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 1009
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 1009
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 1009
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 1009
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 1009
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

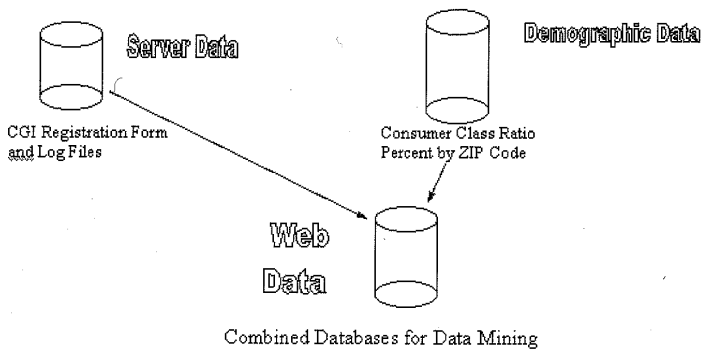


Figure 5–1
The merger of server and demographic data for mining.

Amazon/AlmondNet
Exhibit 1009
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

0,3,0,0,56,0,0,20,0,10,0,5,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, **"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 1009
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 1009
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 1009
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 1009
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 1009
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.

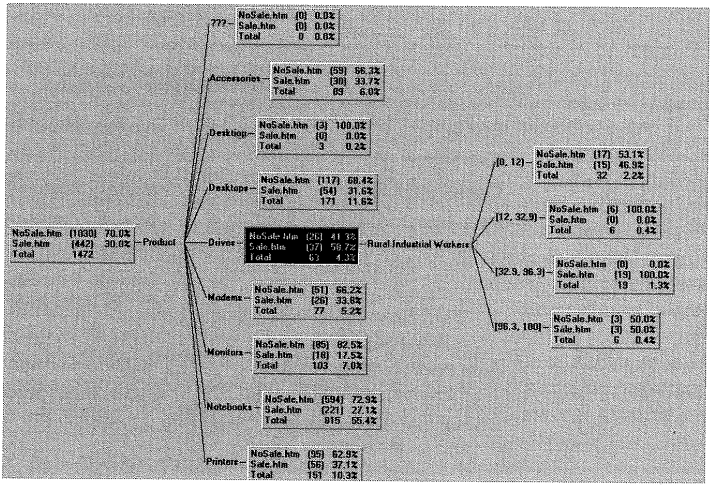


Figure 5–2 *KnowledgeSEEKER provides excellent control over the splits in the trees it generates.*

Amazon/AlmondNet
Exhibit 1009
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 1009
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.

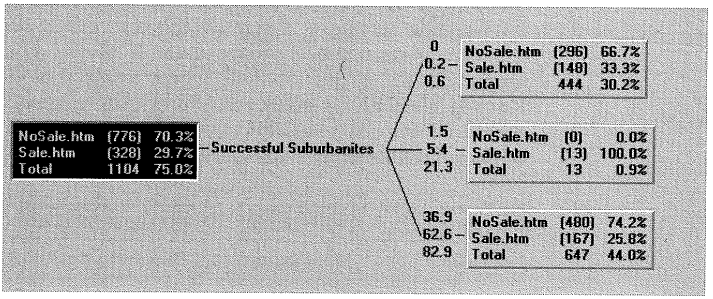


Figure 5–3 *KnowledgeSEEKER's split of the web data of a specific consumer class: Successful Suburbanites.*

Amazon/AlmondNet
Exhibit 1009
Page 181

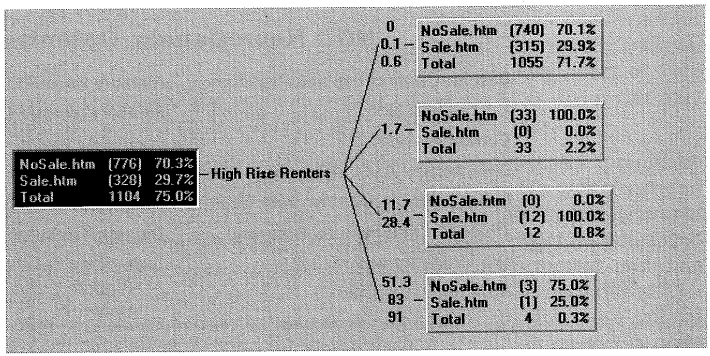


Figure 5–4 *Knowledge-SEEKER's split of the web data of a specific consumer class: High Rise Renters.*

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 1009
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 1009
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 1009
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 1009
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 1009
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.)

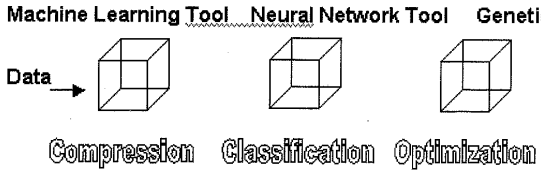


Figure 5–5 *Single technology data mining tools can be linked to maximize their solutions.*

Amazon/AlmondNet
Exhibit 1009
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.

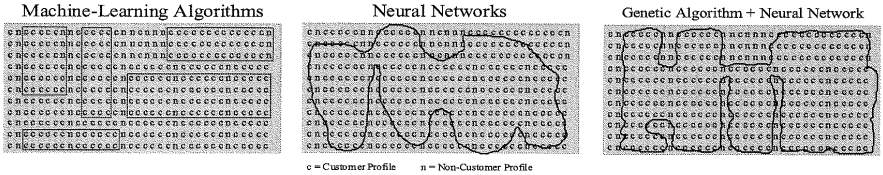
Using one or two of these high-end tools you can do visualization, segmentation, association, classification, prediction, clustering, and optimization. Some are capable of running on scalable servers and take full advantage of their multiprocessor architectures. These data mining vendors recognize that 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 1009
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 1009
Page 189

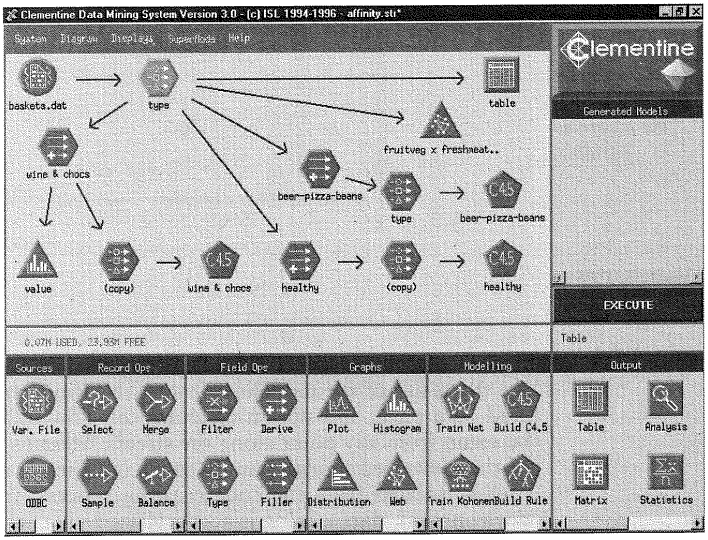


Figure 5–7
The Clementine "canvas" for mining data.

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 1009
Page 190

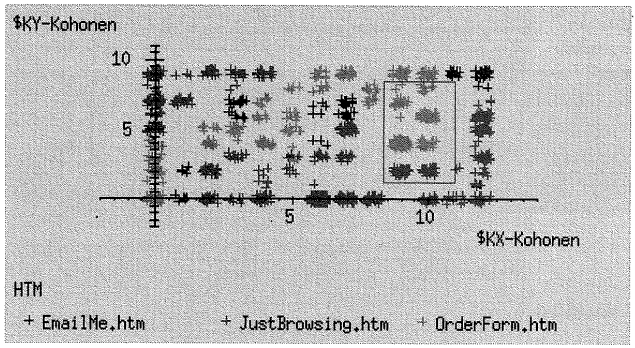


Figure 5–8
By marking the green cluster we were able to create a derived variable.

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 1009
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 1009
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	**Memory Required:** 32 MB
Cost Factor: Yes	**Parallel Processing Support:** Yes
Current Version: 2.1	**Pricing:** $10,000
Data Access: ASCII, Excel, and ODBC	**Solution Format:** API
Data Preprocessing: Yes	**Space Required:** 47 MB
Distinguishing Feature: Clustering	**Training/Consulting:** Yes
File Limitations: None	**Technology Paradigm:** Kohonen neural networks and Fuzzy Logic
Installed Base: 100	

Amazon/AlmondNet
Exhibit 1009
Page 193

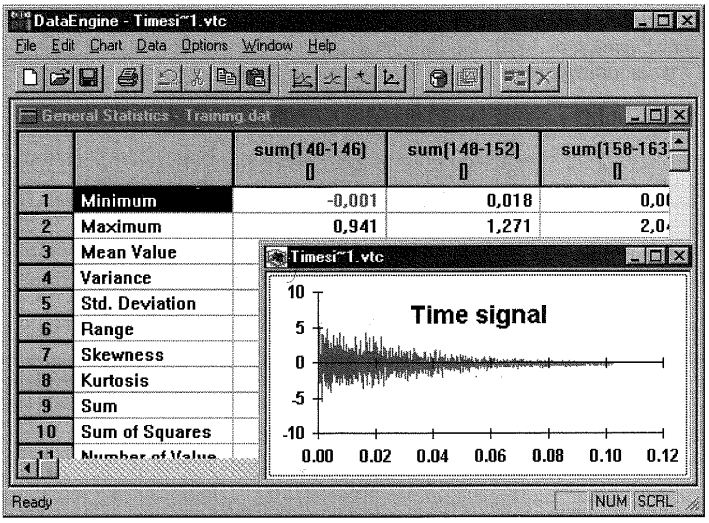


Figure 5–9 *DataEngine readily provides data parameters to the user.*

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 1009
Page 194

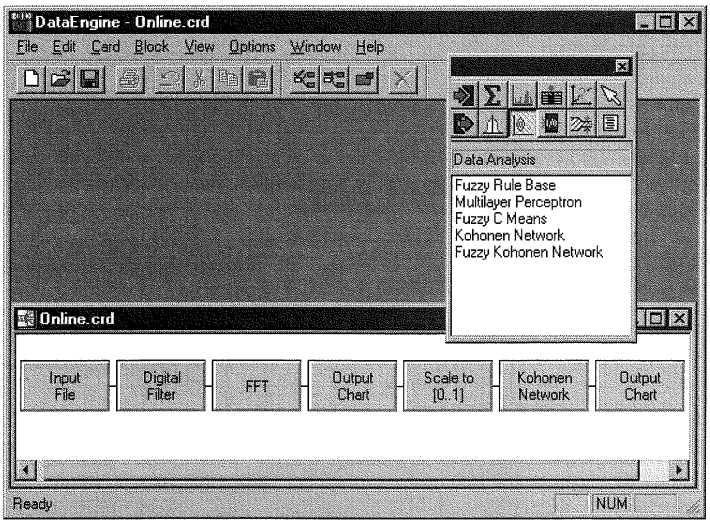


Figure 5–10 *DataEngine provides multiple data analysis processes to the user.*

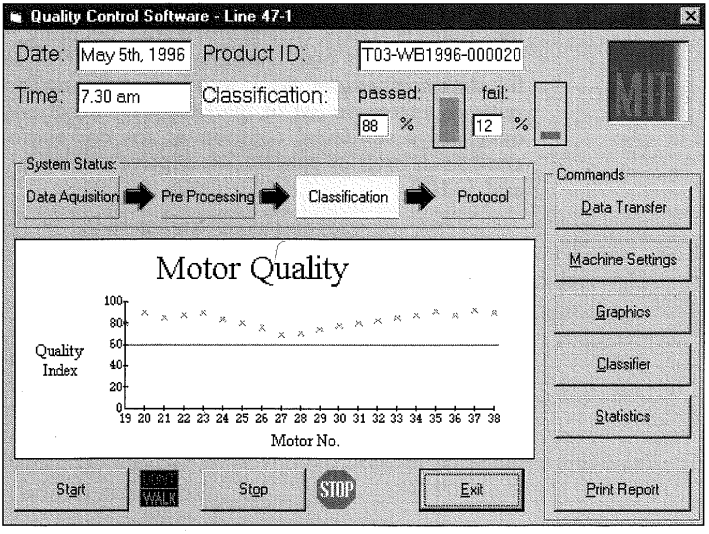


Figure 5–11 *DataEngine enables users to evaluate and map the data mining process.*

