyze call patterns for purposes of discovering how customers differ. The FBI is using a link analysis system to combine information from disparate sources to help the bureau in solving crimes. Link analysis is able to capitalize on relationships and derive

Amazon/AlmondNet
Exhibit 1011
Page 81

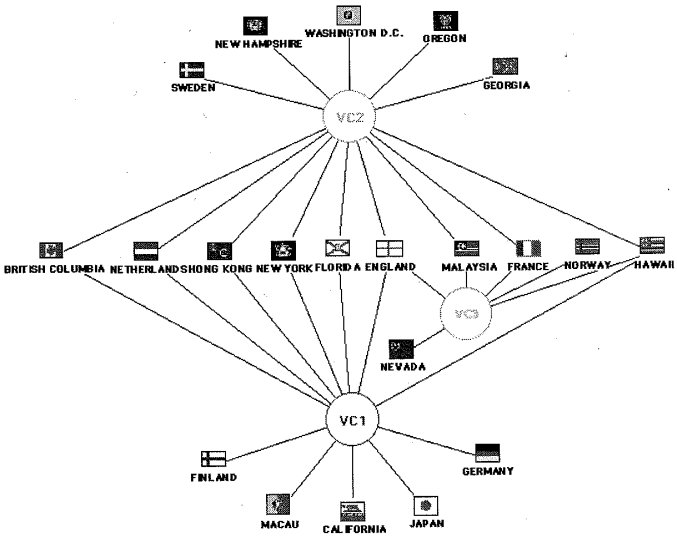


Figure 2–10
A link analysis tool discovered the behavior and relationships in a fraudulent scheme.

characteristics that normally may be missed by traditional methods of analysis.

A link analysis tool can be used to identify relationships between certain products or web pages in your site. It can also provide a visual insight into other hidden relationships based on the demographics of your website visitors. (See Figure 2–10.) However, one major drawback to link systems is their inefficient performance in relational databases. The following tools are representative of link analysis type systems: *NETMAP* (ALTA Analytics), *IMAGIX-4D* (Imagix Corporation), *Analyst Notebook* (I2), *WATSON* (Harlequin), and *GVA* (United Information Systems).

Quantitative Display

Quantitative diagrams can handle extremely large volumes of data, typically containing numerical representations. These diagrams tend to have a traditional statistical look and feel to them for clustering, summarization, and range comparisons. Recent implementations have pushed the limits of these diagrams to perform hyper-dimensional analyses, which make them ideal candidates for identifying *linear* or exponential trends occurring within data sets. (See Figure 2–11.) The following tools represent quantitative display systems: *DIAMONDS* (IBM/SPSS), *CrossGraphs* (Belmont Research), and *Temple-MVV* (Mihalisin).

Amazon/AlmondNet
Exhibit 1011
Page 82

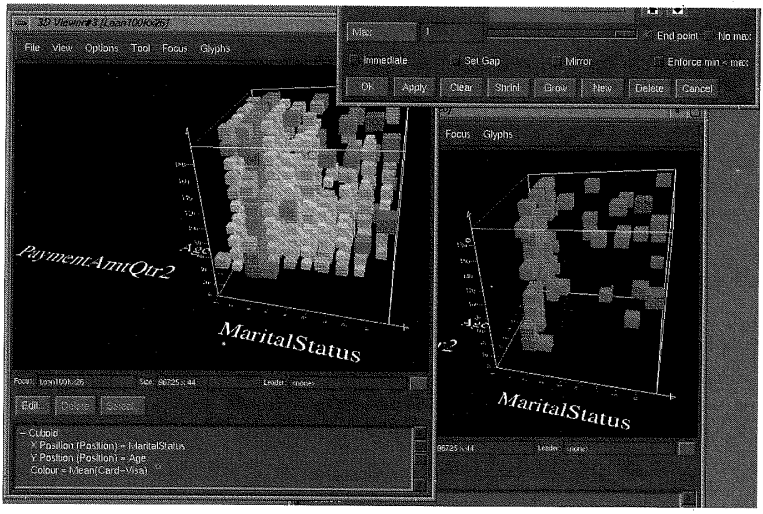


Figure 2–11
In a 3D environment, layers of noisy data can be removed to find clusters of hidden knowledge.

Data Mining Solutions

Data mining can confirm marketing and management observations about your website and its visitors. Through data mining you can quantify these observations so that you know at what point in time sales are likely to occur, or what dollar ranges exist between profitable and unprofitable web sessions. Data mining can lead to subtle improvements in efficiency as well as breakthrough findings in the design and structure of your website. Through the mining of your website you may find that, yes, your best customers are between age 36 and 42, or you may discover to your surprise that your products and services do best with women and not men, as you originally thought! The information that you gather from your registration forms and the mining of the content in your database can tell you this and much more, such as what products to associate for increase cross-selling opportunities, or what information links or advertising to provide your visitors based on their gender, age, demographics, and lifestyle interests. Local sources can also be associated based on ZIP code culled from these registration forms.

As previously stated, data mining is more than the use of software tools, it is a business *process* designed to extract actionable intelligence for product and financial analysis, sales, marketing support, and customer segmentation gathered from your dynamic customer information file. Data mining is about knowledge discovery for

Amazon/AlmondNet
Exhibit 1011
Page 83

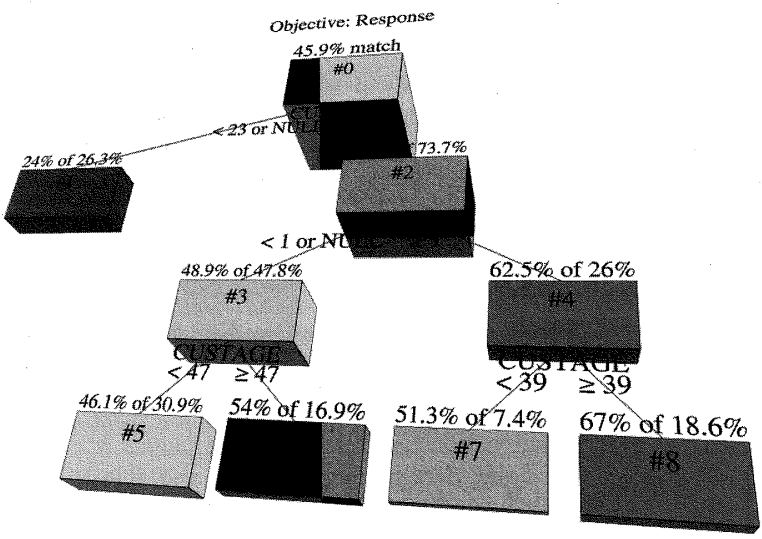


Figure 2–12
One common data mining solution is a decision tree, which segments a database.

strategic advantage. The format of a data mining analysis or data mining tools may vary. For example, it may be in the form of visualization, such the graphical decision tree shown in Figure 2–12.

The results of data mining can also be in the form of IF-THEN rules, which are usually market segments extracted from your website data:

IF	website customer age is less than 47
THEN	response to offer is 46.1%
IF	website customer age is greater than or equal to 47
THEN	response to offer is 54%
IF	website customer age is less than 39
THEN	response to offer is 51.3%
IF	website customer age is greater than or equal to 39
THEN	response to offer is 67%

Still another possible solution available from the data mining process is C code, which usually is representative of a set of weights, or formulas. These weights are usually extracted from a data mining tool incorporating a neural or polynomial network as its core technology:

Amazon/AlmondNet
Exhibit 1011
Page 84

```
**************/
double *webminer(double inarr[])
{
/* node5-Successful_Suburbanites */
    node5  = -1.19493 + 0.032543*LIMIT(Successful_Suburbanites,0,82.9);
    /* node31-Young_Frequent_Movers */ node31  = -0.22708 +
   0.0943097*LIMIT(Young_Frequent_Movers,0,100);
    /* node32-Rural_Industrial_Workers */
    node32  = -0.367938 + 0.0455085*LIMIT(Rural_Industrial_Workers,0,100);
    /* node51-Triplet */
    node51  = 0 - 22.0566 - 40.4871*node5 - 18.3542*pow2(node5)
          + 0.174813*pow3(node5) - 68.848*node31
          - 156.303*node5*node31 - 83.2704*pow2(node5)*node31
          - 44.5781*pow2(node31) - 37.7005*node5*pow2(node31)
          - 0.0442751*pow3(node31) - 64.3699*node32
          - 54.3879*node5*node32 - 185.61*node31*node32
          - 156.467*node5*node31*node32 - 0.270547*pow2(node31)*node32
          - 0.109251*pow2(node32) - 0.272628*node31*pow2(node32)
          - 0.00897233*pow3(node32) ;
```

Some data mining tools are also able to generate SQL syntax, which can be exported from the analysis and used in a production system to extract a set of records in your databases, which meet the parameters of the query:

```
/* SQL 1 */
select * from table
where
    (
       High Rise Renters = 0 or
       High Rise Renters = 0.1 or
       High Rise Renters = 0.6
    );
/* SQL 2 */
select * from table
where
    (
       High Rise Renters = 1.7
    );
/* SQL 3 */
select * from table
where
    (
       High Rise Renters = 11.7 or
       High Rise Renters = 28.4
    );
/* SQL 4 */
select * from table
where
    (
       High Rise Renters = 51.3 or
       High Rise Renters = 83 or
       High Rise Renters = 91
    );
```

Amazon/AlmondNet
Exhibit 1011
Page 85

Keep in mind that the data mining process and the multiple formats of its solutions all have one single goal: insight into the behavior of your online visitors and customers.

Why and Who, Not What

Data mining is about *why,* not *what*. Data mining is about discovering *why* the total sales in your website are up for a particular product or service, or *why* certain website visitors have purchased from your online catalog certain items from certain ZIP codes and age ranges, while others have not. It is about *why* certain website visitors are likely to continuously purchase your product and services, while others will not. Data mining is about discovering *why* the sales figures in your website traffic activity reports are as they are, and as such, it is a much more powerful technology than simple aggregate domain counts.

Data mining is also about *who*. Who visits your website and buys your products and services, who your most loyal and profitable online customers are. Data mining is about ratios, rules, attributes, features, patterns, signatures, and influences hidden in your website data. It is about discovering the reasons or relations behind the numbers in your website traffic activity reports. It is also about uncovering the core differences between your website visitors, online customers, services, products, and sales. Data mining is about *why* and *who,* not *how many.* It can be used to answer such questions as:

- What makes a website visitor profitable or unprofitable?
- What kinds of website visitors respond to your advertising?
- What traits do profitable visitors have in common?
- Why did this online product sell with this visitor?
- What will these website visitors buy next week?

Data mining is about using autonomous pattern recognition technology for achieving specific business objectives, such as increasing your website sales. It is about *perception, learning,* and *evolving*—all human functions that can assist your company in recognizing the fast-changing electronic commerce environment. For example, data mining is about increasing your market share by recognizing who your most valuable online customers are—understanding their features and using that knowledge to identify new website visitors through targeted marketing and advertising.

Amazon/AlmondNet
Exhibit 1011
Page 86

The Online Corner Store

With detailed information about the customer's habits at the point-of-contact, a website can provide individualized interaction, much like the neighborhood store. For example, a bookseller website can display titles matching an individual consumer's preferences. With growing personalized attention during every transaction, a website customer will perceive added value in their relation with that bookseller. The focus of this type of individualized marketing is on owning a greater market share of an individual customer over time. A clear advantage to relational marketing is that during the process of increasing the share of each customer, the company is both building a long-term relationship and learning about that customer with every transaction. This has an additional benefit since it is cheaper to increase sales to existing customers than to acquire new customers.

The website can track customers, it can enable interactive dialog, and it can allow mass customization of products and services created and presented to an individual's unique needs and requirements. The key is to know about as much as possible about each of your website customers. Dialog is the most important aspect of relational marketing. You need to facilitate website visitor communications—act on their feedback and reciprocate by making suggestions and recommendations—in order to build trusting and loyal relationships, which in turn will translate into more online sales and profits.

The use of lifestyle data, or "psychographics," became prevalent in the 1980s, and in the 1990s Regis McKenna introduced the concept of relationship marketing. McKenna suggested moving away from marketing to everybody and instead concentrating on knowing your customers and developing a relationship with them. McKenna's marketing concepts are about developing products and services to serve a specific market, by defining the standards in that market and developing a deep relationship with those customers. More important, relational marketing is a concept ideal for the Web and electronic retailing. The rapid decline of computer prices, coupled with their increased performance and storage capacity, has made the practice of using databases in marketing almost commonplace during the last decade.

A more recent concept further narrows the marketing from segment to that of individuals. Marketing visionaries Peppers and Rogers in their book *The One-to-One Future* actually saw the marketing environment a full year before the Web burst into the scene. In their book they introduced the concept of marketing to an individual on a

Amazon/AlmondNet
Exhibit 1011
Page 87

one-to-one basis over an extended period of time and in the process learning as much about them as possible. Today, like the little corner store, the web merchant can remember what products you like, stock for your convenience, then e-mail you product-specific offers.

Selling as a Service

The goal of selling on the Web is to increase repeat purchases by focusing on brand recognition and customer profitability over their lifetime. Selling on the Web is an opportunity to build loyalty, which will reduce customer attrition and build a higher rate of profit per customer. Because each transaction requires less cost than the initial one, loyal clients are the most profitable, as they tend to spend more over time. Loyalty is fostered by feedback, which should be facilitated either via CGI forms and e-mail so that your clients can tell you who they are and what they want from your website.

In the process of interacting with your website customers, it is important that you assure them of the confidentiality of the information they are disclosing. Keep in mind that customer information is your competitive arsenal and should not be shared or provided to another firm without your clients' approval. On your registration forms, ensure you don't ask for all of the information at once and that you give your website customers an option over what information they provide you. Lastly, give them something in exchange for the information they do disclose to you, such as a special shopping area or a discounted pricing rate.

Data mining performs many functions: it can assist you in assessing the performance of your website and in the integration of your website visitor profiles; it can help you in segmenting your website customers by profitability, loyalty value, overall usage, and number of visits; it can also assist you in the construction of predictive models for targeting future clients.

There is a continuum in this process of data acquisition. For example, the targeting of your products and services is first of all driven by your website's content. A customer will go to an auto site because she wants information on that new sport utility model, or to a sport site because she wants details on that game last night. This is the first level. The second level of the continuum involves some server and browser specifics, and at this point the data is contained in log files, such as access time and day, referring engine, sub-domain and HTMLs requested and viewed. At the third level, the data involves registration and demographic information gleaned from ZIP codes. At this point

Amazon/AlmondNet
Exhibit 1011
Page 88

you are dealing at a market segment level, which can be mined by such data mining tools as a decision trees generator. At the fourth and final level, the data involves individual customer behavior, including page views, clickthroughs, queries, ad impressions, and transactions. At this narrow one-to-one level a neural network tool is required for scoring individuals by propensity to purchase specific products or services.

Web Marketing

Because the Web can track visitors at the granular level, ads can be targeted to those potential customers most likely to respond based on confidential user profiles and their mined behavior. Data mining typically involves a process know as "supervised learning" in which a pattern or a signature is discovered for a particular group of consumers. These are customers with specific tendencies and attributes, which when identified can be used to target new potential prospects with similar profiles. The benefit of relational marketing via data mining is that it gives your visitors and customers the opportunity to teach you what they truly want and desire.

Before you can mine your website data you must plan and design the processes by which you can capture the attributes of your visitors and customers. For starters, your website data has the capacity of targeting ads, products, and services at several levels, the first of which begins with your server log files:

- geographic location
- prior purchases
- browser type
- time of day
- past clicks
- IP address

If you are using forms for purchases, registration, games, guest books, contests, or other types of promotions, you have a second level by which you can begin to capture actual visitor and customer profile information, such as:

- occupation
- profession
- ZIP code
- gender
- age

Amazon/AlmondNet
Exhibit 1011
Page 89

At the third level you have the actual behavior of your customers, whose actions you can mine in order to identify specific patterns or signatures. At this level, an additional data component from cookie files can begin to be assembled and mined in order to discover propensities for clickthroughs and actual online sales. Data from the prior two other levels can also be merged in order to construct models and profiles identifying the attributes of these customers.

When this information is assembled, data mining techniques can be used to infer customer composites about lifestyle similarities, attributes, and characteristics. With these insights you can begin to fine tune your advertising and marketing campaigns. Through the careful design and integration of your forms you can begin to learn from your visitors and customers what they like and don't like about your products and services and those of your competitors. Through this process you can increase your revenue stream from each customer and begin to gain a competitive advantage. In the process of retaining customers, the cost of attracting new ones is reduced, since it costs less to keep them than to acquire new ones. Tracking, capturing, and mining data from advertising and your website promotional efforts are important, especially if you are gathering useful demographic data relevant to your business objective: *online sales.*

The more targeted your ads and online promotions are, the more qualified the leads you're likely to get and the higher a response you are likely to see. Website visitors and customers behave differently and have unique needs for using the Internet, which means you should be aware of their idiosyncrasies. With the data mining of your website you will gradually "learn" to deal with your customers in subtle but effective ways—with different cues, offers, and content—targeted much narrower than traditional media.

Your website has the ability to provide individualized products and services when it captures enough data for mining from assorted log and cookie files and personal information captured from your site's forms. There are additional technologies, techniques, and services unique to the Web which you can use to generate the data for mining, including products and services from BroadVision's Relational, ADSmart, Engage Technologies, and GuestTrack. However, before this can succeed, security and privacy concerns need to be addressed. You need to respond to concerns over silent tracking techniques via cookies, collaborative agents, ad servers, etc., and explicitly inform your visitors and customers how you plan to use the data you are capturing to personalize their experience while they are in your website. Your primary reason for data mining should be to create such a high

Amazon/AlmondNet
Exhibit 1011
Page 90

level of customer service that your clients will not want to defect to your competitors.

Online Coupons

Online coupons are a very cost-effective vehicle for conducting market research over the Web and collecting very accurate consumer product information. To collect online coupons, shoppers choose savings offers from a special selection in a retailer's website. After clicking on their personal choices, shoppers receive securely printed, copy-resistant coupons through the mail, redeemable at any of the retailer's physical locations.

Blending traditional market research and web ad tracking, electronic and physical retailers are beginning to use online coupons to efficiently and cost effectively learn more about their customers. (See Figure 2–13.) Visitors to retailer websites can complete forms in order to obtain coupons, and as they do so they are asked about their gender, the number of people in their household, whether they have a pet, and so forth, in order for the coupons to be tailor-made to their consuming specifications. Online coupons have a high response rate—in some cases, over 20 percent on a variety of national branded

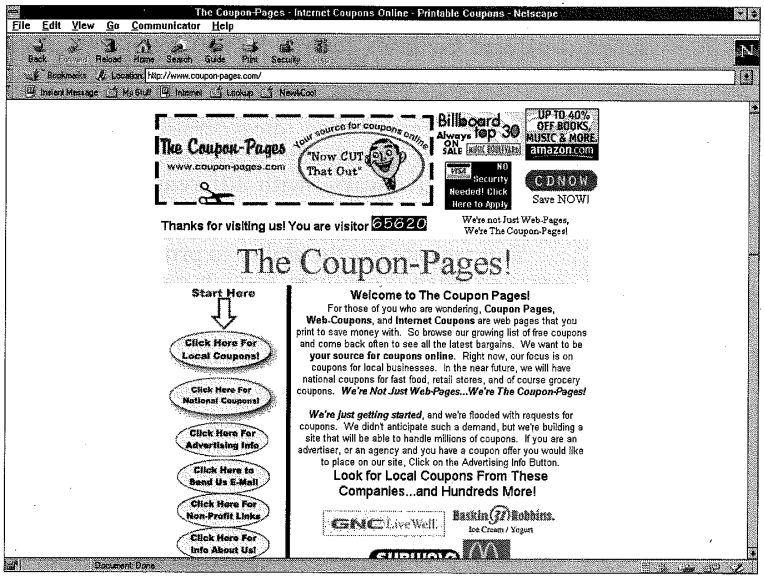


Figure 2–13 *Online coupons can gather important consumer information.*

products from leading consumer-packaged goods manufacturers. Traditional coupons average less than 2 percent redemption.

Mining Customer Value

Most of today's data mining tools are based in part on statistical algorithms such as CHAID (CHI-squared Automatic Interaction Detection) or CART (Classification and Regression Trees) and in part on several artificial intelligence technologies designed to emulate human perception, such as neural networks and machine-learning. Some of these data mining tools search for patterns in data and then organize their solutions as conjunctions or business rules. Data mining analyzers can deliver their solutions in useable business statements that are easy to grasp and implement by a web or marketing teams in an electronic commerce environment:

IF	Total Number of Visits Made =	7 through 9
AND	Category of Book Purchased =	Database Management
THEN	Total Projected Sales for Year =	$90

Unlike website tracking software, data mining technologies are designed to sort through hundreds and even thousands of data fields to find the few significant ones for profiling online customers by potential profit groups or market clusters. The core technology of most of today's data mining tools can split and evaluate hundreds of thousands of records and data fields in a web-generated database in order to arrive at those attributes that are influencing important business factors such as total sales.

The underlying mathematics of most data mining tools is manifested in how they compute the amount of "information gain" in a record relating to its class membership. *Gain criterion* is actually a measurement of information, or the amount of information conveyed in a message. Data mining tools (at least those incorporating symbolic classifier algorithms) actually compute this measurement as they process the records in a database in their search for patterns; they "interrogate" the database via a sequential process of small decisions, with each decision conceptually a yes/no classification based on properties of each customer record. The following is an example of how a data mining analysis may partition a single transaction in a website doing business-to-business retailing:

Amazon/AlmondNet
Exhibit 1011
Page 92

	Low LTV	High LTV
ZIP Code	94501	94502
Customer Since	1994	1996
Catalog Code	V78	B49
Last Purchase	96/04	98/04
Revolving Limit	$7,500	$51,000
Credit/Debit Ratio	45%	34%
Sub-domain	ORG	COM
# of Employees	32	12
SIC	896	456

Here, the transactions are split between Low and High LTV (lifetime value, i.e., the duration of their relationship with a site and the amount of money they spend during that time) on the basis of the data, rather than a hypothesis, hunch, or intuition of the web or marketing analyst, thus automating the process of stratification. Most data mining tools perform somewhat the same process on a database: they split it into classes that differ as much as possible in their relation to a selected output, such as total sales. In other words, a web database is split into subsets according to the results of statistical tests conducted on an output by the data mining algorithm and not a statistician, web, or marketing analyst.

One way to see this process is as a game of "20 Questions," where one player asks questions to determine the kind of object another player has in mind. Presented with a large web-generated customer database with varying "total sales," a data mining tool asks a series of questions about the LTV of each record, with a goal of classifying each sample into a class or group. LTV classes are thus determined by the properties of records in a database and generalized into predictive models from specific examples in that data set.

Another way of viewing data mining technology is as a simplifier. It enables the "pruning of noise" so that the web marketer or electronic retailer can identify the specific customer attributes, such as *establishment date* or *sub-domain*, which have the most effect on predicting online customer LTV. In addition to facilitating and improving web marketing efforts for its most profitable online customers, LTV analysis can indicate which Low LTV customers might warrant additional attention, enabling you to convert them to High LTV, or identifying those that do not warrant additional marketing since they are not profitable.

Amazon/AlmondNet
Exhibit 1011
Page 93

Data mining analysis can also assist marketing in the search for new online prospects, which "fit" the High LTV profile of your current customers. The following website database reflects an average customer sale amount of $200, but a data mining analysis uncovered three clusters of customers with varying LTVs and some insightful features. The three groups of LTV customers reflect total sales ranging from a low of $70 to a high of $900.

The Occasional Web Customer: Low LTV			
IF	Sub-domain	=	AOL
AND	Connection Time	=	4 Through 6 Minutes
AND	Gender	=	Female
THEN	Total Sales	=	$70

The Average Web Customer: Medium LTV			
IF	Transactions	=	3–4
AND	Age	=	40–49
AND	Gender	=	Male
THEN	Total Sale	=	$200

The Premier Web Customer: High LTV			
IF	Age	=	25–39
AND	Sub-domain	=	NET
AND	Gender	=	Female
THEN	Total Sales	=	$900

With this analysis an electronic retailer may concentrate its banners, ads, and other marketing campaigns on retaining all *female* customers between *25* and *39* with sub-domains *.net*, since they represent the highest LTV for this company. At the same time, this website may want to construct a marketing campaign to introduce the Occasional Customers to the additional products they may not be aware of through banners designed to "educate" them, with an effort toward increasing sales. Through this process of segmentation a retailer can begin to construct marketing campaigns on the basis of what each

Amazon/AlmondNet
Exhibit 1011
Page 94

market segment represents to their business. What's important in this example is that the data mining process discovered the unique features and intervals in the data signaling a unique "signature" of High LTV web customers vs. Low and Medium LTV buyers.

A key distinction between manual methods of web data analysis and data mining is that the process used to discover the key customer *attributes*, such as sub-domain and gender, or the *intervals*, the number of transactions 3 to 4, or age ranges 25 to 39 in the data set. The key difference in using data mining tools in the analysis of web data is that the algorithms and not the user discovered the customer attributes and value ranges. The data mining process can also discover the key intervals indicating when a web customer may be expected to go from a Low LTV to a High LTV. The iterative analysis of your web data can also point out such customer profiles as a "brand-switcher" or a "win-back" for a specific product or service, based on historical online trends and web transaction patterns.

Once a LTV profile has been identified by the mining of web-based data, additional enhancements can be made through "appends" from demographic and household external databases. Internal web customer databases can be improved through a "merge and purge" methodology, that is, by merging an internal web customer database with external lifestyle and demographic information. Through this process new features about online customers may be discovered. Next, through the use of data mining, irrelevant attributes can be purged from the database for determining LTV. For example, a website marketer may discover that High LTV web customers have unique attributes regarding specific income levels, or other lifestyle traits. Once the data mining tools have "learned" the dominant features of each LTV group, it can rank them in order of statistical importance:

Name	Ranges	Rank
Zip Code	1211 to 98776	13
# of Employees	03 to 89	49
SIC	000 to 999	03
Customer Since	1996 to 1998	04
Revolving Limit	0 to 90000	01
Credit/Debit Ratio	0 to 100	06

Amazon/AlmondNet
Exhibit 1011
Page 95

Using this type of rank table, a web marketer can purge those customer attributes in a database that have little or no use in identifying High LTV clients. For example, in the above data mining analysis of a business-to-business web retailer, *Revolving Limit* is the most important attribute in the data, while *Number of Employees* was of no importance and may be excluded from subsequent analysis. The merge-purge process can be viewed as a method of "condensing" a very large database, enabling the electronic retailer and marketer to focus on only those factors relevant to the grouping of high revenue producing online customers.