Amazon/AlmondNet
Exhibit 1009
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 1009
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

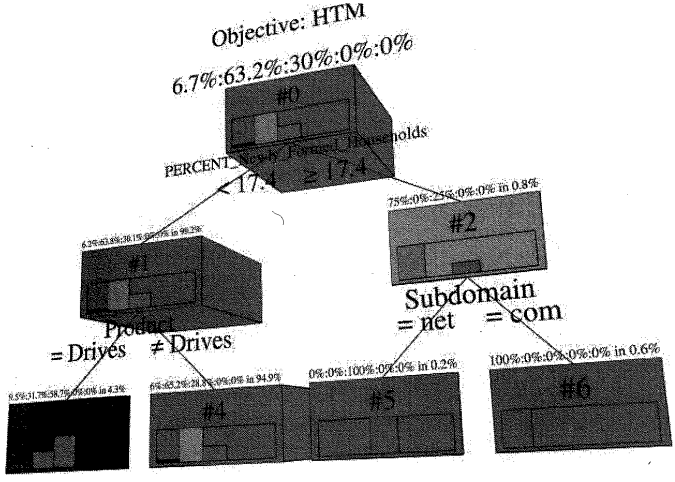


Figure 5–12 *A three-dimensional Decisionhouse decision tree, in which volume and color reflect content and rates.*

Amazon/AlmondNet
Exhibit 1009
Page 197

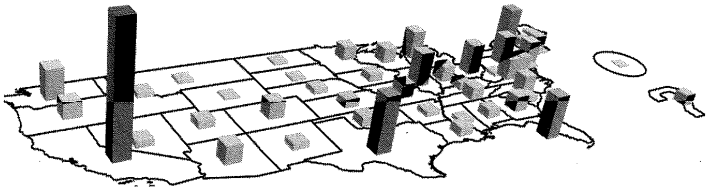


Figure 5–13
Decisionhouse users can map market segments from their website data with ZIP codes.

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 1009
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 1009
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 1009
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 1009
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 1009
Page 202

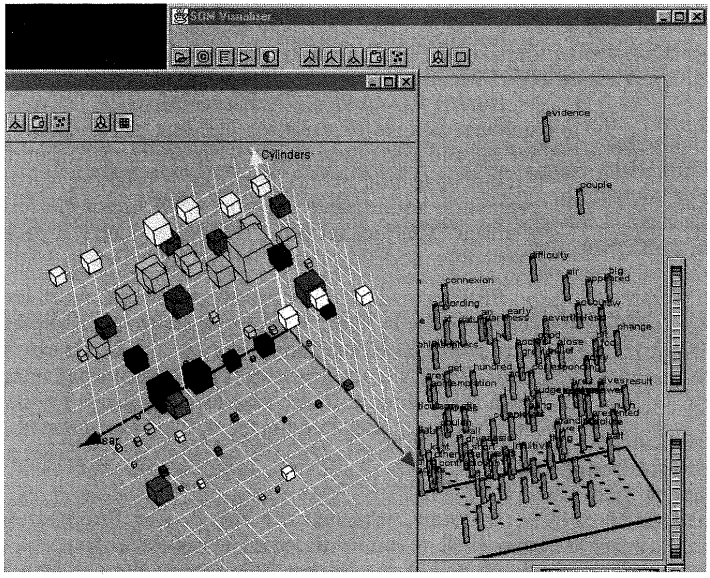


Figure 5–14 *The DecisionCenter Java client provides a view of the data being mined over the Web.*

ity and ease-of-use. Parallel data analysis applications are embedded via CORBA distributed object technology. (See Figure 5–14.)

Final Tool Tip

As you can see, there is a diverse and extensive assortment of features to the current data mining tool products. You will need to do some personal experimentation prior to deciding on which tool(s) are right for mining your web data. It is quite likely that you will need more than one tool and that you will need additional software for the extraction and preparation of your data *prior* to mining it. Likewise, you will likely need some additional tools in the integration of your data mining solutions. This may be in web design and database integration tools or bulk e-mail software. However, the data mining tools that you do use will be very important and integrated components in the operation of your website—especially if you are involved in e-commerce. New tools continue to appear in the market. See Tables 5–1 and 5–2.

Amazon/AlmondNet
Exhibit 1009
Page 203

Table 5–1 *Leading data mining tools.*

Product (Vendor)	O/S	Paradigm
4Thought (Right Information Systems)	WIN95, NT	Neural Network
Answer Tree (SPSS)	WIN95, NT	CART, CHAID, QUEST
BusinessMiner (Business Objects)	WIN95	Machine-Learning
Clementine (ISL)	NT, Unix	Machine-Learning and Neural Network
Darwin (Thinking Machines)	Unix	Machine-Learning and Neural Network
Data Mining Tool (Syllogic)	SGI, NT	Machine-Learning and Nearest Neighbor
Database Mining Marksman (HNC)	NT	Neural Network
DataCruncher (DataMind)	WIN95, NT, Unix	Machine-Learning
DataEngine (MIT)	WIN95, NT, Unix	Neural Network and Fuzzy Logic
Datasage (Datasage)	NT, Unix	Open API
Data Surveyor (Data Distilleries)	Unix	Machine-Learning and Bayesian Network
DbProphet (Trajecta)	NT, Unix	Neural Network
Decision Series (NeoVista)	Unix	Machine-Learning and Neural Network
Decisionhouse (Quadstone)	NT, Unix	Machine-Learning and Regression
Delta Miner (Bizzantz)	WIN95, NT	Neural Network and OLAP
//Discovery (HyperParallel)	Unix	Machine-Learning and Neural Network
Discovery Server (Pilot)	NT, Unix	Machine-Learning
Enterprise Miner (SAS)	NT, Unix	Machine-Learning and Neural Network
Enterprise Modeling Server (BioComp)	NT, Unix	Neural Network and Genetic Algorithm
GainSmarts (Urban Sciences)	NT, Unix	Genetic Algorithm, CHAID, Regression

Amazon/AlmondNet
Exhibit 1009
Page 204

Product (Vendor)	O/S	Paradigm
IDIS Data Mining Suite (IDIS)	NT, Unix	Proprietary Algorithm
KDD Explorer (SRA)	Unix	Proprietary Algorithm
Kepler (Dialogis)	WIN95, Unix	Open API
Knowledge Discovery Workbench (NCR)	NT, Unix	Neural Network and Machine-Learning
KnowedgeSEEKER (AGNOSS)	NT, Unix	CART, CHAID, ID3
Intelligent Miner (IBM)	RS6000, AS/400, MVS	Machine-Learning and Neural Network
MindSet (Silicon Graphics)	SGI, NT	Machine-Learning and Naïve Bayes
Model 1 (Group 1)	WIN95, NT	Networks, CHAID, Regression, GAs
ModelQuest Enterprise (AbTech)	WIN95, NT, Unix	Polynomial Networks
Nuggets (Data Mining Technologies)	WIN95, NT	Proprietary Algorithm
Orchestrate (Torrent)	Unix	Machine-Learning and Neural Network
Partek (Partek)	Unix	Neural Network and Genetic Algorithm
PATTERN (Magnify)	Unix	Proprietary Algorithm, ART
Polyanalyst (Megaputer)	Unix	Proprietary Algorithm
Predict (NeuralWare)	WIN95, NT, Unix	Neural Network
PRW PRO+ (Unica Technologies)	WIN95, NT	Neural Network, k-nearest, regression
Scenario (Cognos)	WIN95	Machine-Learning
Xpertrule Profiler (Attar)	WIN95, NT, Unix	Neural Network, ID3
WizWhy (WizSoft)	WIN95, NT	Proprietary Algorithm

Amazon/AlmondNet
Exhibit 1009
Page 205

Table 5–2 *Leading data mining suites.*

TOOL (COMPANY)	OS	Parallelized	Segmentation	Optimization	Classification
Clementine (ISL)	NT, Unix	No	C5.0	None	Neural Network
Darwin (Thinking Machines)	Unix	Yes	CART	GA	Neural Network
Data Mining Tool (Syllogic)	SGI, NT	No	Proprietary	None	Proprietary
DataEngine (MIT)	NT, Unix	No	None	Fuzzy Logic	Neural Network
DataSage (DataSage)	NT, Unix	Yes	No	No	No
Data Surveyor (Data Distilleries)	Unix	No	Proprietary	None	Bayesian
Decision Series (NeoVista)	Unix	Yes	Proprietary	GA	Neural Network
Delta Miner (Bizzantz)	WIN95, NT	No	OLAP	None	Neural Network
//Discovery (HyperParallel)	Unix	Yes	Proprietary	GA	Neural Network
Enterprise Miner (SAS)	NT, Unix	No	CHAID	None	Neural Network
Enterprise Modeling Server (BioComp)	NT, Unix	Yes	None	GA	Neural Network
GainSmarts (Urban Sciences)	NT, Unix	No	CHAID	GA	Regression
Knowledge Discovery Workbench (NCR)	NT, Unix	No	C5.0	None	Neural Network
Intelligent Miner (IBM)	RS6000, AS/400	Yes	Proprietary	None	Neural Network
MindSet (Silicon Graphics)	SGI, NT	No	C4.5	k-nearest	Naïve Bayes
Model 1 (Group 1)	WIN95, NT	No	CHAID	GA	Regression
Orchestrate (Torrent)	Unix	Yes	Proprietary	None	Neural Network

Amazon/AlmondNet
Exhibit 1009
Page 206

TOOL (COMPANY)	OS	Parallelized	Segmentation	Optimization	Classification
Partek (Partek)	Unix	No	None	GA	Neural Network
PRW PRO+ (Unica Technologies)	WIN95, NT	No	None	k-nearest	Neural Network
Knowledge Discovery Solutions (SRA)	Unix	Yes	Proprietary	None	Proprietary
Xpertrule Profiler (Attar)	NT, Unix	No	ID3	None	Neural Network

Amazon/AlmondNet
Exhibit 1009
Page 207

6

The Data Components

Every visit to your website creates a record of absolutely *everything* that happens during that session. Depending on how your website is configured, a significant amount of visitor and customer information can be captured and stored on your server log files or some other type of database for data mining. As your website increasingly becomes the first point of contact between your company and your current and future customers, a careful analysis of this web-based information can be quite important. For example, most likely your company wants to know who's visiting your site, what draws them there, and how they got there. You certainly would like to know what your website visitors are looking at, as well as what it is they are buying and most importantly *who they are*. Your website data is the most direct source of information about your current and potential future customers. In fact it is your company dynamic customer information file creating itself with every visit and online transaction. In this chapter we will discuss the important data components captured at your website that can provide valuable insight through data mining.

The single most time-consuming part of mining your website will involve the capturing, extraction, aggregation, and preparation of your server data for analysis. This will likely involve some preliminary planning by your web administrator or team in coordination with your sales, marketing, and other business groups in your company. At issue will be the methods and the type of customer and transactional information you want to capture at your website. There are various methods by which you can effectively capture very valuable customer information at the server level. This includes data from several types of web log files, cookies, databases linked to your website forms, dynamic tracking, and even third party vendors. Several technologies can also be used to capture website visitor and customer information, including the popular Common Gateway Interface (CGI), Java, and JavaScript. New web standards and features are also evolving, affecting how new browsers will interact and provide cus-

Amazon/AlmondNet
Exhibit 1009
Page 208

tomer information as electronic commerce transactions occur with new servers from such vendors as Netscape and Microsoft.

Data mining is an elaborate process involving not just the use of powerful pattern-recognition algorithms, but also planning how you will capture important transactional data, how you can append additional information to it, and finally how you can prepare it for actual analysis. In order to gain the most intelligence from your web data, it is important to know what files are available and what programs you can use to collect this information. The brick and mortar of your data mining analysis starts with your website data, more specifically several components that are used to create your server log files and other web-based databases, including transfer (access) logs, cookies, and forms.

Server Log Files

When a visitor first accesses your website, that single request can generate multiple transactions, which get recorded in your log files. First there is a recorded transaction for the requested page, such as info.htm, and then there are multiple "hits" for each graphical (gif) image on that page. Almost all web servers generate these log files as ASCII comma-, space-, or tab-delimited text files. Almost every web server incorporates, at a minimum, these two files: a *transfer log,* which tells you what your visitors saw, and an *error log,* which tells you about disconnects. Most servers also support two additional log files, which can tell you from where your website visitors originated: a *referrer log,* which tells you how your visitors found you, and an *agent log,* which tells you what browsers they used.

Depending on how your website is configured and what program you are using, your log files can be the starting point for finding out who your visitors and customers are. These files can assist you in finding where they came from and what "keywords" they used to find you, what path they followed through your website, and how long they stayed.

Transfer Log File

One of the most important sources of information about your website visitors starts with your server transfer log file, also known as the access log. This is where every transaction between the server and browsers is recorded with a date and time, the domain name or IP address of the server making the request for each page on your site, the status of that request, and the number of bytes transferred to that requester. At the minimum, almost every server records some of this information. The

Amazon/AlmondNet
Exhibit 1009
Page 209

standard method by which servers record access by browsers and other automated spider programs is the "Common Log Format" as created by the National Center for Supercomputing Applications (NCSA). The data fields in the transfer access_log file used by most web servers takes the following format:

1. Internet provider IP address (*webminer.com* or *204.58.155.58*)
2. Identification field (always a hyphen, or "-")
3. AuthUser (usually the ID or password required for accessing a protected area)
4. Date, time, and GMT (Greenwich Mean Time) of transaction (*Thu July 17 12:38:09 1999*)
5. Method of transaction (*"GET"* or *"POST"* with filename such as */index.html/products.htm*)
6. Status or error code (usually *200* (success) of transaction)
7. Size in bytes of transaction (file size)

The following is a detailed breakdown of the standard log file format and its seven standard data fields.

Host Field

This is the first field of the common log format, which is usually the remote server making the request from your website, it is usually in the format of a domain name sun51.cs.standford.edu or an IP address like 207.171.194.2.

Depending on how your server is configured, you will get either the domain name or an IP address in this first field. In order to obtain a name rather than an IP address, your server has to make a request to resolve (translate) the IP address into a hostname using the Internet Domain Name System (DNS). However, even when your server is configured to do these lookups with DNS of IP addresses, not every search will succeed. The reason is that most Internet Service Providers (ISPs) assign IP addresses dynamically to their clients from a reserved pool. This is especially true of large ISPs such as Earthlink, MindSpring, FlashNet, etc.

Identification Field

The second field in the Common Log Format is almost always a hyphen, due to the fact it is never used. The original Request for Comment (RFC 931) was for an authentication service back in 1984 laying out a proposal for a query/response protocol that would identify the owner of a

Amazon/AlmondNet
Exhibit 1009
Page 210

TCP connection between two computers on the Internet, specifically, to identify UNIX systems directly connected to the Internet. This RFC preceded the age of web browsers, when connections were mainly between FTP, ARCHIE, telnet, and mail servers. Several versions of the original RFC have occurred since 1984 with the current version being 1431 The Identification Protocol. The name of the actual program that implements the "ident server" is identd.

Authuser Field

This field was designed to contain the authenticated username that a visitor needs to gain access to a protected directory accessible only by password. For example, an administrator may create a special billing database where only authorized customers can gain access via the use of their usernames and passwords.

Time Stamp

This field contains the date, time, and the Greenwich Mean Time (GMT). The format for the date is DD/Mon/YYYY (i.e., "06/Jul/1999"), and the format for the time is HH:MM:SS. The GMT uses a plus or minus sign to indicate the number of hours the local time for the server differs from the Greenwich Mean Time.

HTTP Request

The transactional and fifth field contains the usual GET command, which is blocked by quotation marks. The GET method tells the server which document the visiting browser is requesting. The other two commands are used less often. They are the POST command, which is used with forms when the *Submit* button is pressed. POST tells the server to expect some data and what program or script can be used for handling it. A less used command is the HEAD, which works the same way as the GET method except that the server only returns the <HEAD> section of any document. HEAD commands are generated usually by special indexing programs, such as *WebCrawler*, one of the earliest search engines. The last part of the HTTP request field is the name and version number of the protocol, such as version 1.1.

Status Code Field

The sixth field in the Common Log Format describes what happened with the transaction, which is usually "GET" and its status code of 200, which stands for "Success." This simply means the server suc-

Amazon/AlmondNet
Exhibit 1009
Page 211

cessfully delivered the web page selected by the browser. There are various classes of status codes, the main ones are:

Class	Purpose	Descriptions
200	Success	201 Created
		202 Accepted
		204 No Content
300	Redirect	301 Permanently
		302 Temporarily
		304 Not Modified
400	Failure	401 Unauthorized
		403 Forbidden
		404 Not Found
500	Server errors	501 Not Implemented
		502 Bad Gateway
		503 Service Unavailable

Transfer Volume

The seventh and final field of the Common Log Format provides the total number of bytes transferred by the server to the client during the transaction. Only the GET requests with status 200 will have a value in this field, all others will only show a hyphen or a zero.

These seven fields are the standard fields generated by most servers. However, additional fields can be appended. Two of the most common ones are the Referrer Log and Agent Log files. A cookie field can also be appended, which can be used to track the activity of visitors while at a website, or upon their return.

Error Log

When a transaction in your website results in an error or failure to download, it gets recorded in the error log file. This can be useful in alerting you to a design problem, that is, situations in which visitors are canceling downloads in mid-process. Here is a sample of a line from an error log:

```
[Tue June 16 06:57:09 1999] send lost connection to client 165.133.159.28
```


Amazon/AlmondNet
Exhibit 1009
Page 212

Referrer Log

This log file contains the URL from which the requests for your web pages originated. In other words, this log captures the location your website visitors came from and how they found you. This can be a link from another page or a search engine "hit." For example, if they found you through a search engine like Yahoo! or Excite, this log can tell you what keywords they used to locate your website. This is very important information, which in a data mining analysis can reveal useful insight into possible strategic marketing initiatives and campaigns. This is a sample output:

```
http://search.yahoo.com/bin/search?p=data+mining+websites → /index.html
```

The complete URL for the search engine (Yahoo!) and the keywords used in the search (data mining websites) are to the left of the → symbol, with the link to your website (usually the opening page) on the right, in this case the common default of index.html.

Agent Log

This file records the name and version number of the browser making the request. It can also contain the name of robots, spiders, and other search engine indexing tracks.

```
Mozilla/2.0 (Win95; I)
Mozilla/2.0 (Macintosh; I; PRC)
MetaCrawler/1.2b libwww/4.0D
```

Extended Log Format

Rather than generating individual and separate referrer and agent logs, your server can be configured to append this information to the original seven fields found in the Common Log Format. For example, a rather common extension is the Referrer log which can tell you where your website visitors are coming from and what keywords they are using to find your site. Additional customer information can be generated about their movements within your site with server-set cookies, which can be appended to your log file.

IF	They were referred from Yahoo! and Keyword was "new film videos"
THEN	The propensity to buy Product Y is 65%
or	
IF	They buy Product A
THEN	The propensity to buy Product X is 74%

Amazon/AlmondNet
Exhibit 1009
Page 213

Configuring Your Server

There are a variety of settings for UNIX and NT servers that need to be set in order to optimize the amount of customer information you are capturing. These settings involved various server software, including the NCSA, the Apache, the CERN, the Netscape, and the NT WebSite. Because there are many settings, some of which are likely to change with each new version of the software, I will not go into detail on the configuration of each. I will note, however, some of the more important settings, which you are likely to want to ensure are correctly set either by you or your technical team. UNIX servers are configured via files containing directives that control basic settings, such as authorization and paths. Netscape is configured through HTML forms, while the WebSite uses a series of dialog boxes and wizards.