Amazon/AlmondNet
Exhibit 1011
Page 96

3

Networks and Algorithms

They're about Learning and Evolving

In the preceding chapters we discussed some of the electronic retailing applications of data mining and how the technology differs from traditional methods of data analysis. Now we will discuss in more detail how each of the technologies operates on a data set and what insight they can provide about your website. How do neural networks learn from data? How do machine-learning and statistical algorithms segment a database? How do genetic algorithms optimize the entire learning process into a method of evolutionary computing? In short, what are the underlying pattern recognition processes of today's data mining tools and toolboxes?

In order to understand and fully benefit from the process of pattern recognition, we need to know how this technology came about. We will take a short historical trek through several branches of artificial intelligence and some key events leading to the developments and acceptance of neural networks, machine-learning, and genetic algorithms, with maybe a couple of short excursions into the area of fuzzy logic and an introduction to the statistical algorithms CART (Classification and Regression Trees) and CHAID (CHI-squared Automatic Interaction Detection). These technologies are at the core of most of the current data mining tools, so knowing something about them will help you know how to use them to your advantage and in conjunction with each other to optimize your mining efforts.

Although we will not go into the gory details or math of backpropagation and Kohonen networks or the Interactive Dichotomizer and C5.0 algorithms, we will instead concentrate on explaining these concepts and the difference between what each network and algorithm does, specifically how each of them operates on the data and what you can expect from each in terms of business solutions. The scope of this chapter is to assist you in knowing how to leverage these

Amazon/AlmondNet
Exhibit 1011
Page 98

technologies to improve your marketing presence and effectiveness on the Web, that is, to use the networks and algorithms effectively to improve your online bottom line.

Data mining is rooted in artificial intelligence. However, AI is a diverse field involving everything from natural language to expert systems. Most of the current data mining tools are based specifically on neural networks and genetic- and machine-learning algorithms—three branches of AI that continue to evolve today, but that experienced some important breakthroughs during the last decade. As such, they are relatively mature technologies that have migrated to commercial software in the form of both data mining tools for the desktop and high-end multiple-paradigm toolboxes for parallel processing servers.

AI has traditionally sought to emulate human processes. Its applications have basically involved attempts to get machines to do what humans do best. From robotics in manufacturing to agents on the Web, both are technologies designed to assist their creators. Likewise, neural networks and machine-learning and genetic algorithms are similar efforts to do through code what humans have been doing for millennia—to learn to recognize patterns. AI seeks to emulate human memory, learning, and evolution. However, not too long ago, the techniques for how to best do this were a matter of controversy and conflict. Two major forces of AI battled with each other not only in their approach on how to best replicate human functions, but also over government funding and corporate attention.

The AI War

Basically, there are two main schools of thought on how machines should learn: *inductive* and *deductive* analysis. (See Figure 3–1.) The deductive approach involves the construction of rules obtained from domain experts through interviews by knowledge engineers. These knowledge engineers go on to construct "expert system" programs containing sets of rules designed to "fire" when certain conditions are encountered by users. The inductive approach, on the other hand, involves the generation of rules directly from the data, rather than from domain experts.

The Top-Down Approach

The deductive approach was the dominant AI branch, receiving the most hype and attention during the 1970s and 1980s. Ten years ago when AI was the "hot" new technology expert systems received the

Amazon/AlmondNet
Exhibit 1011
Page 99

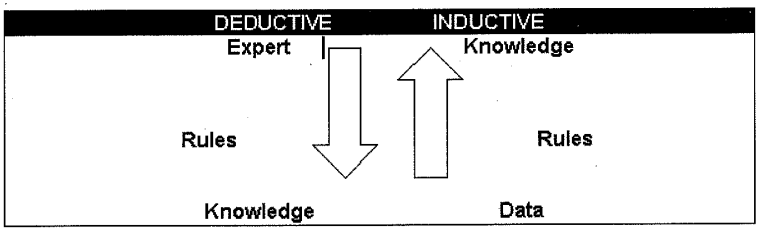


Figure 3–1 *The roots of data mining are based on the inductive approach to knowledge discovery.*

majority of support from both the government and venture funding. Expert systems used special languages, such as LISP and Prolog, and in some instances special workstations. For example, at one time American Express used an expert system to assist them in approving credit, while the Internal Revenue Service tried to use an expert system in the selection of tax returns for audits. A soup company attempted to construct an expert system by codifying the processes of its master cook.

The success of these expert systems, however, was short lived, as the programs suffered from significant shortcomings: they were brittle and expensive to maintain. Some expert systems required hundreds of rules that had to be manually maintained and updated to stay in tune to a rapidly changing business environment. Expert systems eventually proved to be too fragile and rigid to keep pace with the rapidly evolving conditions of everyday transactions. However, it was during this period that two researchers from the Massachusetts Institute of Technology, Seymour Papert and Marvin Minsky, vying for funding for expert system technology, wrote about the deficiencies of neural networks and thus stymied the development of the inductive approach to AI for several years.

The Bottom-Up Approach

The alternate approach to AI is inductive data analysis, which extracts patterns from data and learns from examples. The inductive approach to AI involves generating the rules for a system from thousands of cases, in contrast to extracting the rules from experts via interviews. It was an approach to rule generation that did not take off until the 1980s with the work of several researchers from diverse fields of mathematics, statistics, and computational science. John Hopfield advanced the concept of a feedback neural network. J. Ross Quinlan, a mathematician, developed the Interative Dichotomiser ID3 machine-learning algorithm, and John Holland discovered genetic algorithms.

Amazon/AlmondNet
Exhibit 1011
Page 100

Working independent of one another, these three researchers have had a vast impact on data mining, for it is from their work that almost all of the current data mining tools are derived—decision trees, neural networks, rule generators, evolutionary programs, etc. Through the process of a feedback network, Hopfield demonstrated how a neural network could adjust and eventually learn a pattern from repeated exposure to case samples. Quinland's work on ID3 was one of the first learning systems capable of generating rules in the form of a decision tree. Holland's work at the University of Michigan led to the theory of schema, which provided the groundwork for genetic algorithms and demonstrated how programs can work toward the optimization of solutions. Holland and his colleagues mapped the future of evolutionary computing, which will have a long-term impact on the future of data mining and electronic retailing on the Web.

Neural Networks

Neural networks are so named due to their overall design, which is based on the structure of the brain: both have many highly interconnected neurons. Like the brain, a neural network is a massively parallel, dynamic system of highly interconnected interacting parts. It is able to construct nonlinear predictive models that learn through training and resemble biological neural units in structure. The ability to learn is one of the features of neural networks; they are not programmed as much as trained.

At the core of the network are the hidden connections that contain layers of "weights," which are the formulae to pattern recognition. That is, they work even when given incomplete information. Most neural networks have some sort of "training" rule whereby the weights of connections are adjusted on the basis of presented patterns, and they "learn" from examples. A neural network consists of layers of neurons, which are connected to each other. Some of the neurons will be used to input data into the network and send it to the inner layers of neurons. The output neurons provide the network's response or output. Neurons are connected and combined to generate an output, as Figure 3–2 shows.

In other words, a neural network is a system of software programs and data structures that approximates the operation of the brain. They usually involve a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory. Typically, a neural network is initially "trained" or fed large amounts of data and rules about data relationships (e.g., "AOL sub-domains purchase printers but not scanners"). Neural net-

Amazon/AlmondNet
Exhibit 1011
Page 101

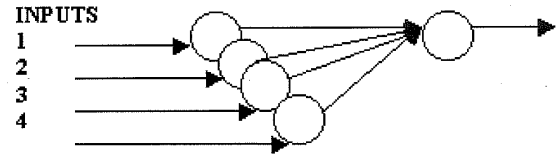


Figure 3–2
This one output network is the equivalent of logistic regression.

works are basically computing memories where the operations are association and similarity. They can learn when sets of events go together, such as when one product is sold, another is likely to sell as well, based on patterns they have observed over time.

In a neural network, after a pattern is learned, such as when a website visitor selects a series of pages, the network examines the possibilities and makes an immediate match, such as a specific ad, offer, price, or link. This is the type of problem for which a web marketer may make a decision based on personal observations. However, what if the problem is not for just one page, but hundreds, and what if the options are not just two or three but 20 or 30? This is where a neural network can be trained to recognize subtle and small differences. It can learn a pattern of interaction and activity over that website and split the information up across its many neurons. It is trained using hundreds or thousands of sessions so that very quickly it learns what path which type of visitors should be offered.

From Slugs to VISA

In 1982, John Hopfield, of Caltech, presented to the National Academy of Sciences the first paper on neural computing presented to any distinguished body since the 1960s. The paper described Hopfield's neural computing system, called the "Hopfield Model" or "cross-bar associative network." Using research based on the olfactory system of a garden slug, Hopfield presented a neural computing system consisting of interconnected processing elements that seek an energy minimum. Hopfield's model represented neuron operation as a thresholding operation and illustrated memory as information stored in the interconnections between neuron units. He also illustrated and modeled the brain's ability to call up responses from many locations in response to a stimulus; he thus modeled how a neural system "associates" information from multiple storage sites for a given input.

Hopfield's enthusiasm, reputation, and clarity of presentation was well received and prompted researchers to become interested again in the fascinating field of neural computing. The increasing attention in the field

Amazon/AlmondNet
Exhibit 1011
Page 102

also opened many sources of funding necessary for the pursuit of neural computing research and the inductive approach to AI. His work led to an explosion in neural networks papers, conferences, and software companies, which continues to this day. Today, neural networks are successful in applications requiring prediction, data classification, and pattern matching. Examples of successful applications include mortgage risk evaluation, production control, handwriting recognition, and credit card fraud. The credit card you carry is being monitored every hour of every day by a neural network for potential patterns of irregularity; the system being used is most likely from HNC, a neural network company from San Diego, which monitors over 150 million credit cards.

Back-Propagation Networks

Neural networks were not seriously considered until Paul Werbo's groundbreaking 1974 Harvard doctorate thesis, *Beyond Regression*, which laid the statistical foundation to the work on back-propagation by several researchers from varied fields. In 1985 the back-propagation neural network architecture was simultaneously discovered by three groups of researchers: D. E. Rumelhart, G. E. Hinton, and R. J. Williams; Y. Le Cun; and D. Parker.

A back-propagation neural network consists of several input nodes, some hidden nodes, and one or more output nodes. Data cycles through the nodes as the net trains and adjusts (Figure 3–3). Back propagation is a learning method where an error signal from an output node is fed back through the network altering weights as it goes, to prevent the same error from happening again. The process involves, as with the other types of networks, a gradual process of training with incremental improvement over time, until a network learns a pattern. This pattern can be anything, including the behavior of website visitors and customers. The back propagation architecture is the most dominant of all neural networks, and it is the most popular in the current crop of data mining tools involved in supervised learning.

To demonstrate how a back-propagation network works we can try it by solving the exclusive-or (XOR) function whose output is logically true (or has the value "1") when the two inputs are opposite (Figure 3–4). The numbers between the neurons are the strengths of the connections between them, or the weights. The weight values will be modified as the network is trained. For training, the network is started with a random set of weights (Figure 3–5).

We start with the assumption of the threshold for the neurons to be .01. This means that if the sum of the inputs is greater than .01,

Amazon/AlmondNet
Exhibit 1011
Page 103

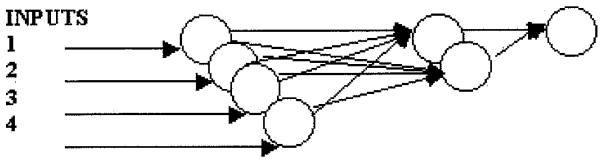


Figure 3–3 *This network with a hidden layer allows it to recognize patterns.*

Figure 3–4 *The Simple XOR Function.*

LOGIC TRUTH TABLE

INPUTS		OUTPUTS
1	1	0
0	0	0
1	0	1

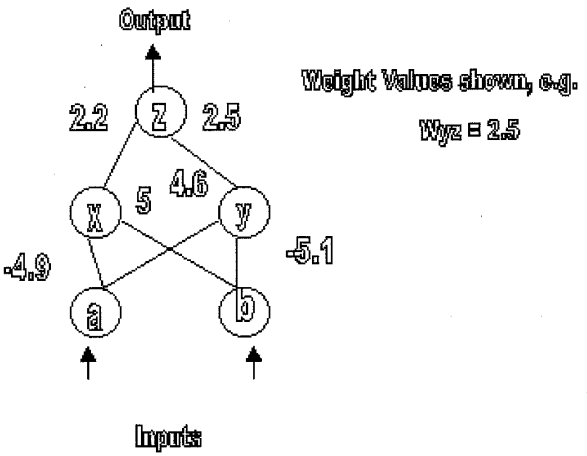


Figure 3–5 *XOR function using linear threshold units in a back-propagation network.*

the neuron will output a 0, and if the sum of the inputs is less than .01, then the output will be 1. After applying a training pattern of 1 on each of the two inputs, the neurons' outputs are computed and adjusted, as Figure 3–6 shows.

Each of the neuron's output is noted for each node along with the sum of the neuron's internal signals. Neurons *a* and *b* are both sending 1's to the next layer. Neuron *x* gets two signals, one from neuron *a* and one from neuron *b*. The weight from neurons *a* to neuron *x* is W*xa*, and its value is –4.9. We multiply the signal times its weight and get –4.9. The input to neuron *x* from neuron *b* has a weight W*xb* that is

Amazon/AlmondNet
Exhibit 1011
Page 104

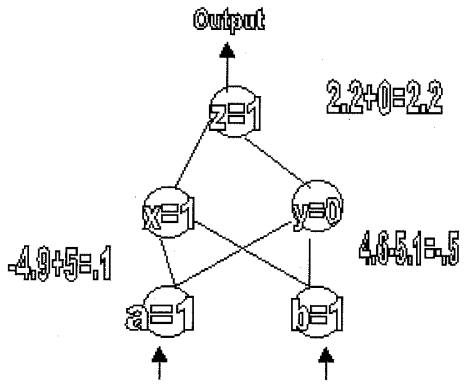


Figure 3–6
XOR function back-propagation network input calculations.

equal to 5. The signal from b is 1. This contributes 5 to neuron x. We next add –4.9 and 5 to get a total internal signal of .1 at neuron x. This is greater than the threshold of .01 so neuron x outputs a 1. In the same way the outputs of y and z can be computed. With the training pattern of the network, the final output should have been a 0, but it was a 1. The network determines an error factor by adjusting the weights from the last layer to the first layer, one layer at a time. The error at neuron z, Ez, is the desired output minus the actual output Ez = 0 –1 = –1.

The weight adjustment is computed by multiplying the error at the neuron times the output of the neuron that the weight is associated with, so that Wzx = –1 + 2.2 = 1.2. This is added to the current value of Wzx with the new Wzx = –1 + 2.2 = 1.2. This new weight is now the error at z times the signal from y, added to the current value of Wzy. The new Wzy is = (–1 * 0) = 2.5 = 2.5. Wzy is unchanged. We next compute the error at the neurons in the next layer back. For neuron x, it is equal to the error at z times the weight of the connection from x to z: Ex = –1 * 2.2 = –2.2. To compute the new Wxv, we multiply the error at x by the signal from a, then add the current Wxa, so that the new Wxa = (–2.2 * 1) – 4.9 = –7.1. After all the weights are adjusted, a new input pattern is applied and if there is an error, the sequence of back-error propagation is repeated. When all four pairs of input patterns produce the correct outputs, the process stops.

Supervised Learning

This is basically how most neural networks learn: by example in a supervised mode. Supervised models, like back-propagation networks, are trained with pairs of examples: positive and negative. A given

Amazon/AlmondNet
Exhibit 1011
Page 105

input pattern is matched with a desired output pattern. Training a supervised network is a process of providing a set of inputs and outputs, one at a time. The network trains by taking in each input pattern, producing an output pattern, and then comparing its output to the desired output. If the network output is different from the desired output, the network adjusts its internal connection strengths (weights) in order to reduce the difference between the actual output and the desired output. If, however, its output matches the desired output, then the network has learned the pattern and no correction is made.

This process continues until the network gets the patterns of inputs/outputs correct or until an acceptable error rate is attained. However, because the network may get one set of patterns correct and another wrong, the adjustments that it makes are continuous. For this reason, training of a network is an iterative process where input/output patterns are presented over and over until the network "gets" the patterns correct.

In your website, the data (inputs or target patterns) may be in the form of continuous value variables or a set or ranges. For example, the connection time of online sessions could be represented by the actual number of minutes (a continuous valued variable) or one of four states (ranges of time): –30 seconds, –90 seconds, –300 seconds, and +300 seconds. In the latter case, four neurons, each with bipolar values, would be appropriate; in the former case a single neuron would be used.

In general, it is easier for a neural network to learn a set of distinct responses than a continuous valued response. However, breaking up a continuous valued variable into artificial distinct categories can make it more difficult for the data mining tool to learn examples that occur on, or near, the boundaries of the groups. For this reason, continuous valued input variables should be carefully used to represent distinct quantities such as connected time from log files, or visitors' age from registration forms.

Training a Network

In most instances you will be using a back-propagation neural network involving supervised learning. This means that the training set you use consists of records whose prediction or classification values are already known. It is imperative that you train your network with good, current data. Selecting the data for the training of a neural network is probably the most important step in data mining you will take. Selecting the wrong type of data will doom your efforts no matter how good or expensive a data mining tool you use.

Amazon/AlmondNet
Exhibit 1011
Page 106

The most important factor in the selection of your data is to ensure you include all the ranges of values for all the variables that the network is subject to encounter. In your website, this means including the least and most expensive items and the lowest and highest amounts of sales, session times, units sold, etc. As a rule, you should have several examples in the training set for each value of a categorical attribute and for a range of value for ordered discrete and continuous valued features.

A trained network has the ability to generalize on unseen data, that is, the ability to correctly assess samples that were not on the training data set. Once you train a network the next step is to test it. Expose the network to samples it has not seen before and observe the network's output. A common methodology is to split your data, training on a portion of your data and testing on the rest. One way to determine the patterns your network is finding difficult to learn is to observe closely the errors it is making. To correct them you can retrain your neural network by giving it more samples of the errors it is having difficulty learning.

Preparing Your Data

Real-world data, including that from your website, may be spotty, dirty, or skewed. That is, it may be distributed in an uneven way, making it difficult for a neural network to converge into a good model. A skewed distribution can prevent a network from effectively using an important attribute in your data.

A common way to deal with this problem is to "discrete" an attribute. Rather than having a single input for dollar sale amount, you might break it down to several ranges. Here is an example. Let's say your website sells software products that range in price from very low to very high:

$10–$19	very low	=	00001	=> Input 1
$20–$50	low	=	00010	=> Input 2
$51–$100	average	=	00100	=> Input 3
$101–$199	high	=	01000	=> Input 4
$200–$999	very high	=	10000	=> Input 5

Most of today's modern data mining tools are able to shift the data into these discrete ranges, but you will need to do some of these transformations yourself should the tools you use fail to have this

Amazon/AlmondNet
Exhibit 1011
Page 107

Figure 3–7 *Common features captured and transformed during a web session.*

Attributes	Range of Values	Value	Conversion
Time(seconds)	1-1254	210	.4646
Age	5-57	9	.3333
Gender	M-F	F	.0000
Sub-Domain Type	com net edu	com	.1000
ZIP Code	01254-95432	94125	.0010
Registered User	Y-N	Y	.1000
Referred From	Engine-Other	Engine	.1000
Sale Price(dollars)	2-567	120	.5263

functionality. Another technique for dealing with skewed and categorical data is to use their logarithmic or square root value, as well as doing "1-of-N" conversions, as the above sample demonstrates.

During the initial phase of using a neural network tool, you will need to train it using examples of website sessions in which actual sales took place. The samples need to reflect as many features as you can possibly capture, such as when the sale was made and who made it (Figure 3–7). However, since a neural network works best when all inputs and outputs are between 0 and 1, certain conversions need to be made from the web-generated data. Both continuous and categorical values need to be converted—a process which good data mining tools are able to do relatively easy.

Webbing Your Features

Date and time fields can be inputted into your website's neural network in one of two ways. Any date can be represented as is and left as a single number for a specific day in a month. A variable such as a length of connection in seconds or minutes can also be represented as a fixed point in time. Because you are looking at patterns of transactions in your website, you may want to do some conversions of dates, in order to detect trends in ads, marketing campaigns, seasonal or network cycles, etc. For example, you may want to extract a date and deal with it as an additional input into your network:

Access Date	Day of the Week	Additional Input
030299	Monday	00001
030399	Tuesday	00010
030499	Wednesday	00100
030599	Thursday	01000
030699	Friday	10000

Amazon/AlmondNet
Exhibit 1011
Page 108

Address fields on the Web present another problem. For example, the ZIP code 94502 is not greater than 94501 by one, or larger by 93269 than the ZIP 01233. It is instead another *class*. A ZIP code is a category, not a continuous value number. One of the best ways of dealing with this type of categorical variable is to use another tool or network for clustering. The time needed to train a neural network is directly related to the number of input features used by the network, which means the more features you input, the longer it takes for a net to converge.

A good way of dealing with these types of inputs is to use a machine-learning algorithm tool to create clusters of similar ZIP codes, which can then be transformed into 1-of-N values for inputting into your neural network. Using another data mining tool to prioritize the inputs into a neural network is a good way of determining which features are the most important for training and constructing your model. Another technique for "data compression" is to do statistical correlation, such as Pearson's. Some of the more advanced neural network tools include features for pruning variables, prior to training on data sets.

Purchasing Propensity

The output of a neural network is in the format of a continuous value. As such it can be difficult to interpret for categorical results, such as "will buy vs. will not buy." However, there is a method by which you can take a normal neural network output and create categorical results, which involves creating multiple outputs, as Figure 3–8 shows.

In these situations, you may want to have multiple outputs with each category being a propensity to purchase unique products or services. You may opt to position certain ads or offers on the basis of how a neural network score corresponds with the maximum possible value, or the top three values, or if it exceeds a certain threshold value.

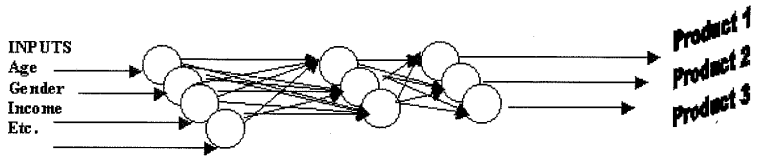


Figure 3–8
A multiple output network for predicting propensity to purchase different products.

Amazon/AlmondNet
Exhibit 1011
Page 109

Training Cycle

Your main task is ensuring you are capturing the relevant and important features from your website, which can be used for prediction and classification. As such, your objectives when using a neural network tool are to ensure you:

1. Identify the input variables—this is *very* important!
2. Convert the variables into usable ranges—pick a tool which will do this.
3. Decide on the format of the output—continuous or categorical?
4. Decide on the training data set sample and a training schedule.
5. Test the model and apply it to your website objectives.

Setting up a schedule for the training of a neural network is important because your model will deteriorate gradually over time. One of the downfalls of any model, especially one designed to predict web behavior, is that it will eventually go stale. The model you create with a neural network will age and atrophy as your website traffic evolves beyond the data used to train your net. As you introduce new products and services, or as you make changes and improvements to your website, changes occur in your traffic and transaction patterns. As such you will need to train your network with fresh data.

Self-Organizing Maps (SOM)

Neural networks perform best in situations whose outputs are nonlinear and with complex problems that do not require a solution of absolute accuracy. They also are ideal in situations where a large amount of historical data is available for training and where a "black box" solution will do and no explanation is required. Most of today's data mining tools incorporate two main types of neural networks, back-propagation and SOMs, also known as Kohonen networks, after its inventor.

In 1981 Tuevo Kohonen proposed and demonstrated a completely different network architecture. An SOM network resembles a collection of neurons, as in the human brain, each connected to its neighbor. As an SOM is exposed to samples, it begins to create self-organizing clusters, like cellophane spreading itself over chunks of data. Gradually, a Kohonen network will organize clusters and can thus be used for situations where no output or dependent variable is known. SOMs have been used to discover associations for such purposes as market basket analysis in retailing.

Amazon/AlmondNet
Exhibit 1011
Page 110

The key difference between back-propagation nets and Kohonen nets is that the former is used for supervised learning where a training sample is needed to construct a classification model. For example, with a back-propagation neural network, training is usually done with a balanced sample of website visitors who made a purchase versus those who did not. Also unlike the back-propagation networks which usually have a single output, an SOM network can have multiple outputs competing with each other. SOMs involve unsupervised learning, a training sample for which no output is known, and is commonly used to discover relations in a data set. This by the way does not preclude using both paradigms—first the SOM to discover a sub-class within the data, then a back-propagation network for classification of that class. The same can be said for using neural networks in combination with genetic and machine-learning algorithms.

SOMs are probably second only to back-propagation architecture in terms of the number of applications and tools for which it is being used. The most significant difference between the back-propagation and the self-organization models is the fact that the SOM is trained without supervision. The SOM network is more biologically oriented than the back-propagation models in that it bears more resemblance to the way humans learn and the brain is organized. If you don't really know what you are attempting to classify (such as the website patterns you are looking for), or if you feel there may be more than one way to categorize your visitors or customers, you may want to start with an SOM.

Calibrating Clusters

Using an SOM to discover clusters in your data can often lead to a problem in trying to extrapolate a meaning for each of the groupings. There are two ways of dealing with this problem. The first involves obtaining the average value for each of the features in the data. For example, if the number of children or customer age is one of the input features, then what you need to obtain is the average value for each of these features in each of the clusters the SOM discovered. The other method involves passing each of the clusters through a machine-learning algorithm in order to generate a set of rules describing the features of that grouping. This second option requires the ability to link the output from an SOM to the inputs of a machine-learning algorithm tool like C4.5, CHAID, or CART, which is possible with some of the new data mining suites, or toolboxes like Clementine.

Amazon/AlmondNet
Exhibit 1011
Page 111

Keep in mind that neural networks are basically memories: they do not compute solutions as much as remember answers. As such, the data samples that are used to train a network are critical in determining the accuracy of the network model. If the memory is organized in the proper way, then the look-up table of that network and its core weights, which is really what networks are all about, will generalize in a manner which will be very accurate. Despite the rap networks take for being black-box wonders from which no explanation to their solutions is forthcoming (which, ironically, may be said for regression), they are very good tools. Neural networks are good data mining tools for classification and prediction and work best when used in conjunction with machine-learning and genetic algorithms.