Server Configuration Tips

All servers can be configured to produce user agent and referrer log output. For example, to configure this on the popular Apache web server, the following settings are needed:

1. Make sure the mod_log_config Apache "module" is compiled into your server.
2. Add the following line to your httpd.conf file in the Apache conf directory:

```
FOR APACHE 1.2 OR BETTER, WITH THE MOD_USERTRACK MODULE:
LogFormat "%h %l %u %t \"%r\" %s %b \"%{Referrer}i\" \"%{User-
    agent}i\" \"%{Cookie}n\""
FOR APACHE WITHOUT THE MOD_USERTRACK MODULE:
LogFormat "%h %l %u %t \"%r\" %s %b \"%{Referrer}i\" \"%{User-
    agent}i\""
```

3. When you have done both of these things, restart the Apache server on the fly using the kill -1 Unix command.

The Apache server will now log in the usual "common log format," with the addition of two new fields, referrer and user agent, and a third "cookie field" uniquely identifying the user if you have compiled Apache with the mod_usertrack module installed. This is an example of the common log format, extended to include referrer and user agent information.

```
foo.bar.com - - [02/Jul/1998:16:48:44 -0744] "GET /webminer/index.html
    HTTP/1.0" 200 5898 "http://www.altavista.digital.com/query"
    "Mozilla/3.01 (Macintosh; I; PPC)"
```


Amazon/AlmondNet
Exhibit 1009
Page 214

Apache and NCSA Server Configurations

These directives should be checked in the *httpd.con* file:

Settings	Function
AgentLog *filename*	Specifies the location of the log file that identifies the client program used for each request, such as: `AgentLog logs/agent_log`
ErrorLog *filename*	Specifies the location of the error file, such as: `ErrorLog logs/error_log`
DNSMode *options*	Controls under what conditions the server will attempt to do reverse DNS lookups of the client browsers, the options are:
None	No DNS lookup.
Minimum	DNS lookup only for resolving access permissions.
Standard	DNS lookup for every request (this is the default).
LogOptions *options*	Specifies how log information should be written to the various log files. The options are as follows:
Separate	Agent and refer log files will be split into two separate files as specified by the AgentLog and ReferLog directives.
Date	Puts a date stamp on both the separate agent and referrer logs.
Combined	Combines and appends both the agent and referrer fields in the transfer log (recommended).
ReferrerLog *filename*	Specifies the location of the referring URL log file. It can be set either as an absolute path or as a relative path from the ServerLog directory, the default is: `ReferrerLog logs/referrer_log`
Script *method cgi_script*	Specifies the type of scripts which may be executed. The methods are: GET, POST, PUT, or DELETE.
TransferLog *filename*	Specifies the location of the transfer log file, either as an absolute path or a relative path to the ServerRoot directory. The default is: `TransferLog logs/access_log`

Amazon/AlmondNet
Exhibit 1009
Page 215

CERN Server Configurations

The following group of directives govern the logging in CERN used to specify the location of log files and their configurations:

Settings	Function
AccessLog *filename*	Specifies the location of the access log file, as an absolute or relative path from the ServerRoot directory. The common default setting being: `AccessLog logs/http:log`
ErrorLog *filename*	Same path and location as the AccessLog, most common setting is: `ErrorLog logs/http:error`
LogFileDateExt *suffix*	This directive specifies the extension to be used for log file names, such as time date format. `LogFileDateExt %H:%T`
LogFormat *format*	Specifies the format of the transfer log: Old, Common, or New.
LogTime timezone	Specifies the timestamp for the log files: `LocalTime or GMT`
NoLog *expression*	Specifies no logs for selected directories: `NoLog /for_your_eyes_only`

Netscape Server Configurations

Netscape uses the AddLog directive to set the transaction log and record the user-agent for server transactions. The options are as follows:

common-log	This records transaction data in the common NCSA log format, with the following parameters: Name=*logname* This specifies the name of the log to use and it must be the logfile name that you initialized with Init fn=init-clf name=... at the top of the *obj.conf* file. Iponly=*n* This parameter sets the server not to perform host-name lookup. Any value will do, since the Server Manager by default creates iponly="1".
flex-log	This records transactions in the extended or flexible logging format. This format is also set in the Int directives with the flex-init function. It uses the same parameters as the common-log function.

Amazon/AlmondNet
Exhibit 1009
Page 216

WebSite Server Configurations

The Server Admin tool is used for administering the WebSite server. WebSite uses the Windows 95 or Windows NT Registry to maintain information about the server. From the *Properties* section of the server's Control menu the configuration for several components can be set, including *Logging*, which specify the location of logging files, the type of access file format, and tracing options for the server log. WebSite has three main logs:

Error Log	This log records access errors, such a failed authentication.
Server Log	This log records each time the server is restarted or a new configuration is updated.
Access Log	This log records each request made to the server and its response in one of three formats: *Common Format:* This is the older NCSA/CERN standard web log format used by most servers. It is the default setting. *Combined:* This format uses a more practical web log format with fields delimited by quotation marks. This format incorporates two additional fields, the referrer and user_agent information. *Window (WebSite Extended):* This format is designed to collect data in a format that can be easily imported into Microsoft's *Access* and *Excite* packages. The data is tab delimited, which tends to create larger files.

The Multiplicity of a Visit

A single visit to your website might look like the following in your server log file:

```
204.58.155.58 - [06/Jul/1999:11:27:45 -0900] "GET /products/ad.html
    HTTP/1.0" 200 2887
```

However, any single visit is actually registered as multiple "hits" all dependent on the number of graphic and Java programs residing on that single ad.html page. Thus the above visit may actually look like this in your log file:

Amazon/AlmondNet
Exhibit 1009
Page 217

```
204.58.155.58 - [06/Jul/1999:11:27:45 -0900] "GET /products/ad.html
    HTTP/1.0" 200 2887204.58.155.58 - [06/Jul/1999:11:27:45 -0900] "GET
    /images/hotstuff.gif HTTP/1.0" 200 327
204.58.155.58 - [06/Jul/1999:11:27:57 -0900] "GET /images/sale.gif
    HTTP/1.0" 200 129204.58.155.58 - [06/Jul/1999:11:27:59 -0900] "GET
    /images/today_only.gif HTTP/1.0" 200 281
204.58.155.58 - [06/Jul/1999:11:27:59 -0900] "GET /images/ad2.gif
    HTTP/1.0" 200 350204.58.155.58 - [06/Jul/1999:11:27:59 -0900] "GET
    /images/and_more.gif HTTP/1.0" 200 275
```

As we found out, the host field in the Transfer Log File can be in the format of the IP address or domain name of the server accessing your website. The reason domain names may not be included in the log files is the extra processing and bandwidth required for looking up the name, which can be obtained via the nslookup command.

Aside from counting visitors to your website, the most important information you want to obtain from the host field is the number of unique *sessions*, determined by the different and unique IP addresses in your access log file. The IP address of visitors is the single best method of determining the number of individuals accessing your site. However, this can be misleading due to the fact that dial-up users are usually assigned "dynamic" IP addresses by their Internet service providers (ISP) from a reserved lot. Also, users of large online services like America Online access the Web from a limited number of "gateways" with their own IP addresses. Lastly, corporate users typically connect through proxy servers, which means groups of individuals use the same computer.

This means that it is very difficult to identify an individual with a unique IP address. This fact leads to two conclusions: (1) sessions are more important than individual addresses, and (2) tracking by other means is necessary whether by cookies or forms.

One of the patterns you may be interested in discovering is the path taken by your website visitors and customers both from outside and inside your website. As we learned, these types of patterns can be revealed from two different sources. To find out how web visitors got your site and what "keywords" were used at a search engine, the referrer log can be used. The referrer log tells how your website visitors found your website and what search criteria they used at Yahoo!, InfoSeek, Excite, etc.

Cookie Files

To find out what paths visitors took while *in* your website, a form of tracking will need to be implemented. This can be done by setting up cookies or some other type of CGI program. *Cookies* are small text files

Amazon/AlmondNet
Exhibit 1009
Page 218

created by servers on visiting browsers' hard disks that contain an identification code. They are passed by servers to browsers so that each time a website visitor returns to a server that passed the cookie it can recognize it and read what it wrote before, such as what pages were visited during the prior session. Netscape developed cookies as a means for servers to identify a specific browser, with the specification that only the server that provided the cookie can be the one that can access it and retrieve it. This means that identification cookie codes sent to your cookie.txt file by another server could not be accessed by a different website. Cookies unfortunately are limited to identifying browsers, not individuals.

Every visit by your customers to your website is a separate event, totally unconnected or related to any previous connection or event. Cookies are a technique by which you can make these unrelated events link into a semblance of a real-world transaction, such as the shopping experience. In fact, cookies were developed by Netscape software developers in an effort to create a shopping cart application. The shopping cart analogy works this way: it allows shoppers to buy more than one item at a session as they wander through the store (website) and pay at the cash register (form) as they leave the store. What is also important is that the store now has a record of which person went where in the store and what items were purchased. Similar to the way you order from a cataloger, a website using cookies is able to record what items you like and purchase over time and how you pay for them.

Cookies are a hack solution to a problem of "statelessness" caused by the nature of the Web and its protocol, the Hyper Text Transfer Protocol (HTTP). HTTP allows for communications to take place instantaneously without the need to maintain open connections, which may not be needed again. An effective client/server design, but poor retailing environment—hence the need to maintain "state"—cookies. (See Figure 6–1.)

The usual way cookies work is when a CGI script identifies a new user, it adds an extra header to its response with an identifier for that web visitor and other information that the server may capture from the client's input. This new header data informs the browser to add this information to the client's cookie file. In Netscape it is added to the cookie.txt file; in Internet Explorer, cookies are kept in a cookie directory under the Windows directory, associated with the username taken from the login name for the operating system (e.g., "mena@excite.txt").

Any subsequent request to that URL from the browser will include the cookie information as an extra header in the request. The CGI script uses this information to return a document targeted to that specific web visitor or returning customer. Since the cookie information

Amazon/AlmondNet
Exhibit 1009
Page 219

Figure 6–1 *A Sample Cookie.txt File*

A Sample Cookie.txt File

```
# Netscape HTTP Cookie File# http://www.netscape.com/newsref/std/cookie_spec.html# This is a
generated file!  Do not edit.
.excite.com    TRUE    /       FALSE  946641600         UID
    867CE396354734CB.preferences.com          TRUE    /        FALSE   1182140421
    PreferencesID       XHphq2JpN264wr8hBTHkSG.focalink.com       TRUE    /      FALSE
    946641600           SB_ID   089385908100001462311090540052.go2net.com         TRUE  /
    FALSE     917824360         preferences     32qQ00hCOAw89U.infoseek.com       TRUE  /
    FALSE     925398460         InfoseekUserId  7BBEDA1C15F0696DCFB611F729A5C544.yahoo.com
    TRUE      /         FALSE   915145200       Y         v=1&n=dnnef14bek7pl.doubleclick.net
    TRUE      /         FALSE   1920499140      id        285bff29www.cdnow.com   FALSE /
    FALSE     942189160         cookTrack       283498007-893877542.netscape.com  TRUE  /
    FALSE     946684799 NETSCAPE_ID     10010408,11fc8400
```

is stored on the client's hard disk, the information remains even after the browser is closed and the computer is turned off. One of the most common methods of using cookies is to "recognize" prior visitors through preassigned headers, enabling the server to present "fresh" content to that visitor or customer, which they have not seen before. Cookies cannot be over four kilobytes in size, and only the issuing domain name server can have access to them. There can be no more than 100 cookies on your hard disk.

Here is the information stored by a typical cookie:

```
.snap.com  TRUE  /  FALSE  946684799  u_vid_0_0  00ed7085
```

.snap.com	This is the domain for the cookie. This by default is usually the server that created and issued the cookie to the client. Only pages from the original server are allowed to read that cookie. This means snap.com cannot read the cookies issued by yahoo.com or infoseek.com.
TRUE	Indicates whether the cookie was set by an HTTP header (TRUE) or a JavaScript (FALSE).
/	This path is variable, allowing for any page in the website snap.com to access this cookie. Same as the domain, it restricts access to cookies to those that created it. Cookies without a set / path will be deleted when the user exits the browser.
FALSE	Indicates whether it is a secured cookie or not. There is an option for encryption using HTTPS, SSL. This attribute has no value; unless it appears when created, the cookie will travel without any security.

Amazon/AlmondNet
Exhibit 1009
Page 220

946684799	The expiration date, expressed in seconds since midnight GMT, January 1, 1970. Another, different format has the expiration date set to Greenwich Mean Time (GMT) in the format "Wdy, DD—Mon - YYYY HH:MM:SS GMT", where Wdy is Weekday (optional), DD is Day, Mon is Month, YYYY is Year, HH is Hour in 24-hour format, MM is Minutes, and SS is Seconds.
u_vid_0_0	This is the critical name which identifies the cookie. The name of the cookie cannot contain commas, semicolons, or white space. The latest cookie replaces the older cookie with the same domain, path and name.
00ed7085	This is the value of the cookie, where the issuer of the cookie stores its information. It cannot contain commas, semicolons, or white spaces.