Neural Network Tools

Most of the work involved in the use of neural networks is in the preparation of the data. Luckily, most of today's data mining tools are sophisticated enough to handle the dirty work involved in such tasks as balancing the data. (*Balancing* is using an equal sample of positive and negative cases (Buyers vs. Non-Buyers) to train a network.) Another task involves the scaling of the data; converting all inputs to a range from 0 to 1 and back in the output. This may also involve using functions for dealing with skewed data sets, such as taking the logarithm or the square root of their values prior to passing the data through a network. Still another task involves converting categorical inputs into either 1-of-N values or thermometer values:

If Number of Children is:	1-of-N Value	Thermometer Value
0	00001	0.5000
1	00010	0.7500
2	00100	0.8750
3	01000	0.9375
4	10000	1.0000

The following data mining tools incorporate back propagation and, in some instances, SOM networks for classification. Most of these tools are mature, third- or fourth-generation versions of their software, which provide some functions for data preparation and balancing of inputs prior to training and model creation. Care must still

Amazon/AlmondNet
Exhibit 1011
Page 112

be taken in their use, since they do require some adjustments, such as the selection of their topography and the setting of learning rates. The next generation of these tools is also beginning to incorporate genetic algorithms to optimize their settings and improve their results. These groups of tools include: 4Thought, BrainMaker, INSPECT, MATLAB NN Toolbox, ModelQuest, NGO, NeuralWorks Predict, NeuralWorks Professional II/PLUS, Proforma, PRW, SPSS Neural Connection, and Trajecta.

A common sense rule for determining when to use a data mining tool based on a neural network involves the following considerations:

- Are you dealing with a complex problem involving much input and which does not require a true model or a clear-cut solution with absolute accuracy?
- Will being close be good enough?
- Are you trying to solve nonlinear problems, that is to say, is there a clear-cut relation between your inputs and outputs?
- Are you dealing with very large amounts of historical data?
- Do you need a solution quickly, and does that solution require an explanation (will a "black box" model do)?

If in your opinion your problem meets these requirements, then the use of a neural network tool will work for you. However, a neural network solution should be avoided in situations where a problem requires a robust model. A neural network should also be avoided in situations where a clear understanding or an insight into the pattern you are trying to discover is required. Some neural network tools provide sensitivity reports, which enable you to see some of the relationships between the inputs and your desired output.

Machine-Learning Algorithms

Machine-learning algorithms are commonly used to segment a database, with their output being either in the form of a decision tree or a series of IF/THEN rules. They commonly work in supervised learning situations where they attempt to find a test for splitting a database among the most desired categories, such as "website visitor will buy vs. will not buy." In both instances these algorithms seek to duplicate the protocol that a skilled statistician would follow in analyzing a data set: to identify important data clusters of features within a database.

Data mining tools incorporating machine-learning algorithms such as CART (classification and regression trees), CHAID (chi-squared

Amazon/AlmondNet
Exhibit 1011
Page 113

automatic integration detection), ID3 (Interactive Dichotomizer), or C4.5 or C5.0 will segment a data set into statistically significant clusters of classes based on a desired output. As noted, some of these tools generate "decision trees" that provide a graphical breakdown of a data set (a sort of map of significant clusters), while others produce IF/THEN rules, which segment a data set into classes that can point out important *ranges* and *features*. Such a rule has two parts, a condition (IF) and a result (THEN), and is represented as a statement:

If Customer code *is* 03
And Number of purchases made this year *is* 06
And ZIP Code *is* 79905
Then PROSPECT *will purchase Product X*
Rule's probability: .88
The rule exists in 13000 *records.*
Significance level: Error probability < 0.13

A Measure of Information

Most of these types of data mining tools evolve from two lines of research and development. The first is from the area of statistical decision trees and the other in machine learning, such as the ID3. Mathematician J. Ross Quinlan, in his ID3 system, introduced and popularized the concept of today's data mining decision tree tools. There are two main types of decision trees: binary and multiple branches. A binary decision tree splits from a node in two directions with each node representing a yes-or-no question. ID3 is a program that can build trees automatically from given positive and negative instances. Each leaf of a decision tree asserts a positive or negative concept.

To classify a particular input we start at the top and follow assertions down until we reach an answer. For example, in the ID3 algorithm, the sequence of fields that are selected as predictors is determined by a measure of entropy. This information measure is best for the attribute whose value, when known, minimizes the departure of your guess, or prediction, from the actual value of the dependent variable. Other fields are then examined for their ability to further differentiate values in the decedent nodes of the resulting tree. The

Amazon/AlmondNet
Exhibit 1011
Page 114

ID3 algorithm works by computing the entropy associated with each attribute. The calculation is made as follows:

$$H(C) = -\sum_{i=1}^{N} p(c_i)\log_2 p(c_i)$$

Rather than work you through the math, suffice to say that ID3 performs a measurement of entropy or noise in order to measure the amount of data an attribute contains. The overall entropy of the classification is the expected amount of information that will be gained when the class is specified. For example, let's say you are trying to determine which attribute contains the most information in the classification of fruit:

"What is it, a banana, an apple or an orange?"		
Data	*Attribute*	*Information Gain*
It's 7.8 oz.	Weight	Little
It's round.	Shape	Some
It's red.	Color	Most

Figure 3–9 is an example of an ID3 decision tree. Notice that ID3 forms a branch for unique clusters with similar income categories. ID3 is the precursor to C4.5 and C5.0, which were developed also by Quinlan and use a criterion known as *information gain* to compare and generate potential splits within a data set. C5.0, the newest and latest algorithm, uses the ratio of the total information gain due to a proposed split to the information gain attributable solely to the number of subsets created, as the criterion for evaluating proposed splits. The ID3 and C4.5 algorithms are based on concept learning, where the number of branches equals number of categories for predictor outputs.

The Rules

Machine algorithms, which are the core technology of most decision tree and rule-generating data mining tools, take a "divide and conquer" approach. In any given data set, such as your log files or forms database, the algorithms look at the attributes (domain,

Amazon/AlmondNet
Exhibit 1011
Page 115

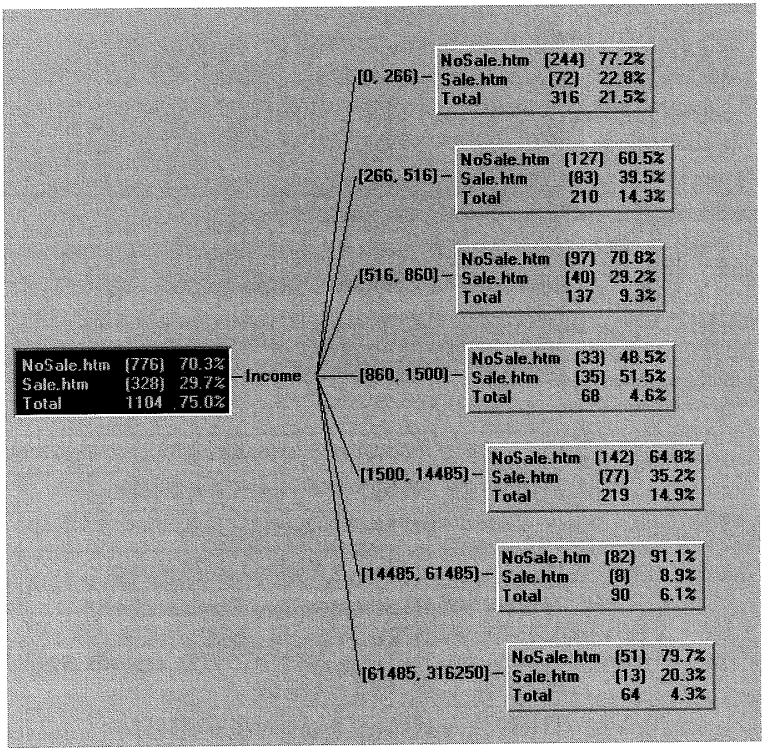


Figure 3–9
A decision tree can segment your website customers based on different attributes, such as their income, age, gender, domains, or location.

referred engine, age, gender) and perform a test at a node of a tree or a rule in order to generate a number of possible branches or IF/THEN statements. If the attribute that is tested at a node is a nominal one, the number of branches in the decision tree are usually the number of possible values of the attribute (for example, for gender, it will be Male, Female, or Unknown). If the attribute is numeric, the nodes in a decision tree usually test whether its value is less than a predetermined constant, giving a two-way split. (Missing values in a data set are treated as an attribute value in their own right. Consideration is given to the fact that a missing value may be of some significance.)

Some data mining tools are able to generate decision rules directly from a database and offer a popular alternative to decision trees. It is also possible to read a set of rules directly off a decision tree, with one rule being generated from each leaf.

Amazon/AlmondNet
Exhibit 1011
Page 116

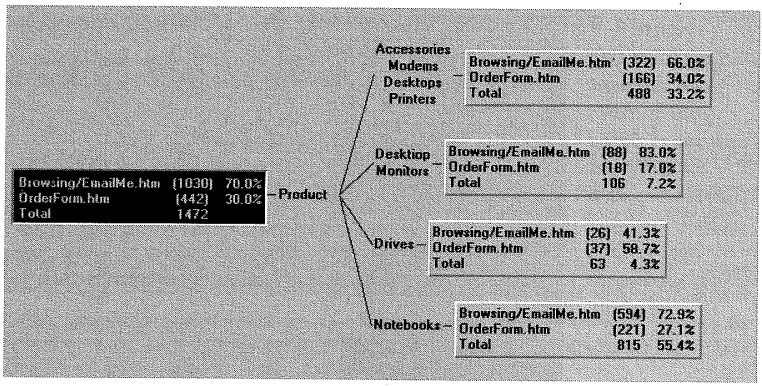


Figure 3–10
Rules can be generated from a decision tree.

IF	Product = Accessories, Modems, Desktops, or Printers	
THEN	HTM = Browsing/E-mailMe.htm	66.0%
	HTM = OrderForm.htm	34.0%
IF	Product = Desktop or Monitors	
THEN	HTM = Browsing/E-mailMe.htm	83.0%
	HTM = OrderForm.htm	17.0%
IF	Product = Drives	
THEN	HTM = Browsing/E-mailMe.htm	41.3%
	HTM = OrderForm.htm	58.7%
IF	Product = Notebooks	
THEN	HTM = Browsing/E-mailMe.htm	72.9%
	HTM = OrderForm.htm	27.1%

Note that these rules are the same as those displayed in Figure 3–10 and can be generated by the software.

The Trees

There are many algorithms for generating decision trees. The CART algorithm generates binary trees (2 splits) it rates high on statistical prediction. It generates two branches from each nonterminal node and uses cross-validation and pruning to determine the size of the tree. Its output or response variable can be quantitative or nominal.

Amazon/AlmondNet
Exhibit 1011
Page 117

CART splits a data set on the basis of diversity, and it looks at all the variables in a database to determine which ones make the best separators or splinters. The best separators are those data fields that do the best job splitting a database into groups where a single class is dominant. Its inputs or independent predictor variables can be nominal or ordinal, with continuous predictors also being supported.

Then there is the AID family of algorithms: AID, THAID, CHAID, MAID, XAID, FIRM, TREEDISC, etc. These algorithms, the most popular being CHAID, are based on the concept of detecting complex statistical relationships. They generate decision trees where the number of branches varies from two to the number of categories of predictor. A key difference between CHAID and CART or C4.5 is that CHAID is highly conservative; it stops growing trees in an effort to avoid overfitting. CHAID also only works on categorical variables. All of these algorithms segment data sets on the basis of statistical significance tests, which they also use to determine the size of their tree. AID, MAID, and XAID were designed for quantitative responses, while THAID, CHAID, and TREEDISC are for nominal responses. Some data mining tools combine two or more algorithms, such as KnowledgSEEKER, which uses CART, CHAID, and ID3, or Clementine, which uses ID3 and C5.0. The algorithm C5.0 is available from Quinland's own company, Rulequest.

The Trees vs. The Rules

Decision trees are popular in many areas of marketing and can also be used to analyze web-based data. The advantages of decision trees are found in their ability to generate understandable business rules in a decision-support environment and the ability to model nonlinear relationships with logical rules. ID3, and its successors C4.5, C5.0, CART, CHAID, and other variations of machine-learning algorithms, perform somewhat the same process on a database: *they split it into classes that differ as much as possible in their relation to a selected output.* In other words, a database is split into subsets according to the results of statistical tests conducted on an output by the *algorithm*—not the user. CART, developed by L. Briemen in 1984, builds binary trees that allow the classification on the basis of the attributes, which are the best splitters and separators. CHAID, which was developed nearly a decade before by J. S. Hartigan in 1975, was designed for purposes of detecting statistical relationships in data and is restricted to categorical variables.

In general, rules are more complex than necessary, and rules derived from trees are usually pruned to remove redundant tests.

Amazon/AlmondNet
Exhibit 1011
Page 118

Rules are not notably more compact than a tree, though, and in fact they are just what you would get by reading rules *from* a tree. In some situations, however, rules are considerably more compact than trees—particularly if it is possible to have a default rule that covers cases not specified by the other rules. One reason why rules are popular is that each rule seems to represent an independent insight into database. New rules can be added to an existing rule set without disturbing ones already there, whereas making an addition to a tree structure reshapes the whole tree. Most rules can often achieve surprisingly high accuracy, perhaps because the structure many real-world data sets exhibit is quite rudimentary, and interactions between the attributes can be safely ignored.

Automatic Segmentation

The process of stratification or classification is automated by machine-learning algorithms on the basis of the data—rather than the hypothesis, or hunch, of the user. Attributes are chosen to split the set, and a tree is build for each subset, until all members of the subsets belong to the same class. Which algorithm is best for you is dependent on the nature of your data mining problem.

Keep in mind what each of the algorithms were designed to do and how they operate on the data. CHAID was designed to detect statistical relationships between variables and is restricted to analysis of categorical variables, such as low, medium, and high. CART is designed to measure the degree of diversity of variables in making its splits—it looks to see which variable is the best splitter or separator in a database (in other words, which website customer attribute is the best "splitter" between your buyers and browsers). ID3 and C4.5 use the concept of "information gain" to make these splitting decisions: which online customer attributes tell you the most about buyers and nonbuyers in your website. The information gained depends on its probability as measured in bits as minus the logarithm to base 2 of that probability. The CART procedures induce strictly binary trees while ID3 partitions by attribute values. CART also uses a statistical resampling technique for both error estimation and cost complexity pruning. Lastly, keep in mind that the output for each of the algorithms differs somewhat—almost all generate some type of decision tree with most also generating IF/THEN rules in varying degrees of numbers. Remember that CART and ID3 can only generate binary trees, whereas CHAID, C4.5, and C5.0 produce multiple branch trees.

Amazon/AlmondNet
Exhibit 1011
Page 119

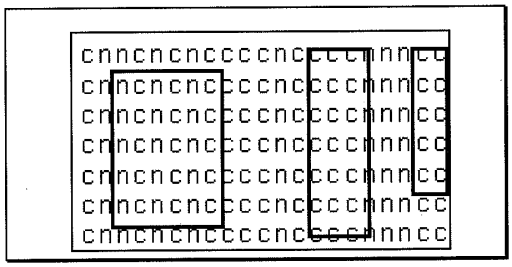


Figure 3–11
C = *Website Customer;* N = *Not Website Customers.*

Machine Learning Your Site

Machine-learning data mining tools are specifically effective in discovering patterns within websites and identifying clusters of profitable customers, as Figure 3–11 shows.

The blocked clusters in Figure 3–11 represent common characteristics within the data and can point to key relationships and patterns hidden in your website data. They also provide to webmasters, marketing, and management easy-to-understand statements about what is occurring in a large website.

IF ZIP CODE	(93011–94123]	
AND REGISTRATION NO.	(0051–0943]	
AND USER ID NO.	(4566JK33–4558DL01]	
THEN	/WEBSELL/PRODUCT8.HTM	87%

This rule can be broken down into the following components. First, the ZIP codes point to a group of California ZIP codes 93011 through but not including 94123, for which demographic data can be matched by percentage of households for multiple consumer lifestyles. Next, a group of registration numbers was found: 0051 through but not including 0943, which represent a specific period in time, for issue dates. Finally, a cluster of cookie identification numbers was isolated that seemingly represent an unknown association. However, additional analysis may discover similarities in browsing habits. In any event, when the combined factors of these conditions occurred, the data mining process discovered a high propensity for selecting PRODUCT8.HTM, which is one of the main pages for this commercial website.

Amazon/AlmondNet
Exhibit 1011
Page 120

A key difference and value of machine-learning algorithms over statistical classification systems such as regression is that they are able to classify separate areas in data space. For example, these types of data mining tools can discover islands of web customers which a regression approach would find difficult to classify. There are several ways that a website visitor can be identified as profitable. One type of client may make multiple purchases of low-dollar items, while another profitable customer may make single high-dollar purchases; in both instances they represent online sales.

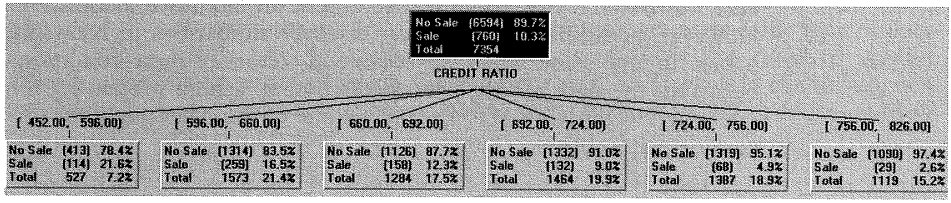
Machine-Learning Tools

A good strategy in selecting the right algorithm is to try multiple tools using a single data set on the same dependent variable and then seeing the results and evaluating the accuracy of the tree. Keep in mind also that you will often encounter more than one data set that may require having access to all of the main algorithms since they work on different data variables and data sets with different results.

The following data mining tools excel in their ability to handle large numbers of attributes and records, estimate statistical significance, handle both categorical and numeric attributes, learn transparent rules, and improve incrementally. This group of machine-learning-based data mining tools includes: *AC2, Alice, Business Miner, Clementine, C5.0 (RuleQuest), CART, Cognos Scenario, Data Surveyor, Decisionhouse, IDIS, IND, KATE-Tools, KnowledgeSEEKER, Preclass, PQ–>R, SAS Enterprise Miner, SuperQuery, Profiler, and WizWhy.*

Using a decision tree tool like *KnowledgeSEEKER*, a database can be segmented into statistically significant clusters very quickly. Like other high-end data mining tools, KnowledgeSEEKER contains multiple algorithms: CART, CHAID, and ID3. A unique feature of this tool is that it runs CART and CHAID concurrently. This means both continuous and discrete dependent and independent variables can be modeled without pre-processing, splitting, or scaling of your data.

Figure 3–12 *The leaves of a decision tree are classes of website customers.*



Amazon/AlmondNet
Exhibit 1011
Page 121

The tool generates Multivariate Decision Trees, IF-THEN rules, and SQL syntax. In the example in Figure 3–12, the data mining tool was able to generate a decision tree identifying a key *attribute* (credit ratio) and *intervals* (dollar ranges) differentiating between customers who purchased a product online and those who did not.

Genetic Algorithms

A genetic algorithm (GA) is essentially a program version of the evolutionary process. The main difference of course is that in a system a program can be modified in microseconds. GAs are really optimization program routines, guided by the principles of natural evolution. They are search procedures based on the concepts of natural selection and genetics. GAs represent a robust alternative to linear programming and mathematical optimization techniques. For an electronic retailer it represents a technology for optimizing the design and arrangement of how its website positions and communicates with its visitors.

Coupled with neural networks and machine-learning algorithms, GAs have a tremendous potential for mapping web visitor patterns and optimizing website design. Because of their inherent design they are highly malleable and well suited for the dynamic environment of the Web. John Holland developed the first GAs at the University of Michigan during the 1970s. Dr. Holland was impressed by the ease with which biological systems could perform tasks that defeated powerful computers. He observed that evolution took place in small increments at the chromosome level and that nature tended to be biased toward some chromosomes that produce a more "fit" organism. The basic concept that Holland developed involves how stochastic searches take place: rather than operating on a single solution to the problem at hand, GAs operate on a *population of solutions*.

Code of Survival

To use a GA you must encode solutions to your problem in a structure that can be stored in a computer. This encoded solution is called a *genome* or *chromosome*. These chromosomes are a population of possible solutions, in essence a set of character strings that are analogous to the base-4 chromosomes that we see in our own DNA. The trial solution—where numbers are used for the input variables—will have a result or output function, which describes the quality of the trial solution. The output is commonly called the *fitness*, since it describes how "fit" the trial solution is. The fitness function typically has a number of different inputs. For instance, the fitness function for

Amazon/AlmondNet
Exhibit 1011
Page 122

a response to an offer in your website may have such input variables as the connect time, the number of days since last purchase, the customer's ZIP code, or the total amount of purchases.

Each population member will therefore have an identical number of genes, or input variables. The GA creates a population of genomes, then applies crossover and mutation to the individuals in the population to generate new individuals. It uses various selection criteria so that it picks the best individuals for mating and subsequent crossover. The key to GAs is the "objective function," for it determines how "good" each individual is. Also important is how genomes or chromosomes are structured as possible solutions (inputs) to a fitness function.

Genetic algorithms are based on the concept of *schemata* as developed by John Holland in the early 1970s. A schema is simply a template for representing the possible patterns of a given solution, which with GAs is represented as a genome or a chromosome. This schema is represented in a series of sequential symbols of 1s, 0s, and *'s in a fixed position. For example, the schema of **01 is the following genomes: 0001, 1001, 1101, and 0101—these chromosomes all end with 0 followed by 1. They are also encoded structural solutions to problems a computer can process. The relationship between a genome and a schema is that when they match, they survive.

The fitness of a particular schema is the average fitness of all the genomes that match the schema in a given population. For example, the fitness of the above sample is the average of the genomes 0001, 1001, 1101, and 0101. Holland found that the shorter the defining length of a schema, the more likely it was bound to survive from one generation to another. The defining length of a schema is the distance between the outermost fixed positions. For example, the defining length of 1*10111 is 6 (7 – 1). Low-order schema are the code of GAs which build from one generation to another, with the fittest surviving and mixing with others to evolve into fitter solutions. Holland's Schema Theorem explains how GAs are really searching through possible schemata to find the most fit solutions for the next generation. Holland worked primarily with strings of bits, but you can use arrays, trees, lists, or any other object. What is important is that you must define the genetic operators: reproduction, crossover, and mutation.

Reproduction

This is the process by which a program evaluates strings copied according to a desired output, with strings with high values likely to contribute an offspring or more to the next generation. GAs work by starting with

Amazon/AlmondNet
Exhibit 1011
Page 123

relatively poor trial solutions, that is, population members with poor fitness. The first process, reproduction, involves an exchange of information between population members. Selection keeps the size of the population constant but increases the fitness of the next generation.

Selection is similar to the process of "natural selection" in nature, where only the fittest individuals in the population survive to pass their genetic material on to the next generation. The chance of the genome surviving to the next generation is proportional to its fitness value. For example, let's say you are attempting to determine the setting of x and y coordinates in order to maximize z on a surface. In this instance, your *inputs* are x and y with z being your *output*. The first step would be to create a population of random possible solutions, in this case some coordinates being better than others:

1. $x = 2.0, y = 4.0$
2. $x = 1.3, y = 3.5$
3. $x = 2.5, y = 6.0$
4. $x = 4.5, y = 5.0$
5. $x = 3.0, y = 2.5$
6. $x = 1.5, y = 4.5$

You next rank your solutions (selection) and keep the best coordinates, which yield the better results:

1.	$x = 4.5$	$y = 6.0$	keep
2.	$x = 3.0$	$y = 5.0$	keep
3.	$x = 2.5$	$y = 4.5$	eliminate
4.	$x = 2.0$	$y = 4.0$	eliminate
5.	$x = 1.5$	$y = 3.5$	eliminate
6.	$x = 1.3$	$y = 2.5$	eliminate

Crossover

When *crossover* occurs, two strings exchange information that yields new combinations of scenarios. Survival is based on choosing the genomes in a random way proportional to their fitness. This involves the selection of two individuals, the selection of a crossover site, and the exchange of values between the two strings. Crossover, which occurs in nature, creates two new genomes from two existing ones by gluing together pieces of each genome. When population members mate, they cross gene values over to their partner, resulting in a "trade" of values.

Amazon/AlmondNet
Exhibit 1011
Page 124

This mixes up the information in the gene values of the population members, creating new and diverse "offspring." Our example of those higher ranked coordinates would follow this process as follows:

	Prior to Crossover	After Crossover
Scenario 1.	2.5, 6.0	4.5, 2.5
Scenario 2.	4.5, 5.0	6.0, 5.0

Mutation

Mutation is a source of variation used to maintain diversity in a population of possible scenarios. In a binary-coded genetic algorithm, it is the occasional alteration of a bit position. Once the crossover has taken place, the offspring undergo mutation. The values of the individual genes of population members can be changed, or mutated. Mutation is important in evolution because unlike crossover (which merely *trades* genes), *new* gene values are introduced. This further increases the diversity of the population members.

Reproduction and crossover do a good job of searching the space for possible solutions, but they depend on initial conditions and randomness that might not happen; mutation provides the additional catalyst for new combinations to occur. Mutation, however, is held to a minimum because radical changes can be as destructive as they are productive. They should only be used to avoid premature convergence to a local optimum.

Once crossover and mutation have occurred, the new population members will replace the old ones if they are better. In this fashion, population members can have improved fitness with each new cycle (or generation) of crossover, mutation, and natural selection. Program users typically want to maximize or minimize the fitness, so the survival of the fittest action is governed by whether the new fitness is better or worse than the previous generation's fitness. The resulting process yields a steadily improving solution to the problem, often identifying the optimum solution in a surprisingly short time.

A Step-by-Step Example Using a Binary-Coded GA

1. Objective: Maximize the function.

$$F(x) = X^2 \text{ (Objective)}$$
$$\text{Where } 0 = < X = < 31 \text{ (Range)}$$

Amazon/AlmondNet
Exhibit 1011
Page 125

2. Code the schema. Use a binary string of length = 5.

   Example: $11000 => 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 0 \times 2^0 = 24$

3. Reproduction of the initial population, e.g., 4 strings:

String Number	String	×	Fitness	% of Total
1	00010	2	4	1.8%
2	00011	3	9	4.0%
3	01110	14	196	87.1%
4	00100	4	16	7.1%
Average Fitness = 56.25				

Initiate reproduction via a roulette wheel four times to get the following:

String Number	String	Number of Hits
1	00010	0
2	00011	1
3	01110	2
4	00100	1

4. Crossover. Members of the newly reproduced chromosomes are mated at random. Suppose string numbers 3 & 4 and 3 & 2 are mated.

For 3 and 4	Split	Before	After
String 3	01110	01110	01110
String 4	00100	00100	00100
String 2	00011	00011	01011
String 3	01110	01110	00011

5. Mutation. Supposing we selected p = .002, we would expect 20(.002) = .04 bits to undergo mutation.

Amazon/AlmondNet
Exhibit 1011
Page 126

6. The new population is:

String Number	String	×	Fitness	% of Total
1	01110	14	196	53.1%
2	00100	4	16	4.3%
3	00110	6	36	9.8%
4	01011	11	121	32.8%
Average Fitness = 92.25				

This process continues to iterate until an acceptable solution is arrived at.

Data Mining with Genetic Algorithms

Genetic algorithms work by evolving successive generations of solutions, which get progressively better by splitting and matching. As with nature, the idea is that only the fit survive. As such, GAs are optimization engines, which are currently being used in data mining applications to improve the performance of neural networks. Most data mining tools are designed for discovery, while genetic algorithms are really tools for optimization. However, when used in tandem a GA can be used to optimize the performance of a neural network in classification and prediction. Although not as common in today's data mining tools, genetic algorithms are beginning to make their way into the standard methodology of optimizing the performance of neural networks. Several data mining tools are incorporating both neural networks and genetic algorithms, using the GA to optimize the settings, the topography, and the learning rate of the neural network in order to improve their overall performance.

Genetic algorithms can also be used in conjunction with rule generators to achieve higher performing prediction rules. The approach involves (as with most GAs) establishing a fitness function, in this case a classifier error rate. The genomes or chromosomes are the set of rules generated by a machine-learning algorithm. The idea is to focus on the generation of IF/THEN rules and then go through the recombination and mutation of these chromosomes (rules) in order to weed out the weakest ones so that the fit rules meet a minimum cluster size (accuracy rate). The process involves establishing classifier *committees*, which are sets of distinct classifiers whose predictions are averaged:

Amazon/AlmondNet
Exhibit 1011
Page 127

Rule 1:	0.93	IF Age 32	THEN will purchase product X
Rule 2:	0.66	IF Gender M	THEN will purchase product Y
Rule 3:	0.87	IF Prior Sale Y	THEN will purchase product Z
Committee Average: 0.82			
Average propensity to purchase product groups X, Y, Z			

Repeated runs of the genetic algorithm system are performed using different random number seed rules. The process involves all three steps (reproduction, crossover, and mutation) to insure poor rules (classifiers) are eliminated by the committees. The concept of using genetic algorithms with machine-learning algorithms is meant to optimize the search space and improve the overall accuracy of their predictive models.

There are few commercial genetic algorithm data mining tools in the market, making it difficult to incorporate the technology in the analysis of your web data. On the desktop level there is Evolver, an add-on to Microsoft Excel. The product has been around for some time and is quite robust and built on solid theory. From the Netherlands there is a newer package called The OMEGA Predictive Modeling System, which uses a GA to optimize the learning process. Another tool known as the NeuroGenetic Optimizer (NGO) incorporates a GA with back-propagation and Self-Organizing Map neural networks. On the high-end there is the NeoVista Solutions software, which includes a genetic algorithm module as part of its data mining suite of tools. Public code is also available from several universities on the Web for those who wish to develop their own applications.

Genetic Algorithm Applications

One of the most intriguing applications of GAs is in the area of retailing, specifically for the optimization of product shelf display size and location. Several retailers have begun to use GAs to design or improve their store and product layouts to maximize revenue opportunity given various store demographics. One of the problems retailers have, much like large e-commerce websites, is dealing with a large volume of products. Coupled with this are the changing and diverse demographics of store shoppers, each requiring a detailed and accurate reorganization of unique products for each store. E-commerce websites selling multiple products and services experience the same problem.

Amazon/AlmondNet
Exhibit 1011
Page 128

The current solutions require extensive planning and resources, but GAs are ideal for this type of logistics problem, which is rapidly replicating itself in a large scale on the Web with large electronic retailers. Coupled with the problem of design is the problem of inventory control, which again GAs can optimize, by reducing overstock and out-of-stocks and increase inventory forecast accuracy.

Through the use of GAs, websites can also maximize their sales by improving their web design and product (web page) organization for improved cross-selling. GAs can improve the use of inventory dollars and increase the affinity grouping of products based on website visitor demographics. Affinity grouping can also assist in the positioning of product items based on web customer profiles, much like retailers are currently doing it with the aid of GAs. The concept of evolutionary retailing is one of using GAs and neural networks to optimize the design and functionality of online storefronts and portals to maximize retail space design.

The AI Cousins: Case-Based Reasoning and Fuzzy Logic

Two other AI-based technologies have the potential of assisting you in the optimization of your electronic retailing efforts and customer service. One is *case-based reasoning* (CBR) and the other is *fuzzy logic.* Although few data mining tools include their technology, you still should know something about how they work, what they are capable of doing, and how you may be able to deploy them in the future.

Case-Based Reasoning (CBR)

A CBR system works by matching new problems to "cases" from a historical database, and then adapting successful solutions from the past to current situations. CBR has typically been used for customer support, quality assurance, and maintenance applications. In these situations, CBR maintains an institutional memory of prior problems so that when a new call comes in, old solutions are retrieved and matched to answer similar new problems. CBR is a simple yet powerful method of indexing, which is more akin to *remembering* than to learning, the way the other data mining technologies operate.

CBR is a collaborative system that stores case samples of solutions it then uses to recommend fixes to new problems. The most common use of this indexing system is with trouble help desks, where CBR technology is used to respond to problems by finding similar cases in its memory, and adapts the solution that worked before to the current

Amazon/AlmondNet
Exhibit 1011
Page 129

problem. CBR technology was developed originally in the 1980s, most prominently from the work of Roger Schank at the Yale University AI lab. Obviously, a similar application can be found in the collaborative filtering software, like Firefly for recommendation of products and services based on case samples.

CBR works by maintaining a database of known records, so for example, you could store a database of prior website customers who've made large dollar purchases. Using this CBR engine you could profile new website visitors who match the prior website customers for classification and prediction. CBR uses distance and combination functions to do the matching between old customers and new visitors. CBR is also quite adaptive, so as new customers are identified new types of neighborhoods are created, permitting new web visitors to be targeted.

The whole concept is that of association: new website customers likely to respond to an offer are probably similar to previous customers that responded in the past to like offers and incentives. CBR works easily with all types of data, including non-relational text and images. Still, there are drawbacks. CBR requires a large amount of historical data for finding neighbors. The classification of new records can require processing all the historical records to find the similar profitable neighbors. It is also highly computationally and storage intensive. The strengths of CBR are of course its simplicity, which boils down to: *"I think this website visitor will respond to this ad because he is a lot like these other website visitors who clicked on the link."*

The number of vendors offering CBR technology is limited. Inference's CBR Express is one. The most immediate application of CBR technology is to the area of customer service, where visitors can interact with a CBR engine in order to come up with solutions that have previously been resolved.

Fuzzy Logic

The name is deceptive; it should really be "continuous value logic," which provides a superior method of information modeling. Conceived in 1965 by Lotfi Zadeh, an engineer and professor at the University of California at Berkeley, fuzzy logic provides a method of reducing as well as explaining system complexity. Dr. Zadeh was concerned with the rapid decline in information afforded by traditional mathematical models as the complexity of the target system increased. He expressed his theory as follows:

Amazon/AlmondNet
Exhibit 1011
Page 130

> *As the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance (or relevance) become almost mutually exclusive characteristics.*

Much of this complexity, he realized, came from the way in which the variables of the system were represented and manipulated. Since these variables could only represent the state of a phenomenon as either on or off, the math necessary to evaluate operations at various "border" states became increasingly complex, which at some point became a morass of equations providing little insight about the underlying process. Zadeh called this state the *principle of incompatibility.*

Fuzzy logic is a calculus of compatibility. In fuzzy logic there is no black or white, there is instead a degree of gray in both. Everything is measured in *degrees of membership.* Fuzziness, the measure of how well an instance (value) conforms to a semantic ideal or concept, describes the degree of membership in a fuzzy set. This degree of membership can be viewed as the level of compatibility between an instance from the set's domain and the concept overlying the set. For example, in the fuzzy set WILL BUY, the value YES has a degree [.31] meaning that it is only moderately compatible with WILL BUY. This is called a "measure of fuzziness" because it is used to assess the degree of ambiguity or uncertainty attached to each fuzzy set. It is important to realize that a website visitor can be a member of different fuzzy sets simultaneously:

IF WILL BUY THEN [.31]

IF WANTS INFO THEN [.35]

IF WILL NOT BUY THEN [.34].

More and more, fuzzy logic is being used with neural networks, much like genetic algorithms. However, it is not being used at the front end of neural networks, to optimize their setting, but instead to explain their "black-box" findings. The key benefit of fuzzy logic is that it allows for imprecise states to be categorized with simple "IF/THEN" relations, which are easy to verify and optimize. Conversely, a key benefit of neural networks is that they learn from data sets; thus a combination of both technologies provides the best of both worlds. One data mining vendor who already incorporates this type of fuzzy logic and neural network mixture is DataEngine from Management Intelligenter Technologien GmbH. For the Web data miner, fuzzy logic offers a powerful technology for data com-

Amazon/AlmondNet
Exhibit 1011
Page 131

pression and insight from the hundreds of thousands of online patterns. Fuzzy logic is also being used as a new method of buying and selling stocks and eventually retail products via the Web by the OptiMark system.

Conclusion

It is important to remember that today's data mining tools stem from the three branches of artificial intelligence designed to emulate human *perception, learning,* and *evolution.* It is equally important to understand how to strategically position these tools in order to enhance the value of your website and electronic retaining presence. Incorporated in today's modern data mining tools are technologies that can extract powerful knowledge and insight about your online customers in the form of business rules, C code, decision graphs for segmentation, classification, prediction, clustering, and optimization.

Don't limit yourself to just one tool or one technology; multiple tools are best for pruning your data and optimizing the process of learning. Use a tool incorporating a machine-learning algorithm to compress your data and to identify the important attributes in your data. Use a neural network to converge on a model, using the attributes identified by the machine-learning algorithm. Use a genetic algorithm to optimize your neural network to improve the accuracy of your model. Use a fuzzy logic system to extract rules from a neural network.

Finally, recognize that these core AI technologies may soon migrate to software other than data mining tools. Some of the segmentation algorithms like C5.0 or CART may soon be incorporated in web analysis software. Likewise, neural networks may soon become a part of new electronic retailing server products. What is important is that you be aware of how each of these technologies can improve your online sales by allowing you to retain and attract your current and future customers. Just as physical retailers like Wal-Mart have used genetic algorithms and neural networks to optimize their stores, so can they be utilized to optimize electronic retailing sites like Amazon.com. The core technologies of data mining are quite mature and robust, having been around for years; how they are used in the new interactive networked environment that is the Web will provide some unique new challenges for e-miners, e-marketers, and e-merchants.

Amazon/AlmondNet
Exhibit 1011
Page 132

4

Ten Steps to Mining Your Data

The mining of your web data starts well before you begin using a data mining tool, and it continues well after your analysis. Several steps are involved prior to the actual analysis of the data in order to ensure the data mining process is a success and a profitable endeavor. Although the data mining tools simplify and automate the analysis of databases, care must be taken in using them or they may lead to faulty findings and erroneous conclusions.

Identify Your Web Business Objective

The mining of your website involves some advanced planning about what type and level of information you intend to capture at your server and what additional data you plan to match it with. This by itself will ensure your data mining efforts will yield measurable business results. For example you need to plan with your web team what kind of log, cookie, and form information you intent to capture from your visitors and at what juncture. Next you need to involve your business, sales, and marketing teams in deciding what kind of demographic and household information you need to purchase to merge with your server data. In addition, you should consider bringing in your information system team for the possible incorporation of your datamart or data warehouse and customer database into the mix.

This advance planning can save you time and money in your data mining efforts. It can also translate into quicker results and thus ensure your success in the mining of your website. Keep in mind the principle that having more data is better than not having enough. Keep also in mind what your business and marketing goals are and what kind of data is required in order to achieve them. For example, if your goal is to prevent customer attrition, then you are likely in

Amazon/AlmondNet
Exhibit 1011
Page 134

need of developing a "churn" model. In order to construct such a model you should secure samples of customer accounts that have recently closed or terminated so that your analysis can examine the patterns and differences between customers who are active and those who are inactive.

This preplanning step may require the use of modeling tools and techniques for building a blueprint for creating a road map to the data mining initiative. Modeling techniques can form the foundation to the mining application and ensure it meshes with the actual business requirements. This modeling initiative can also alert you to possible obstacles, such as the lack of access to certain data for the analysis. Another benefit is that it ensures you document your data mining efforts. Process-modeling tools also provide a benchmark for measuring improvement and a common language for communicating across your firm and the various teams and management. The following are a few of the modeling tools you can use to ensure a successful data mining project:

Product	Vendor	Website
Corporate Modeler	Casewise Systems	casewise.com
ProCarta	Domain Knowledge	domainknowledgeinc.com
Aris Toolset	IDS Scheer	ids-scheer.com
LiveModel	IntelliCorp	intellicorp.com
WorkFlow Modeler	Meta Software	metasoftware.com
Designer/2000	Oracle	oracle.com
Aion	Platinum Technology	platinum.com
ProVision Workbench	Proforma	proformacorp.com
Visio	Visio	visio.com

Before you begin a data mining project, be clear on what you hope to accomplish with your website analysis. The mining of your web data must be looked on as a strategic competitive move, which like all other business initiatives should have a measurable goal. Most likely you will not be performing a data mining project in a vacuum; other divisions or individuals in your firm are likely to be involved in the project. For a large retailing website this may include the following individuals:

Amazon/AlmondNet
Exhibit 1011
Page 135

Server Engineer	Website Analyst
Internet Engineer	Production Engineer
User Interface Designer	Production Content Manager
Web System Administrator	Design Lead
Web Technologist	Web Content Producer
User Interface Analyst	Promotions Manager
Brand Manager	Web Publisher
Internet Application Developer	Forum Manager
Online Documentation Manager	Web Content Editor
Web Partner Services	Web Marketing Editor
Web Account Manager	Web Administrator
Web Producer (Marketing)	Web Designer
Software Developer	Partner Advertising Administrator
Graphic Artist	Production Marketing Manager
Web System Designer	Web Content Publisher

Ensure they participate in the establishment of expectations, whether in terms of online sales, web traffic, or other measurable goals. Most likely you are already aware of the current state of affairs with your website, your Internet presence, and electronic retailing efforts. These are some of the questions you may want to consider at this juncture:

1. How much of an improvement do you need on your advertising or customer response rates?
2. What area of your website do you hope to improve by data mining customer transactions?
3. Is the business goal of your data mining analysis measurable, and if so, what is your baseline for success?

Recognize that you need to do some documentation regarding your data mining analysis both prior and after your analysis. The data mining of your transactional data will definitely have positive results on your website operation and overall bottom line, so the question you need to frame is "by how much?" What you need to be clear on is by how many responses, e-mail requests, online sales, clickthroughs, etc.

Baseline

One way of establishing these baselines is by looking at the current rate of sales for each of your main product lines. For example, if only 80 out of every thousand website visitors click through and purchase Product A on your website, then your baseline for that product is 8

Amazon/AlmondNet
Exhibit 1011
Page 136

percent, and your data mining goal will be to increase that rate for Product A. If you have multiple products, you need to establish baseline rates for each, or at a minimum do it for main product lines, so you can begin to gauge the impact your data mining analyses are having on your overall sales and marketing efforts. You need to establish some general baselines, whether it is in the form of average respond rates or average sales, in order to measure how much better data mining works over random chance. For starters here are some possible business goals for the mining of your web data.

Are you looking to identify potential new website customers? This is probably the most common data mining objective. Classification most commonly involves discovering the attributes, characteristics, or features of your website customers. This objective is best achieved by having as much information as possible about who buys and who doesn't while in your website. Classification typically involves distinguishing between revenue web visitors and nonrevenue web visitors. Classification is a clear-cut black-and-white prediction.

In order for a data mining tool, such as a neural network, to learn the unique features of your website customers, it needs to be trained with a large number of samples. You will need examples of website customers with a large number of features, which you can begin to gather in advance of the actual data mining analysis. Start by designing your registration forms to capture as much as you can about your customers. Explore how you can link your website data with existing internal databases such as your customer information file or your data warehouse. Plan to invest in the purchasing of external demographic or other third party data, which can be overlaid with your web data to enhance its value.

Are you looking to find specific website product sales trends? Are you looking to find revealing online trends or relationships between certain web pages representing individual products or services? In other words, is your data mining goal that of discovering a unique online *association*? This type of discovery can assist you in the positioning of certain web pages, offers, incentives, and links. The discovery of an association between unique products and services can in fact impact your overall website design. If they click on Product001.htm, what is the probability they will click on Product002.htm, Product003.htm, etc.? More importantly, what is the probability that if they buy Product003.htm, they have a propensity to buy Product009.htm?

Are you looking to identify specific buying patterns over time in your website? This may be a sequencing issue, which is usually an association problem with an additional dimension of time. Sequencing

Amazon/AlmondNet
Exhibit 1011
Page 137

involves evolving trends and populations; in the case of your website, this is usually monthly or weekly sales of certain products and services. For high-traffic websites, like a search engine or a popular content provider, this may be accelerated to hourly trends or patterns. Many web marketing and electronic commerce forecasting problems involve sequencing or time-series analyses with data being sampled at fixed intervals of time.

Sequencing is generally more involved, since instead of looking at a single recorded web transaction with one measured value for each feature, are you instead looking at that same feature measured at different times. One of the most common web management problems is attempting to anticipate traffic in the future: How many visits or sales are you likely to have based on observed patterns over time? This is clearly a sequencing problem. Another sequencing problem involves online patterns once a web visitor or customer logs into your site. Here the sequencing problem, which may be tracked by your cookies, is trying to map the movement of your website visitors. This involves mapping the path of your website visitors, which is very much like association, but involving the additional element of time.

Keep in mind that prediction, association, classification, clustering, and sequencing are not exclusive goals and in fact may overlap as a result of your data mining analysis. What is important, however, is to be clear about the business objective, since it will influence both how you select your data and tool and, more importantly, the format of your solution. Have a clear focus on the business goal you are trying to obtain.

- Are you trying to increase response rate to a web campaign?
- Are you trying to identify what products tend to sell together?
- Are you trying to identify the features of large shopping carts?
- Are you trying to identify the features of your most profitable clients?

Select Your Data

Once you've defined your business objective, your next step is to choose the web server and company data for meeting this goal. Here is a quick checklist:

- Is the data adequate to describe the phenomena the data mining analysis is attempting to model?
- Is a common field in your web data being used for linking to other databases?

Amazon/AlmondNet
Exhibit 1011
Page 138

- Can the data from your web be consolidated with your data warehouse?
- Will the data being mined be the same and available after the analysis?
- What internal and external information is available for the analysis?
- How current and relevant is the data to the business objective?
- Are the data sets being merged consistent with each other?
- Who is knowledgeable about the data being gathered?
- Is there redundancy in the data sets being merged?
- What joints are needed for the various databases?
- Is there lifestyle or demographic data available?

Creating Customer Data

One of the advantages of mining your website is that the customer data can be created on the fly. For example, several advertising models provide methods by which to sell your products and services on the web. However, they also provide methods by which to generate customer data for mining. Among the several marketing models, the most common are brand sites, banners, buttons, pop ups, chat, push, e-mail, and agents.

Brand Sites

These were some of the first websites set up—where a company wanted a site simply because everybody else was doing it. Originally, few were doing any electronic commerce. They were simply on the Internet to build brand awareness. Brand sites usually contain useful content with some type of contest, game, newsletter, or promotion, and are usually on the Web for purposes of building customer loyalty and recognition.

These sites span the realm of products and services, as the examples in Figures 4–1 and 4–2 demonstrate. Figure 4–1 typifies the new generation of sites, which are already doing some merchandising. Even these websites have the ability to gather data for mining through the use of log and cookie files as well as forms. They can also use other web-specific commercial software and services such as BroadVision's Relational, GuestTrack, DoubleClick, LikeMinds, and FireFly to further profile their visitors and their propensity or affinity toward their products and services. These branding sites, simply by their presence, can begin to interact and capture some general demographics about their visitors.

Amazon/AlmondNet
Exhibit 1011
Page 139



Figure 4–1 *This site provides games, information, and a shop for making purchases.*

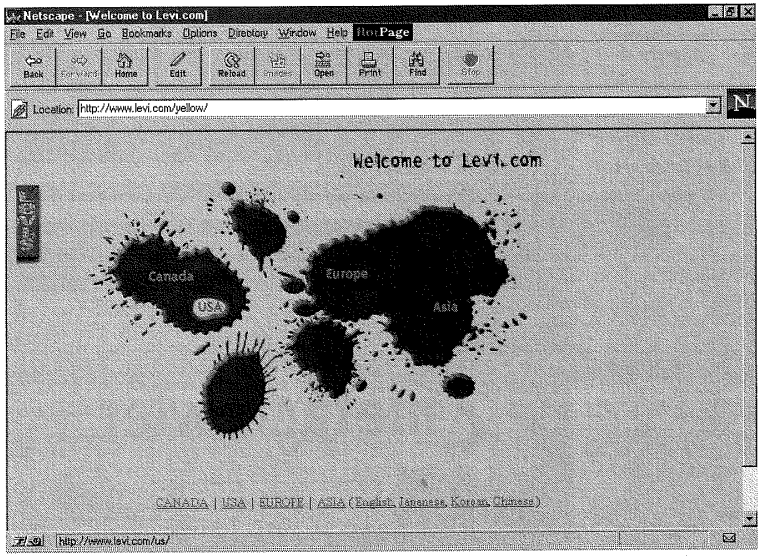


Figure 4–2 *An elegant brand building site.*

Amazon/AlmondNet
Exhibit 1011
Page 140

Banners

Banners are rectangular, static, and dynamic billboards commonly found on the top of search engines, online directories, and other popular websites (Figure 4–3). They commonly entice the users to click through the ad, which will take them to another website. These types of advertisements can attract visitors by interacting with them, which again leads to the generation of visitor information for data mining.

Typically, clickthrough for most banners is around 2 percent through general rotation, but in some instances banners can improve this rate to 6 percent by linking the type of banner to presets on the basis of "keywords" input by the user of search engines or online directories. A rate of around 9 percent can be achieved on the basis of topic directions. Special ad servers like SelectCast from Aptex, using neural network and proprietary indexing technology, claim rates of up to 14 percent. Other techniques for improving clickthroughs are the use of traditional marketing words like "new," "free," or "click here" or by the use of motion or certain bright colors. Studies indicate that animation improves clickthrough rates by as much as 15 percent over static billboards. Other web design techniques include pull down menus and interactive games and ticker messages.

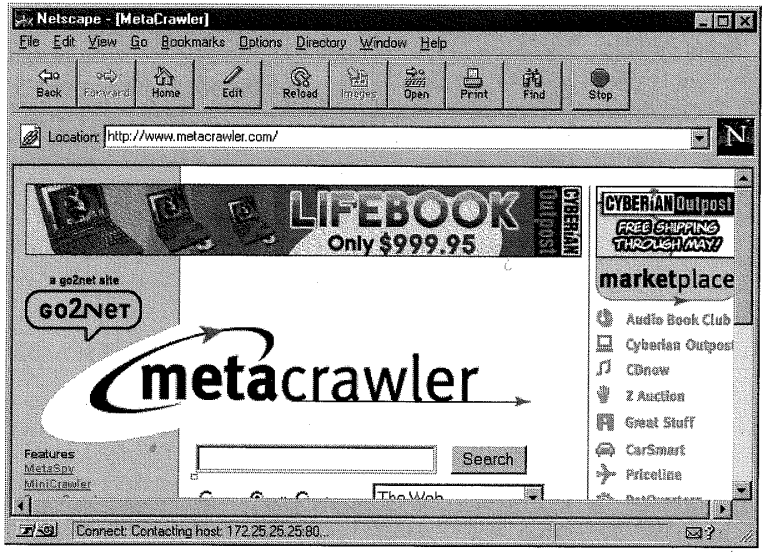


Figure 4–3 *This banner sits atop a popular search engine and rotates with each visit.*

Amazon/AlmondNet
Exhibit 1011
Page 141

Buttons

Buttons windows, usually about a quarter of the size of the normal-sized banners, typically are very limited in the information they provide and are narrowly targeted to a very small audience. For example, you may find them on a website designed to sell software or hardware with a button designed to take the visitor to a computer brand site or a browser site like Netscape or Microsoft (Figure 4–4).

Other ways of getting additional clickthroughs include positioning buttons in popular online magazines or industry-specific websites. They are product- or service-specific and can be closed by the user so that they are not as intrusive as a banner. They are, however, another method of attracting visitors and aggregating customer data for mining.

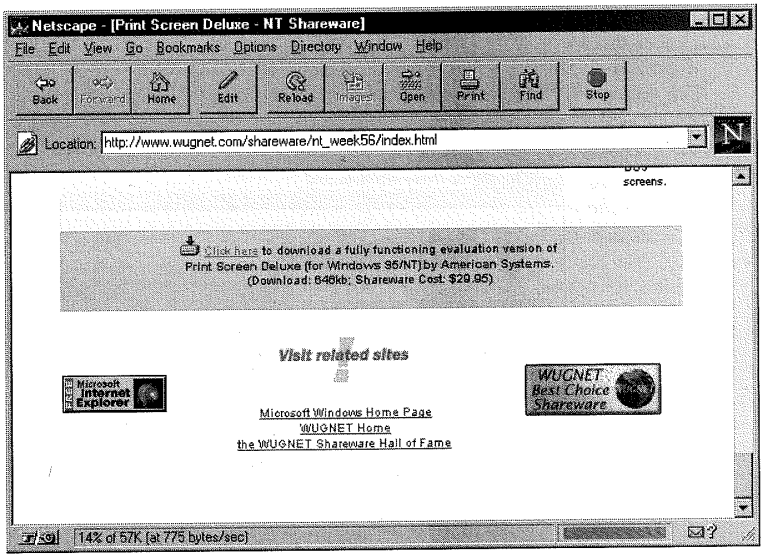


Figure 4–4 *Note the two browser buttons at the bottom of this page.*

Amazon/AlmondNet
Exhibit 1011
Page 142

Pop Ups

Like buttons, pop ups are a novel concept of advertising and promotion. They are often launched within a website for purposes of pitching a special product or service. The difference between buttons and pop up windows is that they are self-contained and generally cover a square of the browser window. They are designed to grab the attention of the visitors while at the same time not taking them away from the main home page (Figure 4–5). Visitors can interact with the pop up either with a self-contained form or banner while at the same time remain at the main website where they can also interact or shop. The pop ups, like the buttons and banners, have the ability to gather personal customer information about potential prospects which can be mined.