A Cookie Future

As the Web increasingly becomes a retailing channel of choice for millions of users, a new proposal has been advanced for improving cookies and through it online shopping, relational marketing, and electronic commerce. A Request for Comment 2109 proposes "a way to create a stateful session with HTTP requests and responses. It describes two new headers, Cookie and Set-Cookie, which carry state information between participating origin servers and user agents." Basically RFC 2109 proposes a Version 1 of cookies which will extend their abilities and provide extra information that will make it easier for users to manage them. This new version of cookies will expand their structure to include the following components:

Name. Same as Version 0, a cookie name is required. Except that for this version it cannot begin with the dollar sign ($), all other restrictions apply: no commas, semicolons, or white space.

Comment. This is a new attribute which allows the issuer of a cookie to tell the browser recipient the purpose of the cookie: *"This cookie* 'group_discount_cookie A1' *is being set to identify you as a frequent and valued customer entitling you to special discounts."* This optional feature allows the website server to tell the recipient browser the purpose of the cookie. If the user has set the browser to ask the dispensing server for the purpose of the cookie before accepting it, a message will appear in a browser dialog box. The user can inspect the information to decide whether to initiate or discontinue a session with the server.

Amazon/AlmondNet
Exhibit 1009
Page 221

CommentURL. Similar to the comment field but with more detailed information on the purpose of the cookie and the processing of them at the server level.

Domain. Although optional, it would work the same as Version 0 with a twist. The domain attribute still specifies the domain for which the cookie is valid, which is usually the dispensing site. However in Version 1 a subsection domain name can be set. For example, the domain *www.isuzu.amigo.com* could specify a cookie for the subdomain *.amigo.com.* An explicitly specified domain must always start with a dot, the same as in Version 0.

Discard. This attribute mandates that cookies be discarded no matter what the maximum age of the cookie may be when present.

Max-Age. The Max-Age attribute defines the lifetime of the cookie, in seconds. The delta-seconds value is a decimal nonnegative integer. After delta-seconds elapse, the client should discard the cookie. A value of zero means the cookie should be discarded immediately, which means that cookies could be discarded during a transaction.

Path. Same as Version 0, the Path attribute specifies the subset of URLs to which this cookie applies. If a path is specified, only URLs that contain that complete path / are allowed to read or modify the cookie.

Port. The Port is an additional feature of the new cookie standard which can be specified for specific system applications, such as FTP for 21, Telnet 23, and web server on port 80.

Secure. The Secure attribute (with no value) directs the user agent to use only (unspecified) secure means to contact the origin server whenever it sends back this cookie.

Version. This attribute is required and should be set to Version=1. The Version attribute, a decimal integer, identifies to which version of the state management specification the cookie conforms.

Cookie Applications

One of the most common uses of cookies is for the tracking of visitors as they navigate through a website. By collecting a visitor's cookie every time a page is requested, say for a different product or service in your site, specific patterns or paths can be mined to determine what parts of your website are the most popular. Also, you may begin the process of profiling your online customers by linking cookies to referrer log and demographic data or registration information on your visitors and customers. This profile can then be used for subsequent analysis and marketing applications. Cookies can also be used

Amazon/AlmondNet
Exhibit 1009
Page 222

to record what type of products and services a visitor has seen and purchased so that subsequent visits can be personalized to match their specific tastes and interests. Cookies can be used to make the visitor feel welcomed and important.

To give you an idea of how cookies work and how they can be used to enhance the value of mining log files, follow this shopping site scenario. When a visitor enters your website, your server issues an identifying user number (USERID=1234) in the following response header:

HTTP/1.1 200 OK
Server: CERN/3.0 libww/2.17
Set-cookie: USERID=1234; path=/
Content-type: text/html
Content-length: 2742
<Entity body>

This tells the browser to save the contents of the cookie, the visitor identification number (USERID=1234) and associated it with the issuing server that sent it. In this example, the server that sent the cookie is www.store.com, meaning the browser associates this cookie exclusively with store.com. Netscape Navigator will store this header identification number in the cookies.txt file, while Microsoft Explorer will put it in a cookie directory. The next request that the browser makes to the store.com server will include a request header that contains the content of the cookie. Here is how it might look:

```
GET / HTTP/1.1
Cookie: USERID=1234; path=/
If-modified-since: Fri 12 Jun 1998 10:57:22 GMT
Referrer: http://store.com/index.html
User-Agent: Mozilla/4.0 (Windows 95)
```

Once the cookie is set, the browser returns a "Cookies" header. Thereafter, the server doesn't need to send any more, except to set additional cookies for special purposes, such as identifying this as a new visitor or for observing what areas of the website this customer wanders into, what this visitor views and asks information about, and of course what she puts in her shopping cart and purchases. Through these streams of ongoing transactions the exchanges of cookies are able to maintain a semblance of continuity that is absent in the HTTP protocol.

Amazon/AlmondNet
Exhibit 1009
Page 223

A Cookies Scenario: The Transactional Exchange of HTTP Requests and Responses

1. BROWSER: Request Headers
   GET /index.html HTTP/1.1
   If-modified-since: Fri 12 Dec 1998 12:34:22 GMT
   Referrer: http://www.kdnuggets.com/companies.html#W
   User-Agent: Mozila/4.0 (Windows 95)

2. SERVER: Response Headers
   HTTP/1.1 200 OK
   Server: CERN/3.0 libwww/2.17
   * Set-cookie: USERID=1234; path=/
   Content-type: text/html
   Content-length: 3284
   <Entity body>

3. BROWSER: Request Headers
   GET /HTTP/1.1
   * Cookie: USERID=1234; path=/
   If-modified-since: Fri, 12 Jun 1999 12:35:23 GMT
   Referrer: http:/www.webminer.com/index.html
   User-Agent: Mozilla/4.0 (Windows 95)

4. SERVER: Request Headers
   HTTP/1.1 200 OK
   Server: CERN/3.0 libwww/2.17
   Content-type: text/html
   Content-length: 2149
   <Entity body>

5. BROWSER: Request Headers
   GET /products/info.html HTTP/1.1
   * Cookie: USERID=1234; path=/
   Referrer: http:/www.webminer.com/index.html
   User-Agent: Mozilla/4.0 (Windows 95)

6. SERVER: Request Headers
   HTTP/1.1 200 OK
   Server: CERN/3.0 libwww/2.17
   Content-type: text/htmlContent-length: 4951
   <Entity body>

The dispensing and tracking of cookies continue until the visitors select the check-out form signaling they are ready for a secured transaction and payment of goods and services. Cookies can store user interests via a unique identification tag, which websites can correlate with a database on the web server or advertising network. This also means that each time a customer visits a website, the server looks for the associated information on clickthroughs, transactions, and registration entries. Cookies can be used to target and rotate ads, for example, so that when you visit a site like your favorite search engine, you don't get

Amazon/AlmondNet
Exhibit 1009
Page 224

hit with the same banner ad. The limitations of cookies, however, are obvious: they track browsers, not people—the cookies in a customer's office PC may be quite different from the iMac at home.

Aggregate cookies can be used for mining visitor patterns while they maneuver through your website, enabling you to make it convenient for them to return again and again. The experience should be as similar to going into a small corner shop where the clerk remembers what you like and how you like it. As part of your privacy policy you should disclose how you are using cookies—to improve your visitors' navigating experience and content value—while ensuring your integrity. As part of this policy you should also use cookies only when needed and for as short a period of time as is needed to accomplish your task. If, for example, you only need to issue cookies for single session tracking, then you should destroy your cookies at the end of your visitors' browsing session. The objective of mining cookies should be to discover online tendencies and to reinforce and personalize your customers' experience every time they visit your website.

Configuring Cookies

In the mining of web data it is important that website activity be captured for analysis, which means configuring your server to write to the Transfer Log file the cookie as well as the referrer and agent information you will need for your data mining analysis. For the Netscape servers this is a simple matter of selecting a configuration option to log cookie information in the transfer log. For Unix servers such as the popular Apache, you use a source code module for implementing cookies. You need to modify the log-reporting module to include the cookie information on each line of the transfer log. This is how you configure your server LogFormat string in the source code to write the referrer, agent, and cookie information:

```
LogFormat "%h %l %u %t \"%r\" %s %b \"%{Referrer}i \" \"%{User-
    agent}i"\"%{Cookie}i\""
```

For Unix servers these cookie-dispensing modules are in C code, which must be compiled prior to integrating. For NT servers the modules are in the format of DLL files, which must be updated on your registry. However, since cookies are set simply by adding an additional response header, they can be done through CGI scripts or other Java or Javascript programs. These "virtual cookies" work well in shopping cart applications, where a purchase selection for a specific product sets up a cookie until a purchase transaction is com-

Amazon/AlmondNet
Exhibit 1009
Page 225

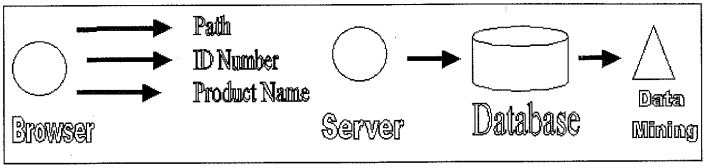


Figure 6–2 *Important customer information can be stored and mined for analysis.*

pleted. The final CGI or Java program to process the purchase will reset the cookie to clear out all information. This application, however, does not write a cookie to the log file, which is really what is of interest in a data mining analysis.

CGI and Java can also be used to track visitors dynamically. For example, a CGI program can be used to create and send HTTP headers to set cookies. Similarly, CGI can be used to dynamically serve specific content based on the type of product a user selects. Numerous websites offer the code for doing this. Aside from serving content dynamically, what is of more relevance to the analysis of your website traffic is the ability to use the same type of CGI programming to issue your visitors a unique identification number, with the objective being to code that number into every link on that html page as an extra argument. This way every time a visitor clicks on a link in that page, their browser sends the path and product name of the file they are requesting with their unique tracking user identification number (Figure 6–2).

Cookies and CGI

Although cookies can be created via several types of technologies, such as VBScript, JavaScript, and Java, both on the server and client-side, most of today's cookies are created using CGI.

The CGI provides a set of web standards that allow the servers and programs to communicate with clients. It takes the information from the browser in an HTTP request format and passes it to the program as a set of variables. When the server receives and processes the request, it parses the headers and transforms them into a list of variables that is sent to the CGI program as standard input when the program first begins. One of the most important uses of forms is the interaction with website visitors and customers. The residue of these transactions is important customer information that is ideal for data mining.

A CGI script consists mainly as a translation program between a browser and server, matching outputs to inputs. When a request is

Amazon/AlmondNet
Exhibit 1009
Page 226

submitted, the CGI script parses the information it receives, extracting the data such as a query string from a registration form sent by a POST request or cookie, which based on a pre-set routine, can generate a response for the website visitor, usually a set of headers followed by some HTML. This response is passed back to the server when the program concludes, with the server passing the response back to the visitor. Every time a visitor makes a call to a CGI program, the server builds the CGI request and runs that program—usually as a separate application. CGI program can be written in any language, although a large percentage of them are written in Perl, a language good at managing arrays and text.