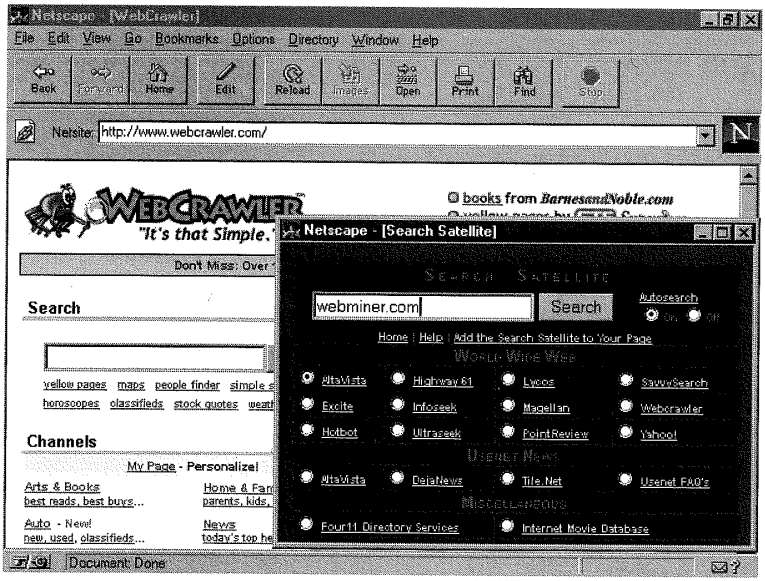


Figure 4–5 *This pop up window can be accessed without the need to leave the website.*

Amazon/AlmondNet
Exhibit 1011
Page 143

Push

Also known as "webcasting" or channels, users are allowed to select the type of content they want "pushed" to them. As with chat, most push sites like PointCast require some initial registration and selection of specific topics, which again is important in gathering information for data mining. Because visitors select and reveal what information they are interested in getting pushed to them in a registration form often requiring gender, age, and ZIP code, important data is compiled for data analysis. When visitors sign up with a channel provider they fill out a menu of topics and interests, which becomes their preference profile. Some of the key players in push technology are BackWeb, Intermind, NETdelivery, and Marimba.

- *BackWeb* is able to push personalized content from multiple sources, including legacy files, in real-time.
- *Intermind* is able to integrate legacy and intranet systems via Notes/Domino. Extranet channels are also available for customer-support applications.
- *NETdelivery* is a special delivery service that can provide mailbox icons on the user's screen.
- *Marimba* and its Castanet technology is designed to reduce bandwidth demand by only updating content that has changed. Castanet is designed to distribute and manage itself automatically.

E-mail

E-mail provides the ability to communicate easily with visitors, prospects, and customers. While e-mail enables a relationship to form and increases purchases or visits, it can also capture data about customer interests or concerns. E-mail enhances a site's ability to form loyal relationships with customers through personalized dialog. The more a site knows about its visitors and customers, the easier it is to provide them with the products and services they desire. E-mail is a very effective and inexpensive way to acquire and maintain a customer. It is non-intrusive and ubiquitous.

Care must be taken, however, when using e-mail to capture client data. You must make sure you allow the customers to sign up voluntarily and to only send them the information they request at the intervals they request, and most important, give them the ability to get off your mailing list. Several mail servers exist allowing you the ability to create a database of recipients based on customer profile information. Servers and software products such as Aptex, Egain, Majordomo, Pegasus, Revnet, ListProc, AltaVista, Arial, InfoPress, and UnityMail allow you to build forms for capturing customer data directly into a database for mining. The Kana Customer Messaging System e-mail software contains a module that uses AI to analyze inbound messages and sort them correctly.

Amazon/AlmondNet
Exhibit 1011
Page 144

Chat

Yet another model for attracting visitors and data gathering for mining are the chat sites found in high traffic locations like search engines and online directories. Since most chat sites require registration information about visitors, data can be gathered and used for targeting ads based on demographics, psychographics, gender, age, income, and of course the chat discussion topic. Chats also tend to be topic-specific, which can provide a method for data gathering about visitor interests.

Agents

Agents or "collaborative filtering" software programs make recommendations based on prior web activity by individuals with similar demographic or psychographic profiles. Affinity or collaborative programs are designed to make recommendations on the basis of what other people have liked with similar likes and dislikes to those of you. They make their recommendations based on your preference input and the activity of a community of users similar to you. The most prevalent of these collaborative filtering sites and software is FireFly, a pioneer in using AI technology for commercial purposes. These types of collaborative filtering applications are usually deployed in websites selling music, videos, books, and other types of recreational products and services.

Some agents act as filters, browsing through large amounts of data and providing their owners only the information that matches their criteria. Most intelligent agents are typically not website-specific. They instead involve the retrieval of the specific type of intelligence you furnish during the registration phase. There are several types of "agents." Some are server-specific while others are designed to cruise the Web in your behalf. For example, some agents are designed to roam the Web for specific information you are interested in. This can be keyword-specific or industry- or company name-driven, or take the form of information on a specific technology, news topic, product announcement, stock price or activity, company name, etc.

The intelligence of these agents varies. Some are simply able to search on specific topics the owner specifies, and others, which are smarter, are able to look over your shoulder and cue on the specific keywords you typically search on, gradually "learning" what your interests are and through these observations suggest or search for information, products, or services in your behalf. Because agents require user registration and specific personal information and topics or keywords they represent yet another method of gathering important personal information for data mining.

Amazon/AlmondNet
Exhibit 1011
Page 145

Dealing with Transactional Data

When you first encounter transactional data gathered from your website, find out exactly how it was done, what guidelines are in place, and how often the logs and forms databases are updated. As in the case of constructing a data warehouse, bringing together data from diverse systems often involves dealing with flat-files, relational tables, and hierarchical and relational systems from multiple operational systems that at times deal with the same attributes (such as dates) in different ways. The coordination and selection of these data sets will involve considerable effort and time. In fact, this may be the most time-consuming part of the entire data mining effort.

Key issues likely to drive your decisions in the selection of your web data are the portability and cost of both internal and external databases. A related issue may include security and privacy, especially in the use of any information provided by your website customers. Keep in mind that some of the information required for profiling your customer might not be available—be prepared to do some data gathering. Be prepared also to deal with inconsistent data definitions, different encodings of data, and inconsistent values when selecting and merging your data from multiple sources.

The data requirements for your mining analysis are directly linked to the output or desired insight that you wish to obtain from your web database. For most analyses, the general rule is the more the better. A second rule in the selection of your web data is that you should strive to have a good sampling of both types of customers or accounts, both positive and negative samples: sales and no-sales. Keep in mind that in your analyses the more descriptive information you have on a website visitor and customer, the better the predictive models you are likely to obtain. First and foremost you must recognize why you are accumulating the data which you will be mining—similarly be prepared to acknowledge that you may need additional data you don't currently have.

The bottom line is simply *know the data!* If you are not conversant with the data, you may want to enlist the service of someone who has knowledge of the problem domain. It is important to select the important variables in the data, such as the dependent (outputs) and independent (inputs) variables. When dealing with files or databases from some other departments, find out why the information was gathered in the first place. It is important to have the participation and involvement of the retailing, marketing, shipping, sales, and of course, the web team members during this process.

Amazon/AlmondNet
Exhibit 1011
Page 146

Prepare Your Data

Once the data has been assembled and visually inspected, some decisions must be made regarding which attributes to exclude and which attributes need to be converted into usable formats.

- What condition is the data in, and what steps are needed to prepare it for analysis?
- What conversions and mapping of the data is required prior to the analysis?
- Are these processes acceptable to the users and the deliverable solution?
- What strategies must be taken for handling missing data and noise or outliers?
- How skewed is the data? Are logarithm or square transformation needed?
- Do you need to do 1-of-N conversion for categorical fields?
- Do you need to normalize dollar fields by dividing them by 1000?
- Do you need to convert purchase dates to continuous values?
- Do you need to convert addresses to sectors?
- Do you need to convert Yes/No fields to 1/0?

A graphical tool or a good file editor can assist you in inspecting the physical state of the data. A visual inspection should provide you an overview of the number and percentage of blank fields in the data set. Also, a statistical tool can assist you in identifying important relations between variables in the data. This, however, may not help in large data sets. Operational data is organized to be compact for speed and efficiency; it is not organized for analysis. Carefully review its format and be prepared to convert it into a format that will yield insight from its analysis. For example, it may be efficient to have an account establish date in the format of MM/DD/YY, but it may be necessary to transform this field to one that equates to "Total Number of Establish Account Days: NNNN."

If you are dealing with a very large database, don't do wholesale conversions on the entire data set. It is safer to first do a random extract, perform the transactions, mine the data, and evaluate the results. If you are considering using a neural network tool, additional conversion of the data will be required so that categorical values are converted to 1-of-N values and all continuous values are converted to ranges between 0 and 1 or into log or square functions. One common method of

Amazon/AlmondNet
Exhibit 1011
Page 147

smoothing data is to reduce the number of distinct values for a feature. A similar technique can be used to "discretize" continuous features into a set of true-or-false features, each covering a fixed range of values.

Missing data presents a unique problem, especially if you are using a neural network tool, since they cannot deal with missing values. Most modern data mining tools provide options on how to deal with missing values. Care should be taken, however, on whether a missing value represents a negative response that needs to be recoded and used. Consider replacing all missing values with a zero or an average mean. The drawback to this is that the replaced value is not correct. Lastly, randomly extract a sample of the data, especially if you are dealing with a large database, and start with a small prototype model. Start small and expand with your data mining analysis. For example, rather than constructing multiple response models for all the products in your retailing site, you may want to test only a few for your major product lines.

Enhancing Your Data

One of the strategies you may want to take in preparing your data for mining is linking it to other databases your company may already have from production files. For example, here are a few file types with their content:

Company File Types	Content
Transactional Data	Type of Products or Services Sold
ZIP Code or Customer Demographics	Gender, Age, Income Level, Children
Outbound Mail	Response, Location, Household Data
Telemarketing	Response Rate, Number or Mailings
Customer Information	Total Number of Sales, Client Since Date
Data Warehouse	Customer Profile by Product and Service
Datamart	Customer Profile by Division or Branch
Consumer Credit Report	Credit/Debit Cards, Occupation, Mortgage

Amazon/AlmondNet
Exhibit 1011
Page 148

Evaluate Your Data

A structural evaluation of your data should be performed in order to determine what type of data mining tools to use for your analysis.

- What is the ratio of categorical/binary attributes in the database?
- What is the nature and structure of the database?
- What is the overall condition of the data set?
- What is the distribution of the data set?
- How skewed is the data set?

Utility Tools for Preparing and Viewing Your Data

Quite often during this phase of a data mining project you will be required to visually inspect large text data sets, such as a server log file. Sometimes you may experience import errors, which can only be detected by physically opening an ASCII file and scanning down to a line item. The following software tools are quite useful for doing these types of tasks:

UltraEdit-32
This is a powerful and useful text editor with the following database handling features:

- Disk-based text editing—up to 2GB file size with minimum RAM used even for multi-megabyte files
- Insert string feature at every increment (ideal for formatting DB records and inserting hard returns)
- Insert/fill columns, insert line numbers and sum columns/selection feature
- Column/block editing—allows insertion/removal of columns of data

When dealing with databases from different sources, you are bound to have to deal with different file formats. As such you will need to import data sets and convert them into a uniform format which you can then import into your data mining tool. The following tool is very good in doing these types of conversions.

Data Junction
This data conversion utility tool permits users to translate data to or from almost any structured file format, including databases, flat files, spreadsheets, ASCII, binary, EBCDIC, reports and text files, SQL, ODBC, COBOL, ISAMs, vertical applications, mailmerge, Stats/Math, and many others. Data Junction has filter, edit, and mapping capabilities for modifying characters, fields, and records into the exact format you required. Specifications can be selected interactively through a powerful GUI, or via batch automation.

Amazon/AlmondNet
Exhibit 1011
Page 149

As a general rule, neural networks work best on data sets with a large number of numeric attributes. Machine-learning algorithms incorporated in most decision tree and rule-generating data mining tools work best with data sets with a large number of records and a large number of attributes. Empirical studies have shown that the structure of the data critically impacts the accuracy of a data mining tool. For example, data sets with extreme distributions (skew > 1 and kurtosis > 7) and with many binary/categorical attributes (> 38%) tend to favor machine-learning-based data mining tools.

Often, derived ratios of input fields may be required in order to capture the impact or the true value of the inputs—to capture the velocity of a client value, such as profit or propensity to buy. For example, a commonly derived ratio is one of debt-to-income: rather than using simply the debt and income attributes as inputs, more can be gained by the ratio rather than the individual values. In your web analysis, the number of site visits or the number of purchases made over time may provide a better insight into the true value of website customers:

Number of purchases/Number of visits:	7/9 = .77 Propensity to Purchase Ratio
Total amount of sales/Total number of visits:	$39/5 = 7.8 Profit Ratio

Format the Solution

In conjunction with the evaluation of your web data and your business objective is the issue of the format your business solution is going to take.

- What is the desired format of your solution, decision tree, rules, C code, graph, map?
- What is the goal of the solution, classification, regression, clustering, segmentation?
- How will you distribute the knowledge gained by the data mining process?
- What are the available format options from the data mining process?
- What does management really need, insight or sales?
- What do you need from the data mining process?

Multiple tools may be required to come up with the ideal data mining form for your website. For example, you may need to extract

Amazon/AlmondNet
Exhibit 1011
Page 150

rules from a clustering analysis. To do so you will need to first perform the clustering analysis using a Self-Organization Map or Kohonen Network. Next you will run the identified clusters through a machine-learning algorithm in order to generate the descriptive IF/THEN rules that "profile" the extracted clusters. Conversely, you may have to first do an analysis using a machine-learning algorithm on a data set with a large number of attributes in order to compress it and/or to identify a few significant attributes, and then run those significant attributes through a neural network for the final classification model.

Tradeoffs may be in order, but you must weigh your options with respect to what your business and website needs are and what tools you will need in your analysis. What are you after, insight or results? Often the format of your data mining solution will determine what data mining tool you will use. If you need to explain how and why you uncovered a pattern in your web data, you may need to use a machine-learning algorithm such as a decision tree or rule generator data mining tool. If all that matters is accuracy and efficiency, a neural network tool will do. For an e-commerce site, most likely both paradigms will be advantageous, since knowing the demographics of customers and increased sales are both desired goals.

Select the Tools

The selection of the right tool is clearly dependent on the task you are trying to accomplish. For example, the following matrix describes what each data mining technology is best suited for:

Tool Type	Classification	Clustering	Description
Genetic Algorithms	Yes	No	No
Machine Learning	Yes	Yes	Yes
Neural Networks	Yes	Yes	No
Self-Organizing Maps	No	Yes	No

Along with selecting the right technology, the characteristics and structure of your data must also be considered when selecting the right tool for the job. Here is a checklist of data-related issues you need to consider when selecting a data mining tool:

Amazon/AlmondNet
Exhibit 1011
Page 151

- number of continuous value fields
- number of dependent variables
- number of categorical fields
- length and type of records
- skewness of the data set

As a rule, machine-learning algorithms perform better on skewed data sets with a high number of categorical attributes and with a high number of fields per records. Neural networks, on the other hand, do better with numeric data. The following are a dozen criteria you need to consider in the selection of the right data mining tool. They involve software issues and hardware requirements:

Scalability

As a website and its log files and forms database get larger, a tool's performance should improve accordingly. Scalability means that by taking advantage of parallel database management systems and additional CPUs, the user is able to work with more data, build more models, and improve overall accuracy by adding additional processors. As the data increases so does the complexity, not only in the number of records but also in the number of attributes, variables, and possible website patterns. If a platform lacks the ability to scale, the computationally intensive nature of data mining will slow and eventually kill the decision support system.

The number of interactions among variables and amount of nonlinearity of parameters also contribute to the scalability of complexity. As the patterns become more subtle among the noise, the need for accuracy rises. At this juncture it is important to evaluate what kind of parallelism the tool supports. Find out if the tool supports a symmetric multiprocessing system (SMP) or massively parallel processing system (MPP). Technical factors affecting scalability include database size, model complexity, performance monitoring and tuning, as well as effective model validation. The issue of scalability in a data mining tool is how well it takes advantage of hardware design, including parallel algorithms and their direct access to parallel DBMSs. A high-end data mining tool needs to be able to run on a scalable hardware platform.

Accuracy

Accuracy is measured in the error rate of the algorithm or network for predictive modeling. There are several ways to evaluation accu-

Amazon/AlmondNet
Exhibit 1011
Page 152

racy. It can be measured as the *degree* of error, that is, determining *how* wrong the incorrect answers were. For some applications, it is of interest to know if the system responded with "no answers" or "wrong answers," and for what values of the condition attributes this happened. The structure of the data set being used to test it must be appropriately noted. See R. D. King, C. Feng, and A. Sutherland, "STATLOG: Comparison of Classification Algorithms on Large Real-World Problems," *Applied Artificial Intelligence* 9(3):289–333, May-June 1995.

- How does the algorithm of the tool determine if the characteristics of the problem match?
- What provisions are made for handling missing values, noise, and cost computations?
- How fast does it train and how does it work on the new data?
- How does it treat the dependent and independent variables?
- How much noise can it handle before accuracy drops?
- How sensitive is the algorithm to noise?
- How sensitive is it to missing data?

Formats

A data mining tool incorporating a machine-learning algorithm should provide a feature for viewing the most important rules or factors it discovers. The tool should not only be effective in detecting patterns, it should also provide you an insight into what is happening in your website and allow for easy and quick interaction with the data. The simplest explanation to the activity in your website is the preferable solution in discovering the nature of the customer transactions under analysis. A data mining tool should generate its results—trees, formulas, weights, graphs, or rules—as simply as possible in order to uncover as many unseen objects as possible. For a data set of some size, such as one from a large web retailing site, reading through all the rules will be an impossible mission for a user, even if each rule is simple to understand. For this reason a tool should generalize a simple solution while at the same time incorporate features to avoid overfitting. That is to say, general rules should describe what activity is taking place at a high level at your website.

You must ask if your tool incorporates a feature for prioritizing its results and, if it does, how it accomplishes this: what factors does it use, what significance level does it use, what is the percentage of its probability and accuracy, how many numbers of cases does it incorporate in its rule, and so forth. Does the tool offer

Amazon/AlmondNet
Exhibit 1011
Page 153

some form of visualization aid, such as histogram of errors by business graphs, or can its market segments be mapped to geographic maps? Is the tool designed to work on a desktop or in a client/server environment? Size is not a true measurement of the complexity it can handle—quite simply the only thing a tool's architecture may indicate is the amount of data it can process and not the complexity it can mine.

Solutions

A data mining tool should help you understand the results of its analysis by providing simple yet exact measurements, such as predicted error rates, level of significance, and/or rate of accuracy. What kind of output (solutions) does the tool generate: decision tree, C code, IF/THEN rules, narrative reports, graphs, visualization, SQL syntax? Some tools produce results that are relatively easy to interpret, such as decision trees that can be expressed as rules. Other tools, such as a back-propagation neural network, may make good predictions but may be difficult to understand. Ideally, both understanding and accuracy should be balanced in the format of the tool's solution.

The tool should also incorporate adequate instruments to allow the user to perform a sensitivity analysis on its output results. Does the tool include some interface for exporting the results of its models to other visualization or OLAP tools? What kind of GUI is provided for model building? Does the tool provide an API, which can be used embedded in a production system, in C, Visual Basic, or PowerBuilder?

Preprocessing

Preparing the data for analysis is one of the most time-consuming aspects of data mining. Your web server log files generally are comma-delimited, so the tool you select should be able to read this format without difficulty. Your forms database is most likely housed in some relational database, which again your tool should be able to connect to. Any special features a tool can perform to ease this process will greatly expedite the development of knowledge extraction and the construction of your predictive models. Some preprocessing functions to consider in evaluating a tool should include its ability to do some of the following:

- data cleaning
- data selection
- data description
- data transformation

Amazon/AlmondNet
Exhibit 1011
Page 154

A key feature of a data mining tool is its ability to cope or handle the data "as-is" in order to do some autonomous data cleaning, such as dealing with missing values, by giving you the options to "discard," "average," "alert," "exclude," etc. The tool you select should be able to provide some descriptive report on the quality of the data to the user.

- How well does it describe the data, row and value counts and distribution of values?
- Can it "sense" the difference between continuous value and categorical fields?
- Can the tool recognize and deal accordingly with integers, real numbers, etc.?
- How does the tool handle missing data or identify integrity violations?
- Can the tool explode categorical variables into dichotomous variables?
- Can the tool "map" fields or change them on the fly?
- Can the tool group continuous variables into ranges?
- Can the tool do calculations of existing columns?
- Does the tool require extensive data preparation?

Connectivity

Some data mining tools require that the data being mined be extracted from the target databases into their internal file format, whereas others can work directly against the database. A scalable data mining tool that can directly access your web server using its native SQL can maximize performance and take advantage of special performance features, such as parallel database access. Consideration should be given to those tools that can optimize performance by allowing the data mining processing at the server level. Although you likely will want to mine your data on a dedicated server set aside exclusively for data analysis.

- Can the tool consolidate data from multiple sources for individual analysis?
- If the tool is currently ODBC compliant, what drivers does it support?
- Does the tool incorporate an HTML or a JAVA interface?

Import/Export

Data mining tools that provide an easy integration path with other products offer additional ways to get the most out of the total data

Amazon/AlmondNet
Exhibit 1011
Page 155

mining and knowledge discovery process. Related to connectivity is the feature of importing data: does the tool support multiple formats such as ASCII, Access, Excel, comma or tab delimited, SAS, SPSS, and other specific DBMS, etc.? What conversion does the tool make with the original data it imports and at what ratio? Does the tool allow for the exporting of code, syntax, rules, etc.? Many database products (including traditional query, reporting, graphics, and visualization tools) can assist in the understanding of the data before and after the data mining process. The tool should provide the capability to easily link its results to an exportable format that can be visually enhanced for a presentation to management or others.

Memory Management

Normally a data mining tool's memory requirement will only be linear, depending on the size of the data set. The time complexity of the learning phase is a very limiting factor for many of today's data mining tools. If the algorithm used, for instance, uses exponential time growth, the maximum size of the training set will be quite limited. When considering the memory usage of a data mining tool, only the complexity matters, as for time usage. Still, the memory usage could give an indication as to what kind of system is necessary in order to handle "normal" amounts of data, in terms of number of records and rows.

On the other hand, whether a complex system uses one or two hours is usually of less importance. For a tool using iteration to achieve better and better accuracy, the time taken into evaluation must be the time to reach a certain level of accuracy. In the evaluation of a tool, consideration should be given to this time/complexity factor. Know in advance that certain tools, such as those based on an SOM network or a genetic algorithm, operate on a data set in such a way that they are very computationally intensive—meaning results may not be available for several hours.

Performance

Speed and accuracy both contribute to the evaluation of a data mining tool's overall performance. Speed is measured by how fast a model is built, as well as how fast a deployed predictive model can evaluate new data. Given the tool's design network or algorithm, how does it process the data, via single or multiple passes? Another factor impacting performance is cost—that is, the cost of providing a learning data set in the development of a model. This cost includes the number of examples necessary and the cost of assuring a needed accuracy in the learning set of a model. In most

Amazon/AlmondNet
Exhibit 1011
Page 156

instances the sheer volume of transactions for a large website dictates that this will not be a factor or concern. The cost factor should be considered in the evaluation of a tool's overall performance. Somewhat related to this criteria is the issue of data limitations, that is, whether the tool restricts the type of variables it is capable of processing.

- What are the limits to the number of ranges or intervals it can group?
- Can the tool assist in the understanding of the results of the analysis?
- Does the tool require numeric or discrete data fields?
- Can the tool present results in a graphical format?
- How many categories can the tool process?

Noise

Tool accuracy is often affected by noise, which is the result of irrelevant columns, missing or incorrect values, or cases that don't conform to an underlying pattern. In evaluating the robustness of a data mining product, you must answer the question of how much of this noise the tool can stand before its accuracy drops. Instead of evaluating a data mining tool based on a bell-shaped training set, one should also measure a system's noise-handling capabilities. Different levels of noise should be added to the training set, and the number of correct classifications with the new rules should be noted. If the noise added to the training set created inconsistencies, the resulting classification errors should be studied carefully, as some systems, for instance, will create conflicting rules with no hints to which one to choose in a particular situation.

Paradigms

A key criterion for a data mining tool is the data mining technologies it supports. For example, can the results of analyses be linked or passed from one classification system to another, in order to develop a hybrid solution? An example of this is being able to do clustering using an SOM (Kohonen Network) and then generating rules from a sub-cluster using a machine-learning algorithm (C5.0)—all within one data mining suite is an ideal feature. In the analysis of the data mining tool, specific references on the type of technologies should be noted in order to optimize its results.

- What are the underlying technologies and techniques of the tool? Are they proprietary or generic? And most importantly, are they linked?

Amazon/AlmondNet
Exhibit 1011
Page 157

- Can the tool split the data into training and testing sets? Can the user select the ratio of this split?
- Is the analysis entirely automated and closed, or does it allow for some degree of customization?
- Does the data mining incorporate more than one classification algorithm or network?
- Does the interface allow for the close evaluation of results of the analysis?
- How intuitive is the tool? Does it require extensive training?
- Is there an application program interface (API)?
- What is the current version of the tool?

Efficiency

Finally, consider the criteria of the data mining tool's efficiency. Will it take much time to take a decision based on the data mining process? For example, in developing a set of weights or rules, what matters is how we find the formula or rule to use. If the rules are sorted somehow, it may be possible to use indexing or search methods to find the most general and robust rule quickly. Otherwise, the complexity of taking the decision could be higher than if, for instance, all rules must be compared pair-by-pair. Some tools offer solutions in the form of hundreds of rules, which defeats the purpose of data mining—uncovering a simple pattern from a massive amount of customer online transaction data. For example in a decision tree, the complexity of decision-making is usually some function of its width and height. In the evaluation of a data mining tool, how it handles this issue of complexity and efficiency needs to be addressed and considered.

Tool Selection

Other nontechnical issues are of course platform and price, as well as the availability of training, consulting, and support. The range in prices for data mining tools vary greatly, starting from a $5,000 desktop single technology system to a multiple paradigm suite with each module costing in excess of $135,000. Keep in mind, however, that both of these systems may be based on the same core technology, such as a variation of the CART algorithm or a back-propagation neural network. The difference in price is really dependent on the platform each of the systems is designed for deployment on, as well as the scope and size of the web data set each system is required to mine. Some data mining companies are new and small and may not be around for long. What is important to note, however, is the technology that their tools

Amazon/AlmondNet
Exhibit 1011
Page 158

are based on, such as neural networks and machine-learning and genetic algorithms, is a mature and robust technology that will survive over time.

Lastly, you need to step back and consider outsourcing rather than purchasing a data mining tool and the related training it may require of your staff. As with other information system projects, outsourcing or consulting is always an option. This may be more cost effective over time; after all, you are not buying software as much as buying

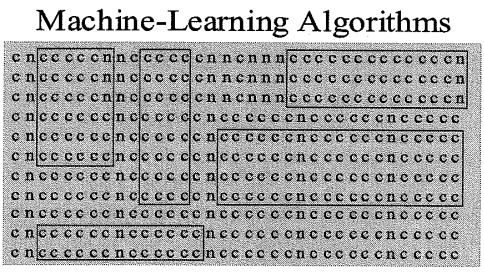


Figure 4–6 *How machine-learning algorithm tools identify customer clusters.*

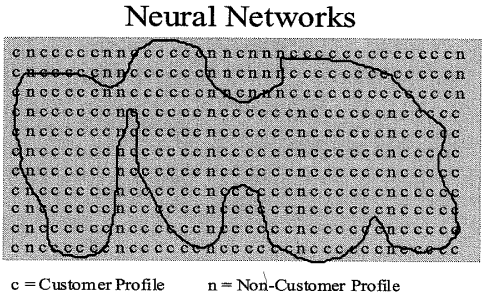


Figure 4–7 *How neural network tools identify customer clusters.*

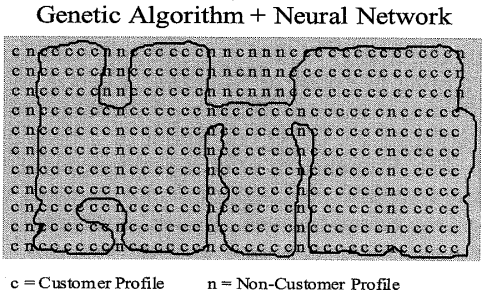


Figure 4–8 *How neural network tools, optimized with a genetic algorithm tool, identifies customer clusters.*

Amazon/AlmondNet
Exhibit 1011
Page 159

the knowledge and insight it provides from your transactional data. However, as previously stated, data mining is an iterative process, which means that you may be churning data on a fairly regular basis from your website, mandating that you incorporate data mining as part of your everyday operation.