The following are some quick examples of cookie settings, so that if you want to reach the HTTP_COOKIE variable for your own direct manipulation you can use either cgi.pm's raw_cookie() method or $ENV{'HTTP_COOKIE'}.

Retrieving a cookie value can be done within a CGI script with the following statement:

```
$myvalue=cookie('mycookie');
```

or

```
$myvalue=cookie(-name=>'mycookie');
```

Setting a cookie involves creating a cookie name and value:

```
$myCookie=cookie(-name=>'mycookie',-value=>'myvalue');
```

or

```
$myCookie=cookie('mycookie'.'myvalue');
```

This will create a cookie, which will expire at the end of the session, while the following call will last for a year before expiring:

```
$myCookie=cookie(      -name=>'mycookie',
                       -value=>'myvalue',
                       -path=>'/'
                       -domain=>'mydomain.com'.
                       -expires=>'+1y'
                       -secure=>1);
```

This is an example of how a cookie may be set up for first-time visitors. Looking for a firstVisit cookie, the CGI script, using Perl, will first check to see if a cookie is set before dispensing one to the visitor:

Amazon/AlmondNet
Exhibit 1009
Page 227

```
#!/usr/bin/perl
use CGI qw(:standard) :
# check for the firstVisit cookie
if (cookie('firstVisit')) {
    #if one exists, this is a returning visitor
    #collect the cookie value
    $firstVisit=cookie('firstVisit');
    $first="not ";
    #print header with no cookies.
    print header();
} else {
    #set firstVisit cookie
    $firstVisit=localtime;
    $cookie = cookie(-name=>'firstVisit',
            -value=>$firstVisit,
            -expire=>'+2y',
            -path=>'/');
#print header with firstVisit cookie
print header(-cookie=>$cookie);
$first="";
}
$currentDate=localtime;
#use cgi-pm's method for creating HTML.
print start_html("First Visit Check"), h1("Visitor Status");
print p("Welcome $first to The Online Shop.");#use Perl's .
    concatenation operator to split
#print p("Check out our daily sales");
print hr();
print p("It is now $currentDate.");
print end_html();
```

Open Profiling System (OPS)

The proposed Version 1.0 of cookies RFC 2109 would allow users more control over what cookies get passed to their browser and would expand their ability to reject cookies. It would also require website administrators and developers to explain why they are issuing cookies to their visitors and customers. Furthermore, the Open Profiling System as proposed by Netscape, Firefly (now owned by Microsoft), and VeriSign would set new standards designed to encourage users of new browsers to provide personal information as part of their setup options. At the same time it would provide them expanded control over who would be allowed access to their personal, home, and business information. OPS provides for the preloading of personal information at the browser level, one time only, then it allows the user to share that information only at those websites where they want to establish a one-to-one relationship—for personalized products, news, and services. The following is from the opening section of the OPS standards submission:

Amazon/AlmondNet
Exhibit 1009
Page 228

Introduction

As more and more users, and more and more services come to the Internet, users are finding themselves increasingly overwhelmed with the richness of new possibilities, and new service providers are discovering that rapidly building systems to fit these users' information needs becomes ever more challenging.

In general, a solution to these problems often involves delivery of highly customized and personalized information to end users, which raises the profound and valid concern about making and keeping explicit commitments to users about how their most sensitive personal information, choices, preferences, and interest will be protected in these exchanges.

Companies and service organizations worldwide want to take advantage of the 1-to-1 nature of communications on the Internet or within intranets to provide their customers, employees and visitors with individualized information, entertainment and services. However, there are two barriers to the feasibility and widespread adoption of such products and services:

1. *The potential threat to individual privacy makes end users wary of sharing any information. Today, there are few measures or safeguards that offer an end user any awareness of or any control over the usage of his or her personal information. This concern often outweighs the incentive of a personalized service and makes a person justifiably cautious about revealing any personal information.*

2. *Gathering the information that makes this personalization possible is inefficient. Service providers need to ask their visitors for information—who they are, where they live, what they do, what they like—in order to personalize the user experience. This data collection can be very time consuming for both the service provider and the end user, and can't be leveraged for their mutual benefit. Furthermore, a single individual might provide much the same information to dozens or even hundreds of parties over time—an inefficient and frustrating experience.*

The Open Profiling Standard (or OPS) is a proposed standard for exchanging profile information between individuals and service-providing parties, with built-in safeguards for individual privacy.

The OPS basically proposes the use of a Set-Profile header similar to the current Set-Cookie header, but with extensively much more information. Instead of a header with a few thousand bytes of name-value pairs for a single site, the Set-Profile header would contain a

Amazon/AlmondNet
Exhibit 1009
Page 229

number of standardized groups of fields (with user permission) to multiple commercial sites with various security levels. The following is from the Netscape press release at the time OPS was announced:

> *Personal Profiles contains a wide range of descriptive information about an individual, such as name, address, ZIP code, phone number, and e-mail address. Personal Profiles can also contain information on an individual's age, marital status, interests, hobbies, and user identification/password. Internet sites and software products that support OPS and Personal Profiles will enable users to be aware of and in control of what profile information they divulge to Internet sites. In addition, OPS enables users to grant permission to specific Internet sites to share their PersonalProfiles with other Internet sites.*

Clearly OPS is aimed at personalizing the Web and making online transactions transparent to the website visitors and websites. The personal information or identity of visitors and customers would reside with the browser settings, much like the setting of your font preferences and other options. OPS would eliminate the need to complete registration forms and the constant retyping of personal information and preferences. For the website administrator it would make available a far greater amount of detailed demographic data for mining, and the enabling of responsive one-to-one service and customer customization.

Browsers such as Microsoft's Internet Explorer already include new options that allow for the setting and storing of user information at the personal, home, and business levels (Figures 6–3 through 6–5).

New server software is also enabling web administrators a variety of methods for personalizing their sites, while at the same time capturing important information about their visitors' and customers' usage patterns. For example, new server software like the Microsoft Personalization System (MPS) allows administrators to capture and store web visitor information between sessions linked by cookies. MPS also allows developers to create globally unique identification cookies (GUIDs) that are identical in multiple websites, allowing for the sharing of cookies between multiple domains. This feature allows for the "networking" of cookie information. The way this feature works is when a visitor appears at a server using this feature instead of sending a cookie, the MPS redirects the user to a different server. This second server then checks for a cookie, assigns a cookie if required, and bounces the visitor back to the original server with a message for that server about the cookie it set up. The original server then assigns a cookie that matches the cookie set by the master server. The result is a set of persistent cookies for two different domains with the same value.

Amazon/AlmondNet
Exhibit 1009
Page 230

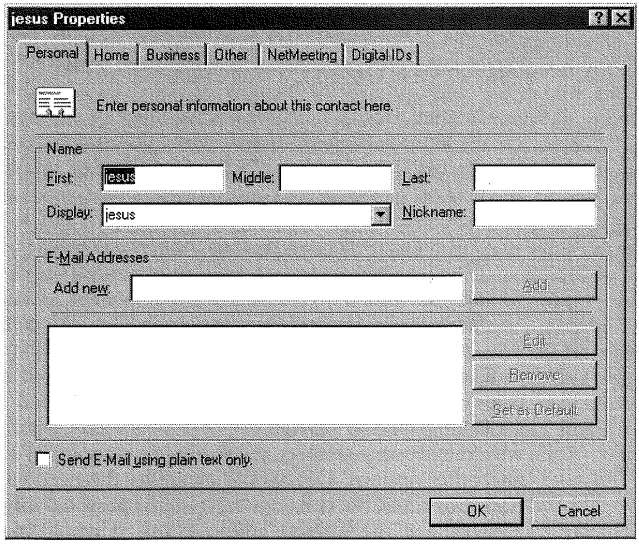


Figure 6–3 *Personal information can be set in Microsoft Internet Explorer.*

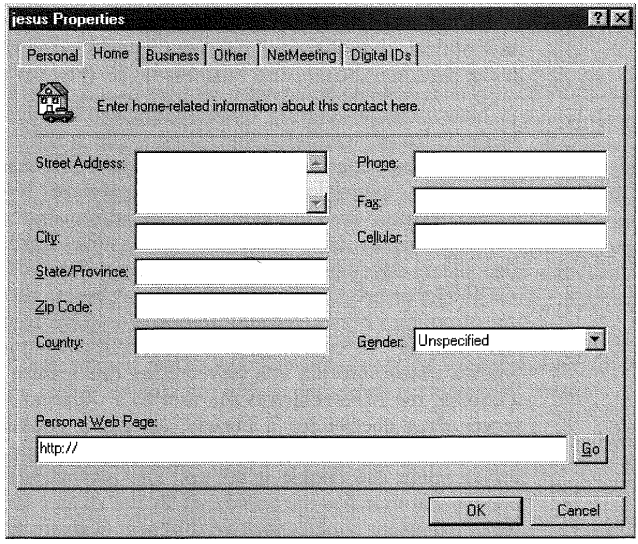


Figure 6–4 *Home information can be preset on Microsoft Internet Explorer.*

Amazon/AlmondNet
Exhibit 1009
Page 231

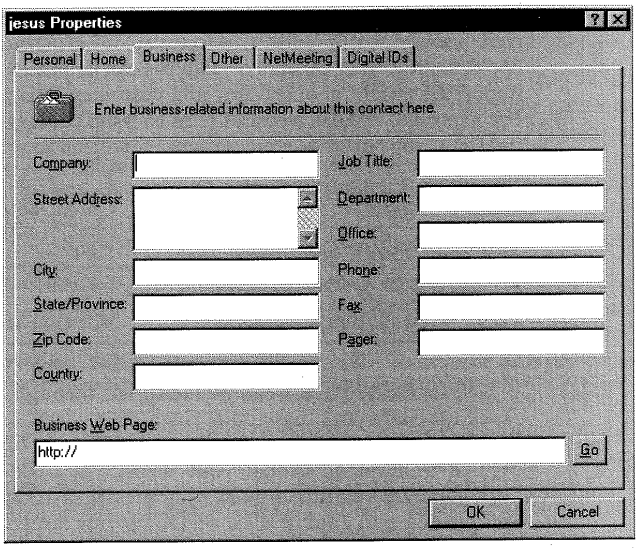


Figure 6–5 *Business information can also be set for Explorer.*

In essence this defeats the concept originally established by Netscape, which limited the setting, viewing, and modifying of cookies to only the original domain issuer. This Microsoft feature in effect allows for the networking of cookies via multiple servers.

The Commerce Server tools from the Site Server Enterprise edition also use cookies to keep track of shopper accounts, affording returning customers convenient access to account information and personalized messages captured from previous transactions. MPS provides for cookies to be set for tracking website visitors quite easily; allowing for the storing of personalized information by specific customers in the PropertyDatabase server component. The Microsoft User Properties Database Object has a number of properties and methods that an administrator can use to add information. Loading the UPD information into a database facilitates the linking, massaging, and mining of the data.