If your data is highly skewed, you may want to use a machine-learning algorithm tool, which can locate profitable customer clusters distinctly independent of one another (Figure 4–6). If your data is nonlinear, you may want to use a back-propagation neural network tool, which can map a data space more efficiently than linear regression (Figure 4–7). You may want to optimize your neural network tool with a genetic algorithm tool (Figure 4–8).

A Tool Shopping Tip: Kick the Tires before Buying

You may want to start by browsing the Web for samples of evaluation copies of data mining tools. A key site is www.kdnuggets.com, a data mining directory of companies, tools, and reference materials. Keep in mind the platform requirements in the selection of tools and the various features each provides. Keep in mind also that you may require more than one tool—consider the benefits of a data mining suite which incorporates multiple paradigms. Finally, ask for a Proof-of-Concept evaluation copy of the software. For example, some vendors will provide special pricing for a 90-day "try-out" deal of their tool.

Construct the Models

It is not until this stage that the actual data mining of your website files begin. This is the process of searching for patterns in a data set and the generation of classification rules, decision trees, clustering, scores, weights, and evaluation and comparison of error rates.

- Is additional data available which could help the performance of the models?
- What are the model error rates? Are they acceptable? Can they be improved?
- Is a different methodology necessary to improve model performance?
- How many models do you require for your entire website?
- Train and test models using a random number seed?
- Output SQL syntax for distribution to end-users?
- Supervised learning or unsupervised learning?

Amazon/AlmondNet
Exhibit 1011
Page 160

- Incorporate C code into a production system?
- Integrate rules in a decision support system?
- Purge noisy and redundant data attributes?
- Classification, prediction, or clustering?
- Monitor and evaluate results?

Will you be developing a model via supervised learning? This is a situation where you have samples of both negative and positive cases: website buyers vs. website non-buyers, web visitor respondents vs. web visitor nonrespondents, etc. In most instances, especially for a large electronic retailing site, you will have thousands of samples from which you can develop your models, in which case you will be performing supervised learning for classification or prediction. However, if you have multiple products and services, most likely you will need to construct separate models for each. For example, the propensity to buy one category of products, say BeatlesCD.htm, will be different from StonesCD.htm or PoliceCD.htm. Furthermore the propensity for each product line will also require a specific model, so that a model for AbbeyRoadCD.htm will be different from WhiteAlbumCD.htm. To what extent you want to refine your modeling efforts depends on your resources and time.

If you don't have a sample of both positive and negative cases, you will most likely be performing clustering, or unsupervised learning. This process is much more difficult to do than classification, but it may be necessary. Clustering is often done to discover consumer patterns. Retailers often do a type of data mining clustering commonly known as "market basket" analysis to better position certain products that tend to sell in unison or to offer special incentives. For example, a clustering analysis might discover that consumers who buy premium wine tend to also buy deli cheese, while those who buy microbrewery beer tend to instead purchase deli meats.

Similar purchasing patterns can be uncovered in a commercial website, for example, where certain products tend to be bought at the same time. Sequencing may be involved between a point in time when one product is purchased, such as an ink-jet printer this month and ink-jet cartridges 30 or 60 days later. An exhaustive amount of purchasing associations and patterns exist in large websites such as one selling relatively inexpensive consumer items like books, CDs, toys, and software. Visual associations can be explored, which can lead to cross-selling opportunities.

Both supervised and unsupervised learning require that you split your data into training and testing sets. This can be in a 90/10, 70/30,

Amazon/AlmondNet
Exhibit 1011
Page 161

or 60/40 ratio. The splitting of the original data must be done using a random seed number; almost all modern data mining tools incorporate a feature for splitting your data and evaluating the results of the models you create. Once the training set has been used to construct a model, the test set is run through the data mining model in order to evaluate its accuracy. Dividing the number of incorrect classifications by the total number of instances produces the model's error rate. Dividing the number of correct classifications by the total number of instances gives the model's accuracy rate.

Almost all of the current data mining tools provide a method by which you can measure the accuracy of the models you can construct with them. They facilitate the process by which you can split your data into training and testing data sets. They each also provide methods by which you can randomly test and compare the results of your training and testing data results. Decision tree tools, for example, provide detailed output designed to show the differences between the current tree and the alternate data set. Some tools allow for the "resubstitution" of data so you can test on the same data that you trained your model on. However, it is always best to test with an alternate "unseen" data set to perform the test. This alternate data set can come from a hold back sample created by partitioning your data or from an imported, external file.

Validate the Findings

As previously mentioned, a data mining analysis of your website will most likely involve individuals from several departments, such as Information Systems, Marketing, Sales, Inventory, etc. It most definitely will involve the administrators, designers, analysts, managers, and engineers responsible for designing and maintaining the day-to-day operations of your website. It is important after you have completed your data mining analysis that you share and discuss your findings with all of them. Domain experts (people who are the specialists in their area) need to be briefed on the results of the analysis to ensure the findings are correct and appropriate to your site's business objectives. This is the sanity check: You need to be objective and focused on your initial goal for mining your website. If your data mining results are faulty—whether it is due to the data, tool, or methodology—you may need to do another analysis and reconstruct a new set of models with their participation and input.

What important relationships were discovered between the independent and dependent variables? For example, do the demographics you discovered about your website visitors and customers match

Amazon/AlmondNet
Exhibit 1011
Page 162

those of prior marketing analyses? A website selling game software will likely have totally different customers from those of an investment brokerage site. Keep in mind too some of the inherent demographics about who uses the Web in general, for this may skew your perception of who your website customers are. People on the Web tend to be better educated and a bit more affluent than average, but this is also slowly changing every month as more and more people get "wired."

Another side issue at this time may be presenting the results to management and the various departments in your company, including your web administrators and designers. It is important that they understand the implications of the data mining analysis. In other words, now that you have identified unique consumer patterns or identified a specific web consumer profile—how are you going to validate it? The validation of the findings often involves testing a model with "real bullets" in field situations. For example, a field test may involve making a special offer on your site, or doing a special e-mailing campaign. You will need to do a test to measure how well your data mining findings or models improve your business bottom line, like response rates and/or sales figures. To minimize risk you may want to do a control test to measure how well your data mining models perform and how well they improve existing baselines prior to full implementation involving all product lines.

Deliver the Findings

A data mining report should be prepared documenting the entire data mining process, including the steps you took in selecting and preparing your data, the tools you used and why, the tool settings, your findings, and an explanation of what the code that was generated is supposed to do. As with any business process, you need to establish both baselines and procedures. In your analysis report you need to comment on the results of the data mining analysis, stating whether it meets the business objective of your website. If for some reasons it doesn't, you should state why not. You may want to include in your report how the data mining analysis results can be improved. Is different or new data warranted? Perhaps external demographic and household information should be brought in, or you could capture better information via newly designed registration forms or cookies in your website.

You should include some recommendation for action stemming from the data mining analysis. For example, if you found that most of your website customers are young professionals, you may want to suggest some subtle changes to the content and design of your site. Data

Amazon/AlmondNet
Exhibit 1011
Page 163

mining is an iterative process, and its documentation will ensure that improvement in the effectiveness of your website, especially if you are involved in electronic commerce, will continue over time. These incremental improvements will occur not only in how data mining takes place in your website and your company but also in the accuracy of the models and in the improvement of your online sales and marketing.

Integrate the Solutions

This final step in the data mining of your website data project is really a commitment to continue the process of learning from your firm's online transactions. This process involves incorporating the findings into your firm's business practices, marketing efforts, and strategic planning. Data mining is a pattern-recognition process involving hundreds, thousands, or maybe millions of daily transactions in your website.

Integrating Your Website and Data Warehouse

The integration of your website data with your company information file or data warehouse can assist your company in knowing how your current customers match in characteristics with those of your website visitors. As with the mining of your data warehouse, the integration of your website data will further enrich your insight about the attributes of your current and potential customers.

Through the integration of your website with your internal production company databases, you can provide website visitors timely information about your inventory, services, and customer support. While doing this you can also begin to learn who they are and what they like and dislike about your products and services. By mining the data collected about your website visitors you can begin to serve them with the type of information their demographics indicate they want to possess.

Your company can communicate to its customers in a highly personalized tone and manner. By integrating web purchasing information with your data warehouse, you can acknowledge purchases, remind customers about similar products, and cross-sell related or complementary services. Several measurements can be mined, including the date of the most recent purchases, the dates of those purchases, and the total amount of money each customer has spent at your site. The mining of these measurements can be used to calibrate customer value, profitability, and loyalty.

Through the merger of your data warehouse and website, several deliverables are possible, including the identification of your most

Amazon/AlmondNet
Exhibit 1011
Page 164

profitable customers and potential future clients. With this profile you can move to target and develop new potential clients while they visit your website. Comparisons of several attributes can be made between your existing customers and those of your site visitors in order to fine tune or alter your ads, offers, incentives, pricing, messages, banners, and other communications. Through this matching you can also make some informed decisions about marketing campaigns and the overall design of your website. For example, you can investigate how your current customers compare to your web visitors in terms of the following attributes:

- Age
- Gender
- Income
- Hobbies
- Occupation
- Marital Status
- Home Location
- Education Level
- Recreational Interest
- Disposable Income Ratio
- Psychographics and Demographics

Psychographics are measurements of individual's attitudes and lifestyle characteristics. Such measurements are excellent indicators on how website visitors and customers are thinking and where they are going with their lives. SRI International, an independent non-profit research institute, has developed eight Values and Lifestyles (VALS) categories for consumer types:

- Actualizers
- Fullfilleds
- Achievers
- Experiencers
- Believers
- Strivers
- Makers
- Strugglers

VALS categorizes U.S. adult consumers into mutually exclusive groups based on their psychology and several key demographics.

Amazon/AlmondNet
Exhibit 1011
Page 165

VALS is unique because it highlights factors that motivate consumer-buying behavior. VALS looks at the underlying psychological make-up of people. It looks at the human side of the equation, for example, preferences for such values as control or freedom, tradition or novelty, information or stimulation, hands-on activity, or intellectual abstractions. One of the SRI's recent developments is iVALS, which is a profiling system designed to enhance the usefulness and quality of online environments for end users, content providers, and intermediaries. iVALS profiles web users and clusters them along the following categories:

- Wizards
- Pioneers
- Upstreamers
- Socialites
- Workers
- Surfers
- Mainstreamers
- Sociables
- Seekers
- Immigrants

This is an example of an iVALS profile:

The iVALS Wizards Profile

Wizards are the most active and skilled Internet users. Computers are a key aspect of their lifestyles, and mastery of technology figures prominently in their identities.

More than 80% of Wizards have been on the Internet for three or more years. Although Wizards report meeting many new people during their years online, it is likely that such friendships often stay virtual. That is, Wizards are not necessarily seeking to meet people in the traditional sense. Rather, their sociability appears as a byproduct to their heavy involvement in so many aspects of Net culture.

Nearly everyone in the Wizard group has performed relatively esoteric functions such as videoconferencing on the Internet or creating Web pages. Half have participated in a MUD or MOO. Most would call themselves computer power users. Many have used WWW retail sites, and all own a lot of games or multimedia, especially for adults.

Amazon/AlmondNet
Exhibit 1011
Page 166

Wizards are nearly all males, and relatively young, with a median age under 30. Despite their age, many work as computer technicians, professors, middle managers, consultants, or industry analysts, and therefore earn a medium to high household income.

Wizards are a prime target for sophisticated technical information, beta test software, authoring tools, computer and software industry conferences and trade shows, and other computer-related professional or product communications. They like software with lots of options and enjoy climbing a learning curve if it leads to interesting new abilities.

Through the data mining of your website a new understanding of your visitors and customers will emerge, especially when merged with psychographic profiles like iVALS. Through the mining of this data, a profile or several composites can emerge which will assist you in tailoring their experience during their time in your site to more closely meet their needs.

This final step of your data mining analysis requires that you monitor the performance of the models that you have generated. All models will age and their performance will disintegrate, so you need to ensure you monitor the accuracy of your data mining models.

The data mining process can involve some if not all of the above steps, and in many cases they may not be sequential. There will be instances where certain steps can be entirely skipped. However, the basic idea in the process is to ensure that the data mining is heavily dependent on recognizing the business objectives of your website and the limitations of the data and the tools.

The integration of the data mining solutions usually means taking action. A data mining analysis can lead to making some organizational changes on how your website resources are deployed, or what recommendations are made to management and what alliances and partnerships you may want to consider. The integration of a data mining analysis may impact how your online inventory is arranged, or how your website promotions are planned, or which website visitors and customers get what offers. It may involve a closely integrated e-mail system incorporating personalized notification of products and services based on your data mining analyses. This integration may involve how accounts are scored, graded, or grouped. The integration of a data mining solution may be done on a case-by-case basis or on a batch process, so that, for example, you profile your entire website customer account file on a weekly, monthly, or quarterly basis.

Amazon/AlmondNet
Exhibit 1011
Page 167

The data mining effort that you conduct will likely include a lingering question: *"now what?"* Are you prepared to use the knowledge of your data mining analysis to change your business and your e-commerce site? Are you prepared to make major marketing and business process changes? Are you prepared and do you have the support of management to re-design your website should your analysis support this action? All or some of these questions are likely to result from the mining of your website and you should plan to deal with them accordingly.

Data mining is a process, not a project. As such you need to consider who will be doing it, when is it required, and what data will you be mining. Data is not static. As such you need to plan on conducting periodic analysis of it, at what intervals depends on how often and dramatically it changes. For example, a bank may need to mine its data warehouse every quarter for its customer acquisition models. A cellular telephone company may want to mine its warehouse every month in order to keep its churn model fresh and accurate. An electronic retailer may want to mine its website every week—why? Because the business environment each of these entities is attempting to anticipate and model changes at very different rates: customers do not change checking accounts as often as they do cellular carriers or online booksellers.

One way of determining how often you need to mine your data is to periodically check the accuracy of your models against current customer data. If the accuracy has deteriorated significantly since the creation of the model, it is time to build a new one. Knowing all of these lingering questions before you start will hopefully lead to a smoother incorporation of data mining into your business and the optimization of your website. Just remember that data mining will definitely benefit any firm with an e-commerce site. A standard process model is being developed by several interest groups known as the CRISP-DM (Cross-Industry Standard Process for Data Mining), which is aimed at moving away from the focus on the technology and instead concentrates on the business problems: http://www.ncr.dk/CRISP/.

Amazon/AlmondNet
Exhibit 1011
Page 168

5

The Tools

In this chapter we will discuss several data mining tools. Data mining tools are commercial programs incorporating an AI technology for visualization, association, segmentation, classification, prediction, clustering, optimization, and modeling. Some tools incorporate multiple technologies, such as neural networks and machine-learning algorithms within a suite, with the capabilities of linking their results.

Criteria

The tools selected represent a cross section of commercial products based on different technologies, each formatting business solutions in a unique manner. This side-by-side discussion will demonstrate how two tools using a variation of the same technology, such as a neural network, can format a solution in a dramatically different manner. In a relatively young industry such as data mining, there are currently no standards. As such the results of their processes can at times be quite different.

All of the software products are designed to run on desktop environments or high-end servers. They vary in price from a couple of thousand to well over a hundred thousand dollars. There are other tools on the market that were not included in this review but are equally good at data mining and modeling.

This sampling will demonstrate the wide range of features, performance, and approaches in today's data mining software vendor market. Our purpose is also to demonstrate that the main criterion in your tool selection should be based on how well it answers your website business needs. Keep in mind the dozen criteria from the prior chapter on tool selection:

Amazon/AlmondNet
Exhibit 1011
Page 170

1. Scalability
2. Accuracy
3. Formats
4. Solutions
5. Preprocessing
6. Connectivity
7. Import/Export
8. Memory Management
9. Performance
10. Noise
11. Paradigms
12. Efficiency

Simply having the "right" technology in a data mining tool is not enough. Having a tool with multiple paradigms has its obvious advantages. For example, you can "link" the results from one to another, enabling you to benefit from two technologies rather than a stand-alone solution. For example, Clementine from SPSS enables you to do cluster analysis using a Kohonen network, then link the segments to a C5.0 machine-learning algorithm for the generation of rules describing the attributes of those clusters. Another tool, Decisionhouse, enables you to use a machine-learning algorithm to segment your data set in a decision tree, and then view the market segment it discovered via a geographic map, for the entire United States down to the ZIP code level. The point is that multiple paradigm tools provide more *useable* business results, which are greater than their single technology parts.

Some large end data mining tool vendors are able to ask and get $150,000 for their base package, because they are able to deliver technologies in a manner a company can incorporate into their business processes and systems. NeoVista, a high-end data mining vendor, will virtually customize their software to meet their customers' needs, most of which are Fortune 100 firms. Simply incorporating an algorithm in a package does not mean it is a very effective business tool. In order for a tool to truly be effective it must be scalable and robust software able to handle large data sets in a client server environment and deliver business solutions in a format that can be deployed over the enterprise. This includes the tools you use for analyzing your website data.

You should select a tool that is going to provide you the business solution you require in leveraging your website to its full marketing

Amazon/AlmondNet
Exhibit 1011
Page 171

and sales potential, and provide you the insight you need about your online customers and the transactions they generate. The tools we look at in this section range from desktop systems to large multitasking, multi-parallel server powerhouses, which can benefit websites of all sizes, from single product/single person websites to major portals of electronic commerce.

To demonstrate the different approaches of some of these tools, we performed multiple analyses on a single data set containing a small random sample of 1,472 records, the output from a small electronic commerce web server for a single day. From this sample data set we attempted to identify specific website patterns leading to online sales and to identify some of the demographics of these visitors and customers. There were three possible outcomes to the sessions, which were identified by specific pages on the site:

JustBrowsing.htm	*This results in a no sale visit.*
E-mailMe.htm	*This is a potential lead to a sale.*
OrderForm.htm	*This is a sale, and the desired outcome*

The primary objective of the analysis then is to identify the characteristics of those website visitors who selected the OrderForm.htm. To do this we linked the forms database, which contained the visitor's ZIP code, and the server log file to a demographic data set from CACI known as the ACORN database.

The ACORN Consumer Classification System provides a percentage of distribution by consumer type for all the ZIP codes in the United States. ACORN analyzes demographic profiles and classifies consumers according to the type of residential area in which they live. ACORN is a database designed to identify consumer profiles on 226,000 neighborhoods sorted by 60 lifestyle characteristics, such as income, home value, occupation, education, age, and other key determinants of consumer behavior. The ACORN is a neighborhood segmentation system containing 43 clusters, or consumer segments, which can be used to determine the lifestyle of customers on the basis of their ZIP code. For example, the ZIP code 94502 is categorized by the ACORN system as consisting of the following consumer segments:

37.4% Wealthy Seaboard Suburbs—This type is located along the eastern seaboard and in California. Residents are married, middle-aged professionals at the peak of their lifetime earnings. They rank among the highest for auto club memberships and travel

Amazon/AlmondNet
Exhibit 1011
Page 172

extensively. Media preferences include Metropolitan, Boating and Yachting, and the *New York Times*. This market's share of disposable income is relatively high.

62. 6% *Successful Suburbanites*—Members of this type have achieved their success with families, home ownership, and prosperity by dint of hard work. Upward mobility characterizes this market. Essentially suburban, they drive new cars and spend money on sports, exercise equipment, and personal computers. They are likely to have loans. They do not watch a lot of TV, but read daily newspapers.

Keep in mind that other types of visitor attributes are available for this type of inductive analysis both from within your website's log files and forms databases and from external data providers such as Equifax, MetroMail, LikeMinds, NetPerceptions, and NetCount, which can further describe the psychographics and household features of website visitors, such as their income ranges or the presence of children. (See Figure 5–1.)

Single Record Sample by Field Name

Top One Percent ZIP %, Wealthy Seaboard Suburbs ZIP %, Upper Income Empty Nesters ZIP %, Successful Suburbanites ZIP %, Prosperous Baby Boomers ZIP %, Semirural Lifestyle ZIP %, Urban Professional Couple ZIP %, Baby Boomers with Children ZIP %, Thriving Immigrants ZIP %, Pacific Heights ZIP %, Older Married Couples ZIP %, High Rise Renters ZIP %, Enterprising Young Singles ZIP %, Retirement Communities ZIP %, Active Senior Singles ZIP %, Prosperous Older Couples ZIP %, Wealthiest Seniors ZIP %, Rural Resort Dwellers ZIP %, Senior Sun Seekers ZIP %, Twentysomethings

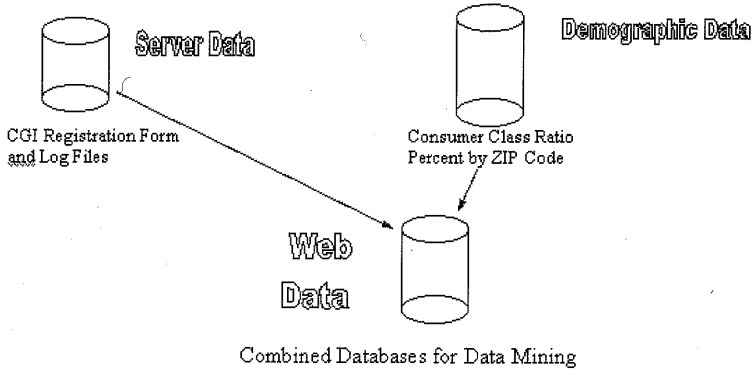


Figure 5–1
The merger of server and demographic data for mining.

Amazon/AlmondNet
Exhibit 1011
Page 173

ZIP %, College Campuses ZIP %, Military Proximity ZIP %, East Coast Immigrants ZIP %, Middle Class Families ZIP %, Newly Formed Households ZIP %, Settle Southwestern Families ZIP %, West Coast Immigrants ZIP %, Low Income: Young & Old ZIP %, Middle America ZIP %, Young Frequent Movers ZIP %, Rural Industrial Workers ZIP %, Prairie Farmers ZIP %, Small Town Working Families ZIP %, Rustbelt Neighborhoods ZIP %, Heartland Communities ZIP %, Young Immigrants ZIP %, SSN Dependents ZIP %, Distressed Neighborhoods ZIP %, Hard Times ZIP %, Urban Working Families ZIP %,**Subdomain,Income,ZIP,Gender,Product,HTM**

Single Record Sample Content

0,3,0,0,56,0,0,20,0,10,0,5,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, **"com", "10265", "10452", "Male", "Notebooks", "JustBrowsing.htm"**

Single Paradigm Tools

We begin by looking at single paradigm tools; these are data mining products incorporating a single technology, such as a neural network or a machine-learning algorithm like ID3.

Tool	Vendor	Platform	Technology
DM Marksman	HNC Software	NT with card	Neural Network
4Thought	Right Info	WIN95/NT	Neural Network
KnowledgeSEEKER	ANGOSS	WIN95/NT/Unix	CART, CHAID, ID3
NGO	BioComp	WIN95/NT	Neural Network, Genetic Algorithm
WizWhy	WizWhy	WIN95/NT	Proprietary Algorithm

Database Mining Marksman

Along with NeuralWare and Nestor, HNC is one of the pioneering neural network software companies to emerge in the mid-1980s. However, unlike the other two, HNC has been very successful and is now a public company. Its success, however, has not come from selling data mining software. In the 1990s the company changed its business strategy from selling tools to selling business solutions based on neural network technology, specifically in the area of credit card

Amazon/AlmondNet
Exhibit 1011
Page 174

fraud detection. HNC's flagship product, Falcon, currently monitors millions of credit card accounts, representing about 80 percent of cards issued in the United States, for fraud. Following in this vertical approach to providing solutions rather than software, HNC today sells ProfitMax, a credit card profitability management system. It has also recently spun off a company called Aptex, which is marketing products like SelectCast for Internet advertising and marketing using proprietary text recognition and neural network technologies.

Database Mining Marksman

Best at: Classification, direct marketing applications.

Cost Factor: No

Current Version: 1.1

Data Access: Can import ASCII and most file formats, also ODBC-compliant interface.

Data Preprocessing: Missing value substitution, conversion of symbol strings, symbol mapping.

Distinguishing Feature: Hardware processor allows for optimal performance of network.

File Limitations: Output must be binary categorical. Inputs: 250,000 rows X 100 columns. Limit of 15 to 20 symbols per categorical variables.

Format of Solution: The software generates reports on *lift*. It also shows which variables impact the most on output results. A sensitivity analysis feature allows for the prioritization of input drivers.

Installed Base: 25

Memory Required: 64 MB

Parallel Processing Support: No

Pricing: $48,750 software and PC with 16 node SNAP parallel processor with 3 days training.

Space Required: 3 MB

Training/Consulting: Yes, $7,500 per session plus travel and expenses.

Technology Paradigm: Back-propagation and Kohonen neural networks.

HNC also produces Database Mining Marksman, a product designed for direct mail marketers combining neural network software with a hardware accelerator card. Marksman contains three components: the Marksman GUI, a link to Microsoft's Excel spreadsheet, and DBMS/COPY. Excel is used for report generation and DBMS/COPY is used for data preparation, extraction, and importing. Marksman supports classification using the back-propagation neural network and clustering using the Kohonen network. Most of the reports generated by Marksman are geared toward direct marketing; hence they are designed to help set cutoff points for direct mail campaigns, based on desired response rate, and maximized return or bud-

Amazon/AlmondNet
Exhibit 1011
Page 175

get limits. The reports include Gains by Score, Gains by Percentile, Acquisition Cost, Lift Chart, and the Rollup Report.

A wizard feature in the software guides the user through each "project" which is the start to finish concept of a complete analysis. From the Analyze button in the main screen the user is provided the following options:

File Definition ... define the data to be mined
Data Preparation ... how is the data to be cleaned and mapped
Relationship Discovery ... what fields are to be used
Modeling ... set the parameter for the model
Analyze Data ... execute the modeling process

Importing data into Marksman involves the use of DBMS/COPY, a stand-alone software product, to export graphs and reports. Models can be binary where the dependent variable must be two-valued, or they can be categorical "Yes" or "No" and lastly continuous with value predictions. There is no way to export code from Marksman; for this you need a companion product HNC sells called DeployNet. This is a distinct disadvantage for a tool that is so expensive.

This data mining tool demonstrates some of the limitations of the breed. It is singular in its approach to a specific data mining task—direct marketing. It is also limited in the technology it incorporates: a neural network from which you cannot export the results of the analysis, that is, the weights in the form of C code. The other limitation of this tool is that it relies on another software product for the importing of data and the exporting of graphical solutions. Clearly it would be difficult to use this type of tool for the analysis of web-based data since its reports are narrowly aimed at the direct mail industry. When selecting your tools, beware of industry-specific applications with designed features that may limit their usefulness when mining website databases.

4Thought

4Thought is like another data mining tool from the United Kingdom (Profiler from Attar Software Ltd.), in that it requires a hardware security key that connects to the parallel port. Although this may be a standard in England where both of these products come from, these software keys can be a deterrent to using the programs. Although vendors' desires for protection against unauthorized duplication are justified, which the companies claim is much more common in Europe, these keys can be a significant problem. When you are working in a net-

Amazon/AlmondNet
Exhibit 1011
Page 176

4Thought

Best at: Time series modeling

Cost Factor: No

Current Version: 3.21

Data Access: Can import ASCII, ODBC support, Excel, Lotus, SPSS

Data Preprocessing: Outlier detection, interpolation, eliminates highly correlated columns, supports functions

Distinguishing Feature: Spreadsheet interface

File Limitations: Dependent variable must be numeric, limited to 256 categorical variables

Format of Solution: Statistical and reports

Installed Base: 500

Memory Required: 20 MB

Parallel Processing Support: No

Pricing: $9,900 plus 20% support service annually

Space Required: 3 MB

Training/Consulting: Consulting services $2,000 daily

Technology Paradigm: Back-propagation neural network

worked environment with both desktops and notebooks, you must disconnect and reconnect the software key to each machine as you do your analysis; crawling under desks to get to the port to get to the software key should not be a task of the data mining process.

4Thought has a somewhat simplistic interface. It presents the data in a spreadsheet format, which after a model is created adds a column with the predicted results. Changes to the data can be made in the spreadsheet environment, so that a graph can easily be created. 4Thought can import fixed and delimited ASCII, Excel, Lotus, and SPSS formats. It is also ODBC compliant and can link to Excel, Btrieve, dBase, and XDB. An ample number of functions are provided for data transformation, which is especially good since it is a neural network tool. It has a wizard-like feature called Autolog, which applies a logarithm to those columns that have a distribution that looks more like a log-normal distribution than a normal distribution. During the construction of a model, dynamic feedback is provided allowing for the user to see how the training and testing are progressing. Once the model is built, a predicted value column is created in the spreadsheet. If new input values are typed in a blank row at the end of the spreadsheet, a new predicted value is automatically computed. In addition, the neural network weights can be exported as a macro function to Excel, Lotus, and SPSS. Several graphs are available for pre- and post-modeling and the comparison between the inputs and the generated outputs.

Amazon/AlmondNet
Exhibit 1011
Page 177

4Thought is clearly aimed at the business analyst and the power spreadsheet user. It provides the functionality of a neural network in a spreadsheet format. In fact 4Thought is similar to other low-end data mining tools that work as add-ons to Excel. These tools, such as NeuralWare's Predict, Braincel, and Evolver use the Excel environment as an attempt to isolate the user from most of the complexities of data preparation and modeling. The strategy is a good one, for it takes advantage of the large installed base and the graphical capabilities of Excel. However, in choosing such a strategy they also limit themselves to the number of rows and cells Excel can import. The mining of a web database with 4Thought would necessitate the importing of it into Excel prior to analysis, which would be highly unlikely for a large e-commerce website with thousands of daily transactions.

KnowledgeSEEKER

KnowledgeSEEKER, which was originally introduced by FirstMark Software prior to its sale to AGNOSS, is one of the most mature data mining tools in today's market, and the current version is one of the industry's most flexible and powerful desktop decision-tree tool builders. Among KnowledgeSEEKER's strengths are its ease of use and its robust method of importing data from multiple data sources in different formats. KnowledgeSEEKER can import from multiple ASCII and DB formats and it is also ODBC compliant. Once the data has been imported into KnowledgeSEEKER, the user can view all

KnowledgeSEEKER

Best at: Ease of use

Cost Factor: Yes

Current Version: 5.0

Data Access: ASCII text fixed-length and delimited, dBase, Excel, Gauss, Lotus, Paradox, QuattroPro, SAS, Sawtooth, SmartWare, Splus, SPSS, Stata, Systat, and ODBC

Data Preprocessing: Emphasis on interaction with data

Distinguishing Feature: Very mature and robust tool

File Limitations: Scalable to hardware

Installed Base: 800

Memory Required: 640 K

Parallel Processing Support: No

Pricing: $5,000

Solution Format: Decision trees, IF/THEN rules, SQL syntax

Space Required: 4 MB

Training/Consulting: Yes

Technology Paradigm: CART, CHAID, and ID3

Amazon/AlmondNet
Exhibit 1011
Page 178

attributes, with the user being given the flexibility to do force splits on any variable in the data to create a decision tree.

The tool attempts to guess at the type of data imported, but the user can change this both during import or once loaded in KnowledgeSEEKER. Attributes can also be changed on the fly through what KnowledgeSEEKER calls "Mapping"—in this process several fields can be merged, renamed, or created. Once a decision tree is generated, the nodes created can be viewed textually and graphically. KnowledgeSEEKER allows for decision trees to be created via three possible ways:

Auto—A complete decision tree is generated until no interesting splits can be found.

Find Split—The user controls the depth and direction of the decision tree.

Force Split—The user forces the splits on a variable they select.

Multiple branch trees can be created with total flexibility, allowing the user to select specific nodes of interest and developing subbranches on the fly. In the decision tree in Figure 5–2, the data splits on the variable "Product" from which the user can further drill down. In this case we selected "Drives." We asked the software to search and create a decision tree, which in this example splits for "Rural Industrial Workers," an ACORN consumer class.

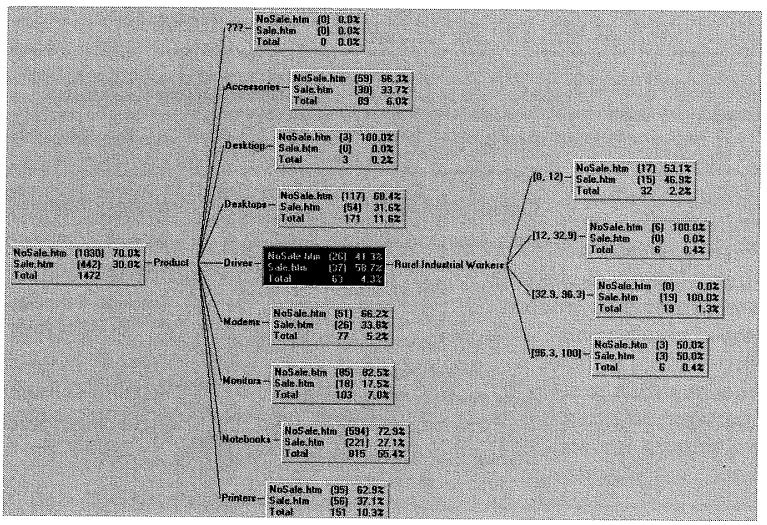


Figure 5–2 *KnowledgeSEEKER provides excellent control over the splits in the trees it generates.*

Amazon/AlmondNet
Exhibit 1011
Page 179

Once a decision tree is created there are two ways it can be validated. The simplest and fastest way is by going to the *Test* menu and selecting the *Resubstitution* option, which provides you both a Misclassification Rate and an Accuracy figure. The other method of evaluating a decision tree is by splitting your data set by whatever ratio you want, and then testing it:

```
KnowledgeSEEKER test results of partitioned data set.
Error Rate Profile :: Test Partition
          Learning    Testing        Dropped
 Category    Freq  %    Freq  %   Diff. Freq %
————————  —— ——  —— ——  —— | —
 NoSale.htm   776 70.3   254 69.0    -1.3 |0 0.0
  Sale.htm    328 29.7   114 31.0     1.3 |0 0.0
————————  —— ——  —— ——  —— | —
          1104        368           0
Misclassification Rate: 0.290761
Accuracy: 70.9239
```

Decision tree settings can be adjusted via the Tree Configuration menu. For example, the *Filter Threshold* (percent error) can be set to the following settings:

Decision	1%
Prediction	5%
Exploration	20%

The *Decision* support setting of 1 percent means that erroneous effects will be produced only 1 time in 100. It is a setting requiring an exceptionally high level of confidence in tree results. The *Prediction* is the default setting confirming that the displayed relationships are valid with a 95 percent certainty rating. This is the recommended setting for most data mining tasks by ANGOSS. The *Exploration* setting is the most relaxed and is designed to let you explore relationships in your data. It allows you to be wrong in about 20 times out of 100 when using this filter level. Lastly, KnowledgeSEEKER gives you the option to specify your own filter threshold for any value between 0 and 1. Other tree configuration options include *Split Search Algorithm*:

Cluster	finds groups that maximize similarity within groups
Exhaustive	finds groups that maximize statistical significance

Amazon/AlmondNet
Exhibit 1011
Page 180

Another setting is for the Split Search Criterion, which lets you change the criterion used to guide the split search for any field as well as change how to display significance values. The default method is the adjusted level of significance. This method is the most statistically sound approach since it reduces the possibility of producing chance results and puts all the fields on a common statistical footing. These are the options:

Adjusted	this setting is designed to ensure the integrity of the classification tree
Unadjusted	this setting is prone to a bias towards fields with a lot of categories
Entropy/Variance	this setting simulates the ID3 machine-learning algorithm

KnowledgeSEEKER incorporates both CART and CHAID algorithms concurrently, which allow the user to segment a data set on either a categorical or continuous value variable. Another useful feature of KnowledgeSEEKER is the ability to *Reshape* any decision tree, which enables the user to page through various splits of their data set. For example, Figure 5–3 is a view on an ACORN consumer class "Successful Suburbanites" from our sample data set. Figure 5–4 is another decision tree of the same data set, split by another consumer class, "High Rise Renters."

KnowledgeSEEKER not only generates decision tree, but can also produce Crosstabs, IF/THEN rules, and SQL syntax for export use. For example, the following rules and syntax were generated from the decision tree in Figure 5–4.

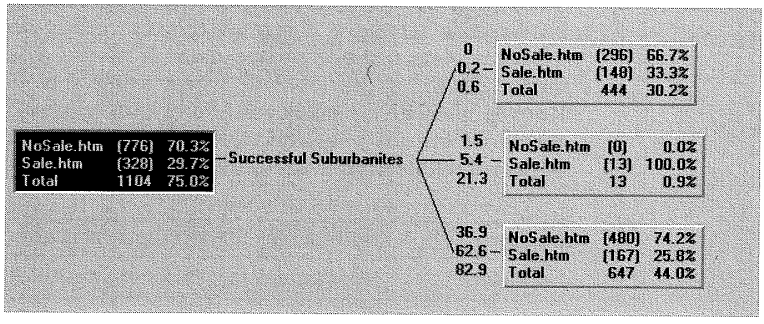


Figure 5–3 *KnowledgeSEEKER's split of the web data of a specific consumer class: Successful Suburbanites.*

Amazon/AlmondNet
Exhibit 1011
Page 181

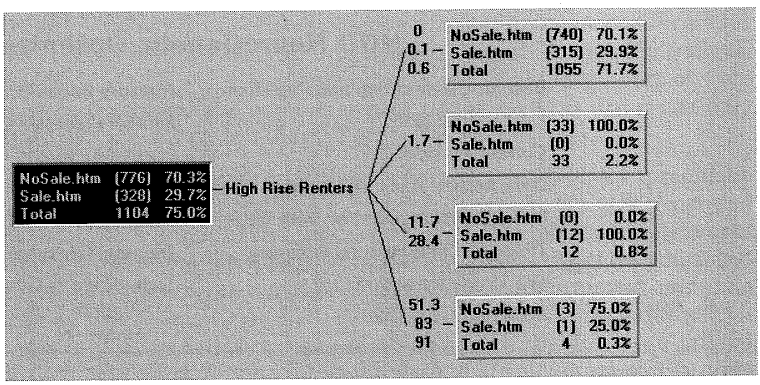


Figure 5–4 *KnowledgeSEEKER's split of the web data of a specific consumer class: High Rise Renters.*

RULE_1 IF	High Rise Renters =	0, 0.1 or 0.6
THEN	HTM = NoSale.htm	70.1%
	HTM = Sale.htm	29.9%
RULE_2 IF	High Rise Renters =	1.7
THEN	HTM = NoSale.htm	100.0%
	HTM = Sale.htm	0.0%
RULE_3 IF	High Rise Renters =	11.7 or 28.4
THEN	HTM = NoSale.htm	0.0%
	HTM = Sale.htm	100.0%
RULE_4 IF	High Rise Renters =	51.3, 83 or 91
THEN	HTM = NoSale.htm	75.0%
	HTM = Sale.htm	25.0%

Overall, KnowledgeSEEKER is one of the best decision tree tools in the market. The tool has a very clean and easy-to-use interface, it is very fast and accurate, and it provides the users plenty of options in modifying their data and in generating different trees and subsequent IF/THEN rules. KnowledgeSEEKER is very robust and makes a very good tool for segmenting web databases, offering multiple solutions in various formats for rapid deployment by web administrators and business and marketing analysts. A new product, knowledgeSTUDIO by AGNOSS, incorporates a neural network.

Amazon/AlmondNet
Exhibit 1011
Page 182

NGO NeuroGenetic Optimizer

Best at: Classification and Prediction
Cost Factor: No
Current Version: 2.1
Data Access: ASCII delimited
Data Preprocessing: Yes, with EMS
Distinguishing Feature: Accuracy
File Limitations: None
Installed Base: 350
Memory Required: 16 MB
Parallel Processing Support: Yes
Pricing: $295 and up
Solution Format: API
Space Required: 5 MB
Training/Consulting: Yes
Technology Paradigm: Genetic Algorithm and Multiple Neural Networks

NGO NeuroGenetic Optimizer

The NeuroGenetic Optimizer (NGO) is a self-optimizing neural network model-generating tool. It uses a genetic algorithm component to select key data variables and optimize the structure of a back-propagation, a Self-Organizing Map, and other neural network architectures to maximize the predictive accuracy of its models.

The concept behind this tool is that rather than building one model at a time, you build multiple ones simultaneously. This particular approach to data mining and model construction lends itself quite well to dynamic business applications requiring multiple forecasting models. For example, BioComp's Enterprise Modeling Server can generate multiple models for each product an e-commerce site may carry. Another feature this tool has going for it is that of combating model deterioration. As with any other types of models, neural network models tend to decay with time. Having the ability to quickly refresh them with new data is clearly an asset, which this tool delivers.

The NGO uses genetic algorithms to evolve neural network structures while simultaneously searching for significant input variables to maximize the predictive accuracy of its neural network models. The effectiveness of the genetic algorithms' search capabilities in optimizing the settings of a neural network is an excellent idea. For example, a problem consisting of finding the best combination (subset) of 20 inputs and up to 15 hidden nodes in a back-propagation neural network is a combinatorial problem with over 16 million permutations. According to BioComp Systems, an excellent solution often appears in less than 1500 evaluations, which is 0.009 percent of the total possible configurations of a neural network. Commonly highly fit networks are often found in the first 30 to 50 neural networks evaluated. This is clearly an efficient

Amazon/AlmondNet
Exhibit 1011
Page 183

means for discovering effective network structure/input combinations. The one drawback to this approach is one of time. A complex analysis, say of 10 or 20 web pages on your site, will likely take hours or even days to complete. The other drawback of the NGO is that it can only construct models on less than 32,000 observations.

This adaptive multimodeling capability is a good pragmatic approach to data mining highly dynamic data such as a web-generated database. The result of this process is higher accuracy neural networks that are optimized for the entire website with less effort on your part. The NGO automates much of the modeling development and data mining process. Some of the tedious tasks like the splitting of the testing/training data sets, determining which input variables to use and the overall neural network architectural design, are all automatically done for the user by the software. The NGO uses its genetic algorithms to evolve the neural network structures and select which input variables are the most important to what you are trying to predict, such as total sales for each product on your website. Once NGO has produced a model that fits the data, that is, learned the relationships between multiple inputs and a desired output, you need a separate product called the Predictor to view and use the resulting network. Predictor is shipped with the NGO and allows you to view graphs resulting from the network. It also allows you to enter data for predictions or load files to create scores that can be saved to disk.

Using the NGO can be as simple as opening a data file and running the software. The tool prompts you for your data format type and loads it into memory. Next it prompts you for the identification of inputs and desired output. Since you are likely going to be dealing with an ASCII comma-delimited flat file, generated from your server, you must tell NGO which fields to use as inputs and what field will be your output. From that point NGO goes on to build and validates the training and testing data sets, providing a variety of methods for viewing the modeling processes. NGO provides multiple graphs for monitoring its modeling activity, such as *Predicted vs. Desired* or *Accuracy vs. Error Trend* results, *Status* and *Report* screens—all of them can be viewed from multiple screens. This is a good feature since once you start the mining process you need to know that the training and testing of the data are taking place in the background.

The NGO also creates a population of candidate-input variables and neural structures, training them and evaluating them. It next selects the top networks and pairs up the genetic material representing the inputs and neural structure of the networks and does the exchanges of the genetic material between them. As with all proper genetic algorithms it also throws in a few mutations for a flavor of

Amazon/AlmondNet
Exhibit 1011
Page 184

random searches before it goes back into the training/testing cycle again. These processes continue for a defined number of cycles (generations), for a defined period of time, or until a neural accuracy goal is reached.

If you leave the training at the default settings, NGO will continue running, splitting, pairing, and recycling for several hours. In our sample data set of 1472 observations with seven fields as inputs/outputs, the NGO ran for over six hours before it converged on an answer. This process is clearly computationally intensive, but the effectiveness of genetic algorithms combinatorial search capabilities with neural networks is truly great. During a run, the NGO provides you the ability to view the status of what is happening, view the evolving population, see the configurations and the statistics of the top ten networks. You can also observe learning curves and watch the neural outputs match your desired data for the network being trained and view and/or print reports on the specifics of the system setup and the resulting top networks. After several hours NGO found that being from a ZIP code with a high percentage of "Successful Suburbanite" consumer class was the most important attribute for predicting online sales.

WizWhy

WizWhy is a data mining tool originally developed and marketed in Israel. It uses a proprietary algorithm to generate IF/THEN rules and predictions. This algorithm is very extensive, meaning that the more data samples you have the longer it takes to generate rules. Another minor concern about WizWhy is that it generates a very large number of rules, which luckily, the user is able to reduce at the end of the data mining analysis. The interface is fairly straightforward, prompting you

WizWhy

Best at: Classification
Cost Factor: Yes
Current Version: 1.1
Data Access: ASCII/dBASE
Data Preprocessing: None
Distinguishing Feature: Provides rule probability, count, and significance level
File Limitations: None
Installed Base: 100
Memory Required: 640 K
Parallel Processing Support: No
Pricing: $5,000
Solution Format: IF/THEN rules
Space Required: 1.5 MB
Training/Consulting: Yes
Technology Paradigm: Proprietary Algorithm

Amazon/AlmondNet
Exhibit 1011
Page 185

for the location and type of file format to import. Once imported, the tool lists all the fields, prompting you for the field to predict and an option to "Analyze if Empty" and "Ignore Field" for all the other fields in your imported database. Once you are done with your selections, you click on the *Rule Type,* where you can select the parameters for the rules you want WizWhy to generate. If you like, you can simply leave WizWhy at the default settings and select the *Issue Rules* button.

The following is a sample of the type of Rule Report generated by WizWhy, using the default settings. The settings can be modified to increase the accuracy of the rules and to reduce the number of rules. For example, this analysis generated 33 rules, which we reduced to 20 for brevity; the settings for reducing them are also adjustable. After some time a report and a group of rules will be generated directly from your web data. Here are the top five rules extracted from our sample data set:

WizWhy REPORT

Total Number of Records: **1472** Minimum probability of the:

1) if-then rules: **0.410**
2) if-then-not rules: **0.820**

Minimum Number of Cases in a Rule: **20**

Field to Predict: **Webminer.HTM**
Field to Predict is analyzed as Boolean in reference to the following value:

OrderForm.htm

Prediction error costs:
Error of the first kind: 1
Error of the second kind: 1

Mean probability of the predicted value is 0.300

PREDICTION ON THE PRESENT FILE Decision point:

Predict OrderForm.htm when conclusive probability is more than **0.449**

Number of errors of the first kind (misses): 339

Number of errors of the second kind (false alarms): 64

Total number of errors: 403

Success rate when predicting **OrderForm.htm: 0.600**

Success rate when predicting **NOT OrderForm.htm: 0.736**

Total cost of errors: 403

Number of records with no relevant rules: 26

Average cost (per record): 0.279

1) *If* **Subdomain** *is* **net**
and **Product** *is* **Printers**
Then
Webminer.HTM *is* **OrderForm.htm**
Rule's probability: **0.793**
The rule exists in **23** *records.*
Significance level: Error probability < 0.01

2) If **Income** is **789.00 ± 617.00**
and **Product** *is* **Drives**
Then
Webminer.HTM *is* **OrderForm.htm**
Rule's probability: **0.639**
The rule exists in **23** *records.*
Significance level: Error probability < 0.1

3) *If* **High Rise Renters** *is* **1.00**
Then
Webminer.HTM *is* **not OrderForm.htm**
Rule's probability: **1.000**

Amazon/AlmondNet
Exhibit 1011
Page 186

WizWhy REPORT *(continued)*

The rule exists in **45** *records.*
Significance level: Error probability < 0.1

4) *If* **Enterprising Young Singles** *is* **98.00**
Then
Webminer.HTM *is* **not OrderForm.htm**
Rule's probability: **1.000**
The rule exists in **45** *records.*
Significance level: Error probability < 0.1

5) *If* **ZIP** *is* **94105**
Then
Webminer.HTM *is* **not OrderForm.htm**
Rule's probability: **1.000**
The rule exists in **45** *records.*
Significance level: Error probability < 0.1

Because of WizWhy's relative slowness, it would not be appropriate for a large website. However, a small to medium site that does not require daily analysis will find WizWhy to be an excellent tool for generating rules with probability rates, record counts, and significance measurements. Note also that it has the ability to generate IF/THEN as well as IF/THEN NOT rules, and is able to tell you the probability of the rules and the number of instances as they appear on your training data set.

An important point to consider in the evaluation and selection of data mining tools is that these products and their results can be linked in order to obtain a hybrid solution. For example, you may want to use a decision-tree tool like KnowledgeSEEKER or WizWhy to segment and identify the key variables in your data set, especially if you are dealing with a large number of data fields. Then using the more compressed data set, with a reduced number of variables, you can use a neural network tool like NGO for classification—which with its embedded genetic algorithm can further optimize the accuracy of the final models for your website. (See Figure 5–5.)

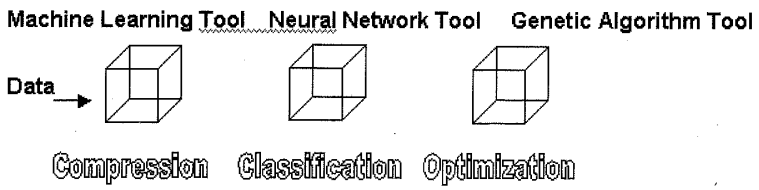


Figure 5–5 *Single technology data mining tools can be linked to maximize their solutions.*

Amazon/AlmondNet
Exhibit 1011
Page 187

The Data Mining Suites

The sum is often greater than its parts, and in data mining this becomes quite obvious when multiple technology tools are combined in a suite of paradigms. The following section showcases this new breed of multiple paradigm products.

These tools can combine features and technologies and are capable of linking analyses and their results in formats that can segment your website visitors by their demographic or lifestyle features, then present them in the form of SQL syntax, C code, or a 3D ZIP code map. Incorporated in these new tools are multiple technologies including statistics, neural networks, genetic algorithms, decision trees, and visualization modules enabling users to view relationships, extract patterns, and construct predictive models from their data.

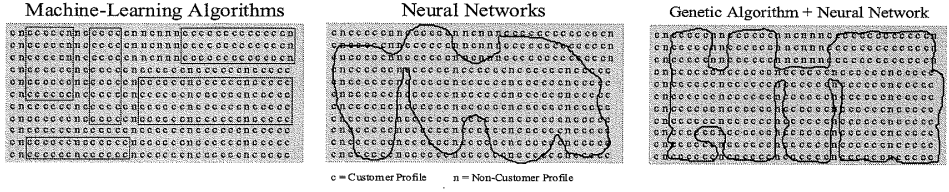
Using one or two of these high-end tools you can do visualization, segmentation, association, classification, prediction, clustering, and optimization. Some are capable of running on scalable servers and take full advantage of their multiprocessor architectures. These data mining vendors recognize that no single technology can answer all the business questions a data warehouse or a large electronic commerce site may have about their data and as such, their products are designed to provide multiple solutions. Specifically built to deliver business results, most of these tools are able to generate C code, SQL syntax, three-dimensional graphs and maps for presentations and exporting into production systems. Some have the ability to base their analysis on millions of observations.

Tool	Vendor	Platform	Technologies
Clementine	SPSS	NT/UNIX	Back-propagation, SOM, C5.0
DataEngine	MIT	WIN95/UNIX	SOM, Fuzzy Logic
Decisionhouse	Quadstone	WIN95/NT/UNIX	CART/ID3, Regression, GIS

One of the advantages of these types of tools is that you are able to search through data space in a different way. (See Figure 5–6.)

Amazon/AlmondNet
Exhibit 1011
Page 188

Figure 5–6 *Slicing and dicing data spaces with multiple processes.*



Clementine

Best at: Clustering and Classification
Cost Factor: Yes
Current Version: 5.0
Data Access: ASCII and ODBC
Data Preprocessing: Yes
Distinguishing Feature: GUI Interface
File Limitations: None
Installed Base: 500
Memory Required: 40 MB
Parallel Processing Support: Yes
Pricing: $43,000 single license, one-year maintenance
Solution Format: C code
Space Required: 65 MB
Training/Consulting: Yes
Technology Paradigm: Regression, back-propagation/Kohonen neural networks and ID3/C5.0 machine-learning algorithm

Clementine

Clementine from SPSS combines several data mining technologies in one product, including visualization, statistics, neural networks and machine-learning algorithms. Data can be imported as a flat file or connected via ODBC, from which several manipulations and processes can be performed simply by connecting icons from the main panel. The most impressive feature of this tool is the ability to perform all aspects of data preparation, graphing, analysis, and modeling by dragging, connecting, and dropping icons on its main panel—through what SPSS calls "visual programming." The Clementine icon-based programming environment is accessed from the main window (see Figure 5–7).