Although cookies are an excellent way to track individual sessions by browsers at the server level, they provide little value in data mining, unless they are used in aggregates. This is because cookies generally represent an individual identification number, which like your social security number is useless in a data mining analysis of a customer database. A social security number, like an individual cookie user identifi-

Amazon/AlmondNet
Exhibit 1009
Page 232

cation number, is a single observation that is too unique to be generalized into a useable pattern. However, by grouping certain types of cookies, important clusters or patterns can be mined, which can lead to important insight about visitor and customer activity in your website. For example, retailer websites tend to issue cookies in several sequential steps, for purposes of identifying first-time shoppers:

Website Activity	Cookie Settings Issued
Visitors arrive at the site's home page	=on
Visitor selects a product category	=1stVisit=Welcome
Visitor selects a specific product	=Session_ID=1BABV317MEN

Sequential segments can be identified through the use of various types of data mining tools, including decision tree programs designed to discover key clusters in databases. The important concept to understand is that your website has the unique ability to both generate relational marketing messages and information when you use log and cookie files, or CGI forms designed to capture visitor profiles.

Forms

By far the most effective method of gathering website visitor and customer information is via registration forms. However, care must be taken in asking only for the most essential information; nobody likes to complete lengthy and intrusive questionnaires, especially those pertaining to income and other personal information. Keep in mind that there are various optional methods and sources of gathering demographic information rather than asking for it directly from your customers. In any event, asking too many nosy questions will only ensure that you will get no information or intentionally incorrect information from respondents.

Server processing of form submissions can take several directions, the most common of which are the launching of a CGI program that executes and returns a response to the website visitor. This is usually some form of response to the website visitor, such as some text or HTML. However, from this simple request and response mechanism a complex array of website and customer interaction and relationships can evolve. The most important is the circular feedback from consumer to retailer and the gathering of real-time vital client information. The transmission of information back and forth between client

Amazon/AlmondNet
Exhibit 1009
Page 233

and server in hidden form fields establishes a relationship with your website visitors and customers and can track their browsing actions.

The two most important HTML tags are <FORM> and <INPUT>. The <FORM> tag is used to map the section of the HTML file to be used for user input. It is how most HTML pages call a CGI script—the tag tells the program the type of encoding being used and the method being used to transfer the data:

```
<FORM ACTION="url" METHOD=[POST|GET] ENCTYPE="...">
```

The ACTION attribute references the URL of the CGI program so that after the web visitor completes the form, all of the information is encoded and passed to the CGI script for decoding and processing. The METHOD attribute tells the CGI script how to receive the input: GET or POST (in most forms it will be POST). The POST method is the recommended way of getting information from forms since it sets no size limitations. Forms use text-input bars, radio buttons, check-boxes, and other means of accepting input by using the <INPUT> HTML tag. This is a sample of a text-input field in a form:

```
<INPUT TYPE=text NAME="..." VALUE="..." SIZE=...MAXLENGTH=...>
```

NAME assigns the value entered by the user. VALUE places the default text in the input field. The SIZE specifies the length of the input field, MAXLENGTH sets the maximum number of characters allowed to be input. To submit the information the following tag is used:

```
<INPUT TYPE=SUBMIT>
```

When the web visitor has completed the form they can submit its content to the URL specified by the form's ACTION attribute when the user clicks on the *Submit* button. When the content of the form is sent, the browser encodes the information before uploading it to the server and the CGI application. Once the user enters the values for every field, the symbolic names are passed with the following specifications:

- Spaces are replaced with pluses (+). Fields are separated by an ampersand (&).
- Name and values are separated by equal signs (=), name on left, value on right.
- All "abnormal" characters are also replaced with a percent sign (%) followed by the two-digit hexadecimal character code.

Amazon/AlmondNet
Exhibit 1009
Page 234

An encoded string will result in this output:

```
name1=value1&name2=value2&name3=value3...
```

This is a sample of a very simple form:

```
<html> <head>
<title> Name, Age and Gender</title>
</head>
<body><form action="/cgi-bin/profile.cgi" method=POST>
Please provide us your name: <input type=text name="name"><p>Please
    provide us your age: <input type=text name="age"><p>
Please provide us your gender: <input type=text name="gender"><p><input
    type=submit>
</form>
</body></html>
```

If your website visitor enters his name as Joe Blow, age 34 and gender Male, the input will be encoded into the following string:

```
name=Joe+Blow&age=34&gender=Male
```

Form data consists of a list of name/value pairs. Before the browser transmits data to a server and a CGI program it encodes the information using a scheme called URL encoding.

```
name          Joe Blow
e-mail        jblow@doc1.com
```

First all non-alphanumeric characters (@) gets encoded to (%40):

```
name          Joe Blow
e-mail        jblow%40doc1.com
```

Next, it replace all spaces with plus signs:

```
name          Joe+Blow
e-mail        jblow%40doc1.com
```

Then, each name and value get paired with an equal sign:

```
name=Joe+Blowe-mail=jblow%40doc1.com
```

With the final encoded string of pairs separated by an ampersand:

```
name=Joe+Blow&e-mail=jblow%40doc1.com
```

After a CGI program receives the encoded form input, it parses the string and stores it so that you can use the data, which is in the structure of name/value pairs. This string can be manipulated so that, for example, pluses (+) can be replaced with spaces, commas, or tabs. URL-encoded characters can be translated back to their original char-

Amazon/AlmondNet
Exhibit 1009
Page 235

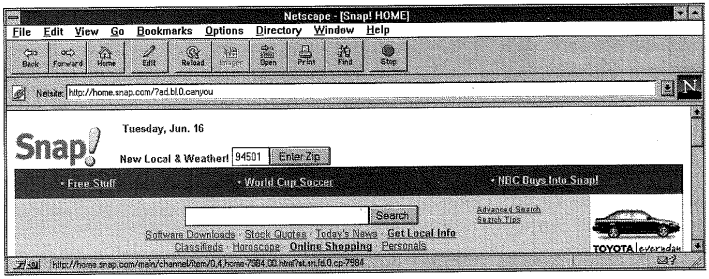


Figure 6–6
A simple one line form can provide targeted information to website visitors.

acters. The purpose of doing this of course is so that the encoded strings can be in a format suitable for merging with other internal and demographic databases and subsequently mined. Most of today's data mining tools are able to import text files delimited by commas or tabs.

```
name=Joe+Blow&age=34&gender=Male&jblow%40doc1.com
```

can become a comma delimited flat-file:

```
Joe Blow,34,Male,jblow@doc1.com
```

which is a format readily importable into most data mining tools.

A simple one-field form requesting your visitors' ZIP code for purposes of providing links to local classifieds, news, services, and weather can also be used to collect the all-important ZIP code (Figure 6–6). A ZIP code can be used to not only localize a website to meet the need of its visitors, but for matching with demographic and psychographic databases for subsequent data mining analysis.

Common Gateway Interface

The CGI is the most popular protocol by which browsers interact with web servers. It is a robust and versatile method by which visitors and customers communicate with online stores and other providers of products and services. CGI is a method by which web servers are able to interpret input from browsers and in some cases return information based on their input. The protocol defines a standard way for programs to communicate with the web server. CGI communications is handled over the standard input and output using any programming language, although Perl is the most common.

Perl is a high-level language that excels at parsing text, which is what most CGI scripts do. CGI programs usually involve getting

Amazon/AlmondNet
Exhibit 1009
Page 236

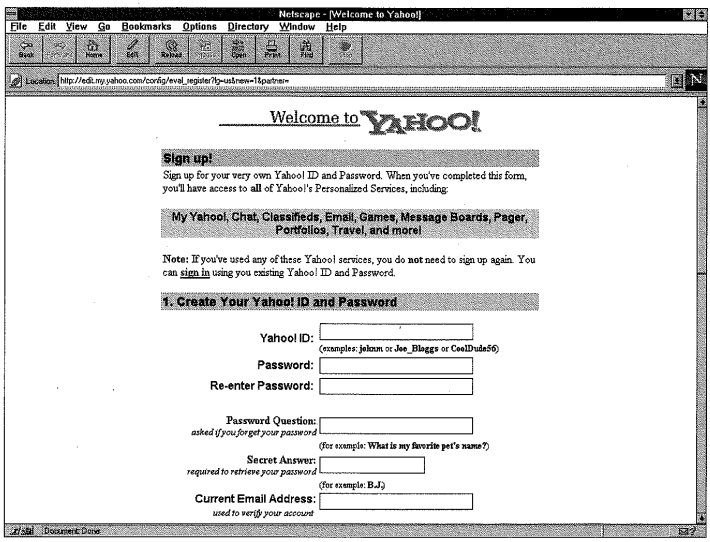


Figure 6–7
A form used to capture website visitor information for personalized content delivery.

information from the browser and sending it to the server. CGI programming can be reduced to (1) getting input from the browser, (2) processing that input at the server, and (3) sending that output back to the browser. The most common method of doing this is, as we discussed, through HTML Forms, which for data mining is an important component because unlike logs or cookies they are being provided by the website visitor or customer themselves. Often, the website visitor or customer is presented with a form to complete, such as their ZIP code, age, gender, preferences, and so forth (Figure 6–7).

Once the visitor or customer fills out the form and submits it, the information is sent to a CGI program, which processes the input, such as appending it to a database. Next the program will acknowledge or send something back to the browser, like a user identification or password. For example, in exchange for customer information you may want your CGI program to provide a code for accessing a special location in your website, or you may want to grant a special discount. The point is that via the use of CGI and the forms that they support, you are able to interact with your visitors and customers. More importantly you are also able to capture important information about who they are and what they want and need.

Amazon/AlmondNet
Exhibit 1009
Page 237

CGI programs can read and write to files and databases, allowing the capturing of important customer information at the server level. How you accomplish this can be done either by server side processing of compiled programs or scripts executed by an interpreter. Most commonly this is done using a scripting language such as Perl, although C and C++ can also be used as well as JavaScript or Java. The problem with using C and C++ is that the administration requirements increase in complexity. Programs need to be recompiled any time a modification is made. In a highly dynamic environment, such as a large website, this can be a problem.

Java and JavaScript provide an option to using forms to capture customer information with just the traditional CGI. Java is an object-oriented programming language, while JavaScript is an object-based scripting language with a set of built-in objects. JavaScript is an extension of HTML and allows for some of the processing to be shifted from the server to the client, thus reducing processing time and server load. Because JavaScripts are embedded in HTML pages, web visitors do not have to wait for servers to respond because when pages are received by the client the scripts immediately execute for faster response time. JavaScript can also be used on the server side with Microsoft's Active Server Pages and Netscape's LiveWire.