Data preparation is accomplished using nodes from three palettes: Source, Record Operations, and Field Operations. These nodes give the user the capability to link and import data, to view the data in a table, and to transform the data. The Record Operators allow the user to select, merge, sample, and balance the records used for the analysis. They can be used to select a subset of the records based on a specific condition, to join records from multiple files, or do random

Amazon/AlmondNet
Exhibit 1011
Page 189

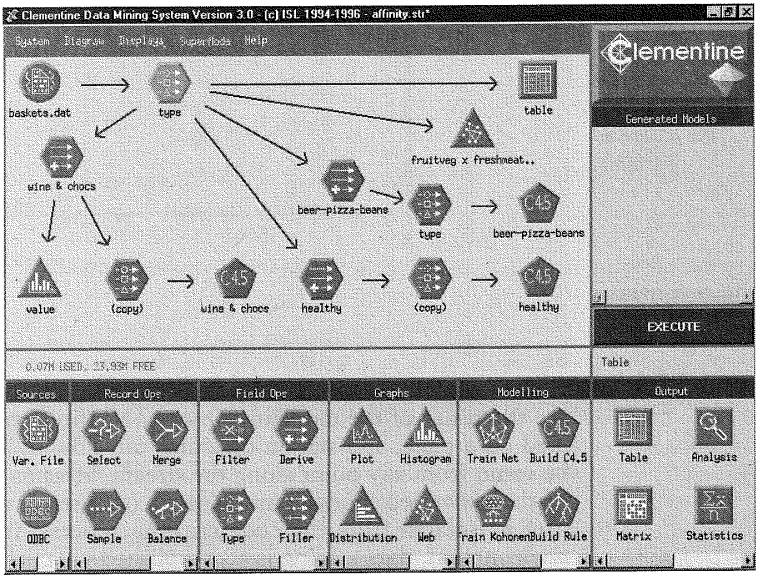


Figure 5–7
The Clementine "canvas" for mining data.

selections and other types of data manipulation. The Field Operators allow the user to filter records, create record types, and derive fields as well as other preprocessing transformations. A graphical group of nodes allow the user to generate plots, histograms, distribution charts, and generate a link association graph it calls "the web."

The modeling nodes allow the user to perform several operations including classification, segmentation, and clustering. Classification models can be constructed either via standard regression or a back-propagation neural network. Segmentation can be done via the C5.0 or ID3 machine-learning algorithms. The clustering is done with the Kohonen (SOM) neural network and the standard statistical nearest neighbor method. Several Output nodes are included to explore the results of the modeling processes as well as to generate C code for use by production systems.

To build a model the user simply selects and links the various Records, Field, Modeling, Graph, and Output nodes. Through this linking process procedures are created labeled "streams" by Clementine. These streams control data preparation, selection, partitioning, model building, testing, evaluation, and deployment. The stream pane is the primary work area, and the palettes across the bottom of the main panel contain the nodes and the "nuggets" that

Amazon/AlmondNet
Exhibit 1011
Page 190

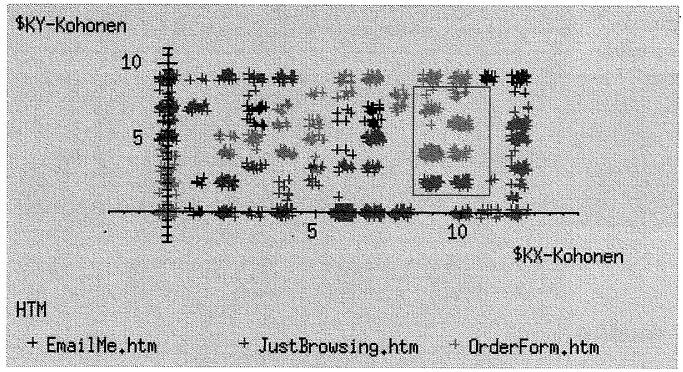


Figure 5–8
By marking the green cluster we were able to create a derived variable.

are the result of the modeling process. Multiple streams can be created on a pane, and stream diagrams can flow in any direction and be rerun from any point along the stream. Streams plus the contents of the generated models can be saved (as a state) and recalled. The user is isolated from the code creating process; in the Clementine environment, data mining is accomplished via click, link, and run.

One powerful feature of Clementine is its ability to link a clustering network to a rule-generating algorithm. For example, let's say you want to do an exploratory analysis of your website visitors from your log files—but you have no clear idea of where to start. Using one of the modeling nodes in Clementine known as Kohonen or Self-Organizing Map, you can build a cluster from your data, then "grab" it in the graphical environment and create a "derive" node. From there the derived node can then be routed through the C5.0 machine-learning algorithm, which can generate rules describing the features of the cluster discovered by the SOM. Using the SOM component, the sample data set was used to train the Kohonen Network, and several clusters were created (Figure 5–8).

Using the derived variable created from the clustering process we next generated a set of rules describing that grouping:

Rules for T:

Rule #1 for T:

if HTM == OrderForm.htm

and Subdomain == com

then –> T (172.0 Instances, 1.0 Confidence)

This rule is indicating a high number of orders occur with .com domain types

Amazon/AlmondNet
Exhibit 1011
Page 191

Rule #2 for T:

if HTM == OrderForm.htm
and Subdomain == gov
then –> T (0.0 Instances, 1.0 Confidence)

This rule is indicating no orders are being made to government: .gov domain types.

Rule #3 for T:

if HTM == OrderForm.htm
and Subdomain == edu
and Product == Desktops
then –> T (54.0 Instances, 1.0 Confidence)

This rule is indicating that academic .edu domain types are making desktop orders

Rule #4 for T:

if HTM == OrderForm.htm
and Subdomain == edu
and Product == Monitors
and % of Wealthy Seaboard Suburb Consumer ZIP is > 16.7
then –> T (9.0 Instances, 1.0 Confidence)

This rule is indicating some monitor orders are being placed by schools in specific geographic neighborhood areas in Connecticut and New Jersey, where the consumer types "Wealthy Seaboard Suburb" reside.

Through this process the user can literally discover rules from a blank data space in a totally autonomous and unbiased manner. This process is especially effective in situations where the objective of data mining is exploratory in nature. It is also effective in situations for which there are no positive and negative samples from which to train a neural network or construct a decision tree. This feature clearly illustrates the power of integrated data mining tools—where the user can pass the results from one technology to another for refinement or for conversion into the desired format. In this example, we did not want the weights (formulas) from the neural network; we instead wanted the insight in the form of easy to understand rules, which described the attributes of clusters in the website database.

Clementine's visual programming interface is unique and powerful. It is intuitive enough so that nonstatisticians can construct and compare multiple models very quickly on website data. Very complex data mining processes, which if done manually would take time

Amazon/AlmondNet
Exhibit 1011
Page 192

and experience, can be accomplished with the click-and-drag functions of the Clementine environment. The preprocessing abilities and the ability to construct parallel models and compare results side-by-side are very powerful features, which most single paradigm tools are not able to do.

One of Clementine's drawbacks is its inability to generate graphical decision trees. This makes for a very poor interface when viewing the results of the machine-learning algorithms. The most powerful feature of this tool is the ability to link the clustering Self-Organizing Map (Kohonen) neural network with the machine-learning algorithm C5.0—enabling the user to generate rules describing clusters. This is a unique feature that few tools can do: *the linking of unsupervised learning to supervised learning.* Clementine runs on Sun and other major Unix platforms, and it uses Hummingbird for emulation on Windows workstations where is runs on NT.

The ability to link data mining processes makes tools such as Clementine worthwhile. Clementine's easy-to-use visual interface enables web administrators, marketing, and business users to build models for prediction, forecasting, clustering, visualizing, estimation, and classification directly from their website data.

DataEngine

DataEngine is a tool from Germany incorporating several clustering types of technologies, including the following modeling techniques:

- Fuzzy Rule Base SystemMultilayer Perception
- Fuzzy C-Means
- Kohonen Network Fuzzy Kohonen Network

DataEngine

Best at: Clustering and Association
Cost Factor: Yes
Current Version: 2.1
Data Access: ASCII, Excel, and ODBC
Data Preprocessing: Yes
Distinguishing Feature: Clustering
File Limitations: None
Installed Base: 100
Memory Required: 32 MB
Parallel Processing Support: Yes
Pricing: $10,000
Solution Format: API
Space Required: 47 MB
Training/Consulting: Yes
Technology Paradigm: Kohonen neural networks and Fuzzy Logic

Amazon/AlmondNet
Exhibit 1011
Page 193

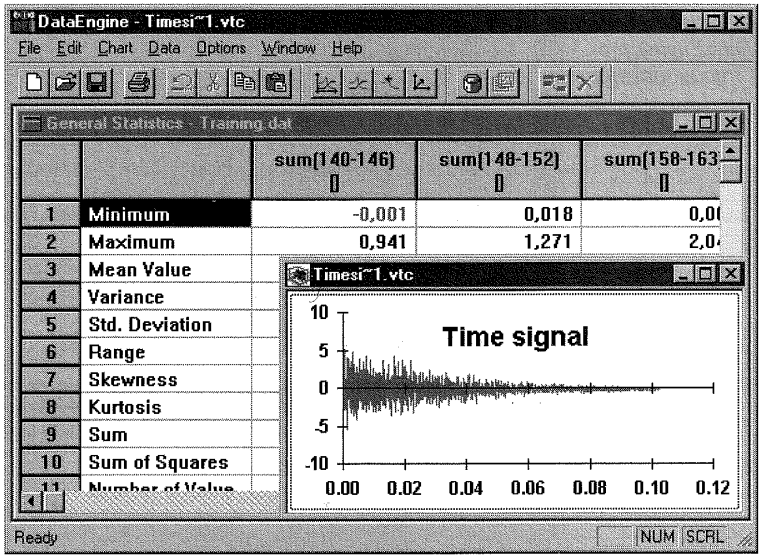


Figure 5–9 *DataEngine readily provides data parameters to the user.*

Each of these modeling methods can be configured to perform an analysis of your web databases, but although the importing of the data is fairly straightforward, deciding which clustering method to use is not. DataEngine has several useful components for data acquisition, preprocessing, and visualization. It also contains an extensive list of mathematical functions for scaling your data and a statistical component for correlation and regression analysis as well as general descriptive statistics. (See Figures 5–9 and 5–10.)

Several statistics functions are included in DataEngine for the treatment of your data, such as: moving average, moving minimum, moving maximum, moving variance, and moving standard deviation. This type of tool is optimized for control problems rather than the exploratory mining of web data. For example, because of the need to deal exclusively with numeric data, when it comes to using a Kohonen network there is a need to convert the original data. This can be done within most data mining tools or outside of them such as in a spreadsheet or database program. One common method of converting "string" fields like domain types, is as follows:

Amazon/AlmondNet
Exhibit 1011
Page 194

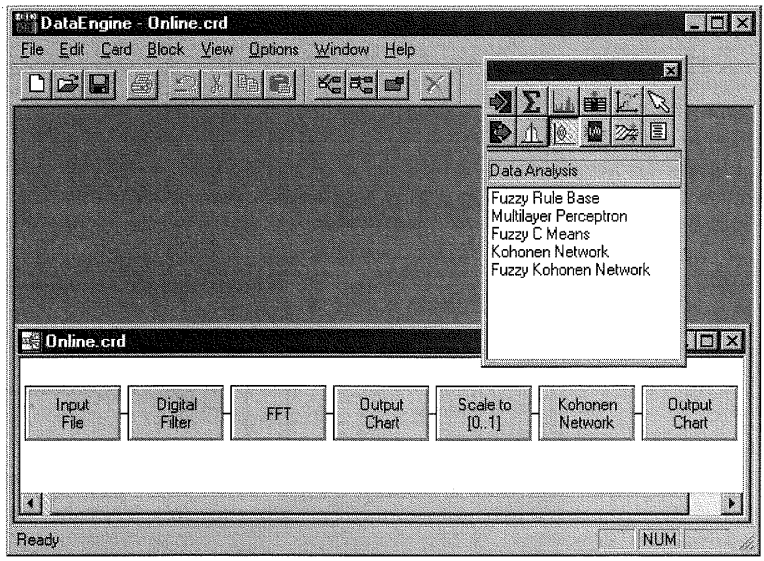


Figure 5–10 *DataEngine provides multiple data analysis processes to the user.*

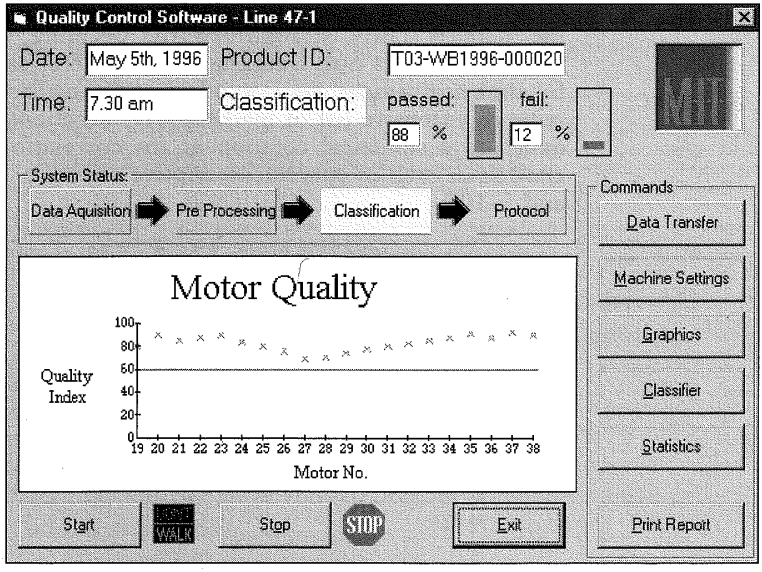


Figure 5–11 *DataEngine enables users to evaluate and map the data mining process.*

Amazon/AlmondNet
Exhibit 1011
Page 195

Original Subdomain	Converted Value
COM	0000001
GOV	0000010
JP	0000100
NET	0001000
ORG	0010000
UK	0100000
EDU	1000000

The same holds true for our desired output:

Possible Outputs	Converted Value
E-mailMe.htm	001
JustBrowsing.htm	010
OrderForm.htm	100

Although DataEngine contains a suite of very powerful clustering and association techniques and features (see Figure 5–11), the user should be well versed in their use prior to attempting to mine web data with this tool.

Decisionhouse

Decisionhouse is a powerful multiple-paradigm data mining tool from Quadstone Ltd. in Scotland. Decisionhouse incorporates several components

Decisionhouse

Best at: Visualization
Cost Factor: Yes
Current Version: 1.2
Data Access: ASCII and ODBC
Data Preprocessing: Yes
Distinguishing Feature: Tool Following
File Limitations: None
Installed Base: 50
Memory Required: 40 MB
Parallel Processing Support: Yes
Pricing: $150,000
Solution Format: Scores and graphs
Space Required: 65 MB
Training/Consulting: Yes
Technology Paradigm: Machine-learning algorithm and regression

Amazon/AlmondNet
Exhibit 1011
Page 196

for data preprocessing, segmenting, visualizing, reporting, and classifying in a seamless integrated system. The tool is highly scalable, capable of running on advanced parallel computing platforms, and operates in a three-dimensional environment. All the components in Decisionhouse are fully interoperable, making it easy to pick out data of interest from any view, and to "drill it" further using other parts of the software:

- Binning Editor
- Decision Tree Builder
- Crossdistribution Viewer
- Crosstab Viewer
- Map Viewer

Scorecard Builder

A powerful feature of this tool is the ability to let the user interact with the data from a single record view to a view incorporating millions of records. The interface allows for a quick method by which the user can view, rotate, and drill down on very large databases. In the Decisionhouse environment, the user can quickly look at the distribution of any field in the database and analyze it with a high degree

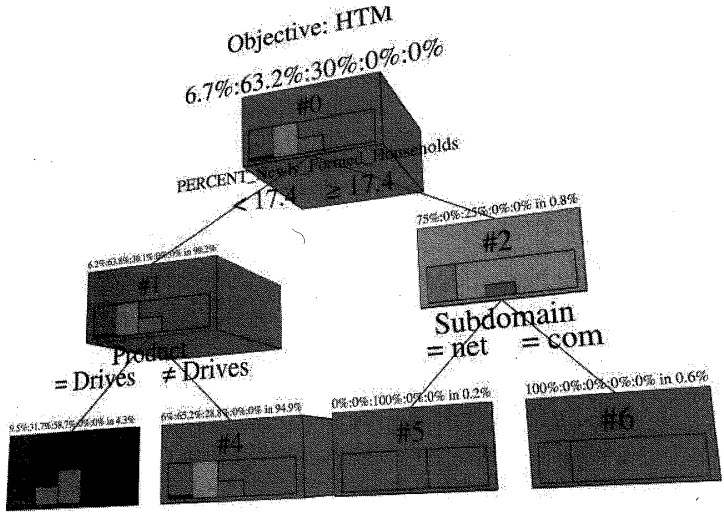


Figure 5–12 *A three-dimensional Decisionhouse decision tree, in which volume and color reflect content and rates.*

Amazon/AlmondNet
Exhibit 1011
Page 197

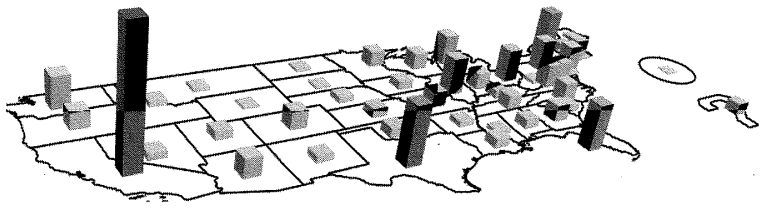


Figure 5–13
Decisionhouse users can map market segments from their website data with ZIP codes.

of flexibility. The tool provides the ability to segment the data, to create scenarios, and to pose quick questions against a database.

In the Decisionhouse environment the graphics and segments are represented by color and mass to reflect a high frequency rate. This enhances a "feeling for the data" as the user interacts with it during the mining analysis. From several of the Decisionhouse components the user has the options of selecting a group of attributes in the data and generating ad hoc views on millions of records almost instantly. This can be done either from the Binning Editor, the Crossdistribution Viewer, or the Map Viewer. The Decision Tree Builder performs the segmentation of the data, with the ability to link its results to the other components. For example, in the three-dimensional decision tree in Figure 5–12, two market segments (in red) were identified by the consumer class "Newly Formed Households" and domain types NET and COM.

The market segment identified in the decision tree can be mapped on a U.S. map or on one of a state, such as Texas or California, and viewed down to the ZIP code level. This market segment can also be viewed by other components, such as the Crossdistribution Viewer to explore more closely these newly discovered clusters. Because of its integrated designs and functionality a user can "fly over" the data from a web database and at any time drill down for further exploration using a variety of tools, including the generation of a map and/or SQL syntax.

For classification, Decisionhouse uses another component called the Scorecard Builder, which uses several types of regression. A retailing site that wants to predict propensity to purchase scores for site visitors could use the Scorecard Builder module. As with other integrated software suites, the power to link all of the components enables the user to segment, classify, visualize, and map the data in one seamless process. Another innovation of this product is its "tool following" feature. That is, the result of one tool can be passed or followed by another, so that for example by selecting from the above

Amazon/AlmondNet
Exhibit 1011
Page 198

decision tree the node #6, that market segment can be viewed in a geographic map. See Figure 5–13.

Other Tools

Darwin (Thinking Machines). Darwin is a set of tools oriented toward classification and regression. Like other suites it incorporates multiple paradigms. StarTree is used for building decision trees using a CART-based algorithm. StarNet is included for creating models using several neural networks. It supports five possible training architectures: back-propagation, steepest descent, conjugate gradient, and modified Newton. Darwin also contains a genetic algorithm for optimization problems called StarGene. Another component called StarMatch produces clustering models using the k-nearest neighbor algorithm. A final component, StarData, is used for data preparation and analysis. A scripting facility lets users record and play back data mining analyses.

Darwin has a GUI for building models using each of these components. The algorithms are parallelized for handling large amounts of data. Darwin's use of multiple algorithms for model building broadens the range of problems on which the product can effectively work. Although developed by Thinking Machines to take advantage of scalable platforms, the weakness of this tool is its output, specifically the format of its solutions—it lacks visualization. For reporting its modeling results Darwin relies on the spreadsheet Excel, through integration with MS Office.

Decision Series (NeoVista). Decision Series is basically a custom built set of applications bundled with their Professional Services consultancy. Their Decision Series is a suite of customized data mining tools that provide different models and technologies. DecisionNet is a back-propagation neural network tool, employing a proprietary technique closely related to radial basis functions, and can be used for classification and regression. DecisionAR can be used for association and sequencing discovery. DecisionCL is used for clustering, employing a proprietary technique closely related to k-means. DecisionAccess provides data-preparation functions, and Decision Series accesses data in ASCII files.

NeoVista offers Decision Series only in a bundle package with its knowledge engineering services. Consequently, NeoVista does not have a GUI for model development, although the company develops

Amazon/AlmondNet
Exhibit 1011
Page 199

a custom GUI for most applications. Each of their four applications—Neural Nets, Clustering, Genetic Algorithms, and Association Rules—costs $135,000. NeoVista is currently providing decision support software and services to large retailers, banks, and telcos. A typical NeoVista install involves several weeks of on-site interviews by its knowledge engineers for purposes of identifying the business problems, which their data mining solutions have been purchased to solve.

Enterprise Miner (SAS). Enterprise Miner is an SAS System entry into the data mining market. As with everything SAS sells, Enterprise Miner is an add-on module to the SAS base system. It provides a GUI to three mining modules: the SAS Neural Network Application and the SAS Decision Tree Application, for building CHAID-based decision trees, and an SAS Regression Application. Bundled with the product is the SAS/Insight visualization tool for data exploration. A useful feature of EM is its integrated methodology that allows users to run regression, a back-propagation neural network, and a decision tree side by side and then integrate their results in a visualization module. The technology is very basic with no special algorithms or features in the neural network or decision tree modules. The most powerful feature of EM is the installed base of SAS users, which is sizeable, especially with financial related firms.

GainSmarts (Urban Sciences). GainSmarts provides data analysis, statistical modeling, and financial measurement for vertical market segments, including the automotive and financial services industries. For the automotive sector it lets dealers and sales managers "see" their best opportunities in an array of useful charts, maps, and reports. Their tool provides "what-if" scenarios for new and used vehicle sales departments. Gain guides users to the hidden potential in automotive dealerships through advanced data visualization, mapping, and intelligent performance benchmarking. For the financial sector it provides a turnkey system specifically designed for prospect targeting, response prediction, customer retention, loyalty measurement, cross-selling, and offer management. Gain uses decision tress, genetic algorithms, and gain tables to chart expected profitability for mail campaigns. It uses genetic algorithms to optimize the CHAID decision trees, a technique, although computationally intensive, that does improve results.

Intelligent Miner (IBM). IBM's Intelligent Miner is a set of tools for classification, association, and sequence discovery, time series, clustering, and regression. For classification it uses a decision tree and

Amazon/AlmondNet
Exhibit 1011
Page 200

neural networks and for clustering it uses a Kohonen neural net. Most of the algorithms have been parallelized for scalability. Models can be built using either a GUI or an API. Intelligent Miner is tightly coupled with DB2, which also must be installed, but Miner supports input from sources such as ASCII files.

MineSet (SGI). MineSet from Silicon Graphics Inc. is a set of data mining tools that combine classification and association algorithms with a high degree of visualization. Their strongest and most notable features are in the integration of data mining analytic tools with high-end visualization tools for user exploration and navigation of data sets and mining results. The product's data mining tools include the Association Rule Generator, Decision Tree Inducer for classification, Evidence Inducer, and Column Importance determination utility. All the algorithms, except association, are based on the Machine-Learning Library in C++ (MLC++) code developed at Stanford University.

Visualizers, such as the Map Visualizer and the Scatter Visualizer, work directly with data to explore relationships and trends. Models are presented to the user for analysis using 3D visualization. Model results are displayed with visualizers such as the Tree Visualizer, which lets the user explore a decision tree model in a 3D format. The tools present several dimensions of data simultaneously by using color, size, and animation. The visualizers support filtering, querying, rotation, zooming, and panning, and can be customized by the user. SGI has added k-means and iterative k-means algorithms for clustering and has developed graphical outputs for clustering in its visualizer. MineSet 2.5 also supports regression trees for the generation of decision trees. The software can now create decision tree models on continuous variables.

Orchestrate (Torrent Systems). Orchestrate is an excellent object-oriented application environment designed for building and deploying scalable data processing software systems. Based on the data-flow model of application programming and execution, Orchestrate consists of two parts: the Orchestrate Application Framework and Runtime System and Orchestrate Components. Developers build Orchestrate Applications using standard UNIX shell-scripting techniques to combine existing sequential programs with Orchestrate components to construct a sequential data-flow Orchestrate program. Developers build Orchestrate Components using Orchestrate's robust C++ Component Development Interface. The Orchestrate Application Framework and Runtime System is a parallel software infrastructure which maps the sequen-

Amazon/AlmondNet
Exhibit 1011
Page 201

tial Orchestrate data-flow application to the underlying hardware architecture and manages run-time execution of the system as a single UNIX program.

Pattern Recognition Workbench/Model 1. PRW from Unica Technologies is a set of tools for building classification, clustering, time-series, and regression models. In addition to building models with back-propagation neural networks, it provides algorithms for logistic regression and linear regression. PRW builds clustering models using k-means. The tool set is the basis of Model 1, a package for database marketing modeling and analysis sold by Group 1 Software and Unica.

PRW provides a GUI based on a spreadsheet-style interface. The data must be brought into one or more spreadsheets, after which the data is prepared for mining via the product's extensive set of functions. PRW will automatically generate alternative models and search for a best solution. It also provides a variety of visualization tools for monitoring model building and the interpretation of its results. A model can be deployed as a spreadsheet function, as a dynamic link library, or as C code.

DecisionCentre and the Future. The future of data mining tools may well resemble those of DecisionCentre, an experimental tool developed at the Imperial College in the U.K. It uses a three-tiered architecture to perform parallel data mining anywhere on the Web. The tool uses Java and CORBA to provide a web-centered approach to data mining. The DecisionCentre server and the command and visualization client will run on any platform that supports Java, and can access data sources anywhere on the Web or on a corporate intranet. DecisionCentre supports several parallel data mining algorithms, implemented with the standard MPI. The Data Mining services currently available include:

- Decision Tree-Based Classification (four different parallel options)
- Bayesian Classification with Parallel Autoclass
- Parallel Self-Organizing Maps for Clustering
- Back-Propagation Neural Networks Association Rule Discovery

The Java-based visualization tools are integrated in the client interface. They can be used to view the data source as well as the output models from the data mining process. The visualization tool set contains a Self-Organizing Map Visualization Classification Tree Browser and a 3D Visual Data Mining.

The DecisionCentre three-tier architecture offers scalability, accessibility and extensibility. It is written in pure Java, providing portabil-

Amazon/AlmondNet
Exhibit 1011
Page 202