Assembling the Components

To take advantage of each of these data components you likely will need to use both log and cookie files in combination with forms. A possible method for doing this involves using both server-side CGI scripts along with client-side JavaScript. The basic JavaScript used to set a cookie for both Netscape Navigator and Microsoft Explorer can be embedded in an HTML file, and it may look something like this:

```
document.cookie = "myCookie=A Value; expires=Tuesday, 06-Jul-199908:00:00 GMT"
```

When a visitor accesses your website, their browser will load the cookie set forth in the document .cookie object as one long string. When you assign a value to this object, the browser updates the named cookie if it exists, or creates the named cookie if it does not exist. The name of this particular cookie is myCookie, the value being stored is set to "A Value," with an expiration date of July 6, 1999 at 8 am. In order to store a website visitor's name, age, and gender when they submit a form along with a cookie you can write a JavaScript to look something like this:

Amazon/AlmondNet
Exhibit 1009
Page 238

```
data = document.FormName.visitor_name+"|"+
   document.FormName.visitor_age+"|"+
      document.FormName.visitor_gender;
document.cookie =cookie_name+"="+data+
      "; expires="+expiration_date;
```

Through the use of CGI, Java, and/or JavaScript, you can link the tracking ability of cookies with the important customer-provided demographic information captured with a form to gain an understanding of *who* is going *where*. In incremental steps you can begin to assemble a composite of who your website visitors and customers are:

```
Log File Information:
ntt.com.jp - - [08/Jul/1999:16:44:14 -0744] "GET /webminer/index.html
    HTTP/1.0" 200 5898
    "http://search.yahoo.com/bin/search?p=data+mining+websites"
    "Mozilla/3.01 (Macintosh; I; PPC)"
plus Cookie File Information
Session_ID=1BABV317MEN
plus Form Information
Joe Blow,34,Male,94502,jblow@doc1.com
equals Visitor/Customer Record String
ntt.com.jp - - [08/Jul/1999:16:44:14 -0744] "GET /webminer/index.html
    HTTP/1.0" 200 5898
    "http://search.yahoo.com/bin/search?p=data+mining+websites"
    "Mozilla/3.01 (Macintosh; I; PPC)", Session_ID=1BABV317MEN, Joe
    Blow,34,Male,94502,jblow@doc1.com
```

This comma-delimited string can easily be imported into a database such as Access for pre-processing and preparation for data mining (Figure 6–8).

One of the key strategies for capturing customer information at your website is to identify certain demographic features from their ZIP, which can be commingled with commercial consumer and census data. Analysis of both external and internal patterns can provide you an insight into the type of consumers that you are attracting to your website. These market clusters are individuals who have a propensity to buy your products and services. Your analysis of the type of individuals attracted to your website can reveal some common psychographical profiles with common behavior and values. Through the analysis of your website data you will be able to identify individuals with an interest in what you are selling.

You may in fact discover different types of market segments for different parts of your website, especially if you have a large website. Because of the dynamic nature of the Web you can create and test different content and marketing campaigns and promotions almost on the fly. The testing of offers and marketing concepts on the Web flows

Amazon/AlmondNet
Exhibit 1009
Page 239

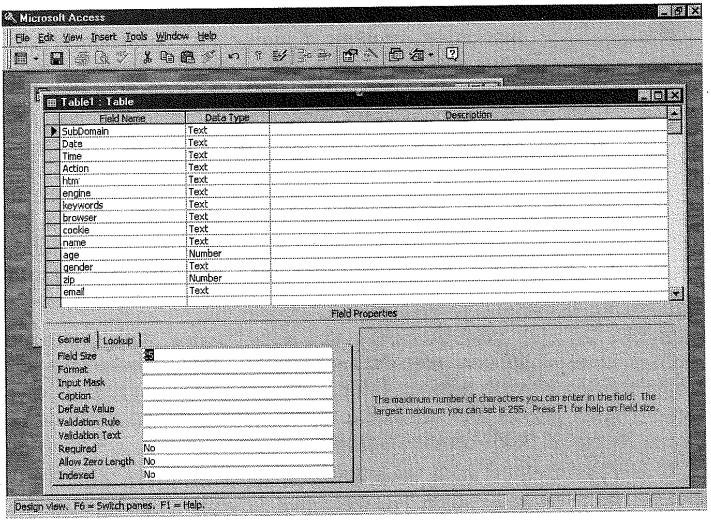


Figure 6–8
A customer database can be constructed from several web data components.

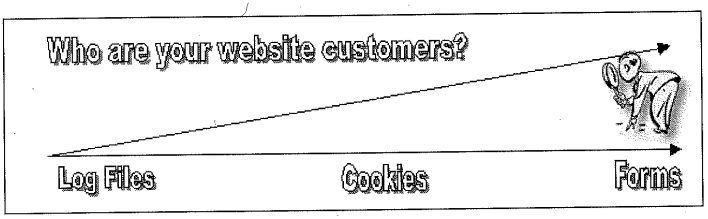


Figure 6–9
Your customer profiles will range in detail based on the components you use.

very quickly and works very effectively—without expensive focus groups and delays. In fact, through the use and creation of dynamic content today you can very easily test different product descriptions, offers and promotions and then do quantitative comparisons to see which work best.

Components Considerations

You need to consider the location of the data you want mined, that is, its storage, prior to analysis. Will you store it on the same server? Probably not. Like other types of data mining projects, issues over data variables you want to capture and convert as well as sample size need to be considered. Additional issues include the intervals of these analy-

Amazon/AlmondNet
Exhibit 1009
Page 240

ses and what type of customer attributes you need to capture for sales, marketing, strategic planning, etc. For example, do you want to mine your data on a daily basis, or will a weekly sample of 10 percent of your online transactions suffice? Keep in mind that irrelevant, redundant, or covariant variables in a database constitute noise to any data analysis and can weaken your data mining predictive models. The Common Log Format was not designed for data mining; it was designed to measure server traffic, and as such it contains a lot of redundant and useless information that can be discarded from the mining analysis. In some instances you will need to append or merge log or database information created from registration forms into a data set more suitable for inductive data analysis (mining).

Another issue you need to deal with is dirty data. You must fix inconsistent data whenever possible including deciding how to deal with missing values. Some data mining tools provide the options and utilities for fixing spotty databases and some do not, so be prepared to do some cleaning.

Lastly, consideration must be given on how you will represent your data. For example, how do you want to deal with categorical attributes such as gender or ZIP codes? Do you want to group them into classes? How do you want to deal with continuous value variables? Do you need to create ratios from them? Compounded by all of these issues is the fact that data from a website is highly dynamic. Keep in mind that these elements are not set in stone, and that their customization is the norm rather than the exception. The data components generated from your website are only the starting point in the data mining process, for their value increases exponentially when merged with your data warehouse and third party information.

Amazon/AlmondNet
Exhibit 1009
Page 241

7

The Providers

In this chapter we will discuss several *external* sources of information that can be linked or appended to your website data for additional insight into the identity, attributes, lifestyle, and behavior of your visitors and customers (Figure 7–1). These sources of information fall into two general categories: the traditional marketing database vendors and a new breed of web-based software and network service providers. The first group represents the traditional marketing packagers and resellers of demographic, consumer and household information. The second group, on the other hand, represents a whole new genre of web companies offering software and services (sometimes both) designed to capture and generate "webographic" data about Internet users' behavior and preferences. These new players use a myriad of solutions to track and profile visitors and shoppers, everything from collaborative filtering technology, to proprietary software and the commingling of cookies via server networks.

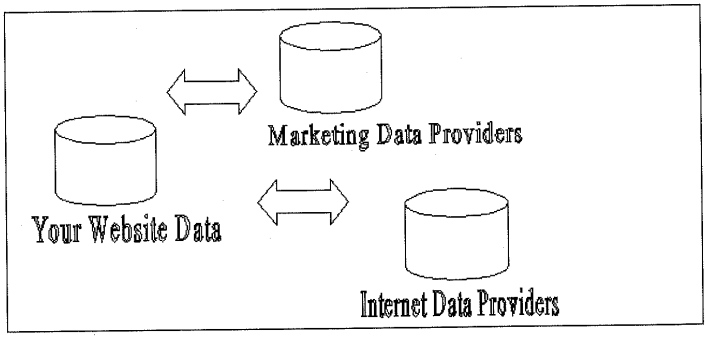


Figure 7–1
Your website data can be enhanced by linking it to offline and online external sources.

Amazon/AlmondNet
Exhibit 1009
Page 242

Marketing Data Providers

There is an entire industry of information resellers who accumulate, organize, and package data for companies and marketers wishing to improve their sales, retain their customers, and enhance their relationships with their clients. The use of demographics, household, income, and consumer data can also be used to add value to the vast amount of customer and transactional information already being captured at your website. Through the appending of this third party data you can gain a new insight about your online customers.

A common methodology in data warehousing is the merging of various internal company data with external commercial information. Data warehousing is the process of bringing together disparate data from various business units scattered throughout an organization for purposes of providing decision-support. A key strategy is often to use data mining technology to identify likely customers for products and services by integrating customer information with external demographic and lifestyle data. Similarly, in database marketing it is quite common to "append" household information from such companies as Acxiom, Equifax, Experian, and others in order to understand what type of individuals are responding to mail campaigns or other promotions and sales incentives.

The merging of server-based data with external demographic, psychographic, and household information can provide an electronic retailing site important insight into the identity of its online visitors and customers. As with the data warehousing and database marketing applications, this additional information can assist a website position, communicate, relate, and retain profitable online customers.

Several firms specialize in aggregating public, census, public records, directory, and consumer information and repackaging it in a format that companies and marketers can use with their own internal data. These data providers sell various levels of information about consumers and households. The following is a description on some of the largest data integrators and providers in the United States—all of which you can use to enrich your perception of your website customers. This can be accomplished by using either the ZIP code or physical address as completed by visitors and customers in your website's registration and purchase forms.

Applied Geographic Solutions

Applied Geographic Solutions provides current year demographic estimates, five- and ten-year projections, consumer expenditure esti-

Amazon/AlmondNet
Exhibit 1009
Page 243

mates, and the MOSAIC neighborhood segmentation system. Data is provided for all standard geographic levels including Block Group, Tract, and ZIP codes, including:

- Basic population and household data, including group quarters population
- Labor force status, employment by industry and occupation
- Household and family income distributions and averages
- Income distributions by age of head of household
- Household size and age of head of household
- Population by age, sex, and race
- Educational attainment
- Vehicles available
- Marital status

The AGS consumer expenditure database covers 340 base variables, arranged in a hierarchical structure to permit ready cross-category and cross-geographic analysis and is available at the block group and higher level. The CEX database is available for the current year and both five- and ten-year projection periods. The main expenditure groups included in the AGS database are:

- Household Furnishings
- Household Equipment
- Food and Beverage Personal Services
- Transportation
- Entertainment Contributions
- Apparel Shelter
- Gifts

www.appliedgeographic.com

Acxiom

Sequestered in the Ozark hills of Arkansas, Acxiom is a company with more information on Americans than the Internal Revenue Service. It certainly has the information better organized and in a more usable format for such tasks as marketing, risk management, fraud prevention, and other knowledge-based applications. Acxiom provides consumer information it has gathered through acquisitions, alliances, and partnerships to verify, identify, and segment customers and prospects on nearly every American today. Acxiom also compiles a variety of public file data from such sources as

Amazon/AlmondNet
Exhibit 1009
Page 244

telephone directories and county recorder and assessors' offices. Axiom acquired DataQuick, one of the largest compilers and resellers of real estate information in the country; thus, they can identify home ownership and related transactions.

Acxiom's Consumer InfoBase is one of the largest multi-sourced database in the United States. It contains:

- property information such as age of the structure, size, and physical characteristics
- household characteristics such as marital status and type of vehicles owned
- individual characteristics such as gender, occupation, and education
- lifestyle information such as interests, hobbies, and activities
- geographic characteristics such as census demographics
- address information including full mailing address
- contact information including telephone number

Consumer InfoBase lists combine consumer data and public record data with information from other sources to create a database of over 150 million individuals residing in over 100 million households. Acxiom also contains data on consumers from Europe, Asia, and South America. The company has also moved to make their vast information available via the Web through their Acxiom Data Network (ADN) system (Figure 7–2).

ADN provides household demographics such as marital status, occupation, and estimated income ranges, as well as other useful

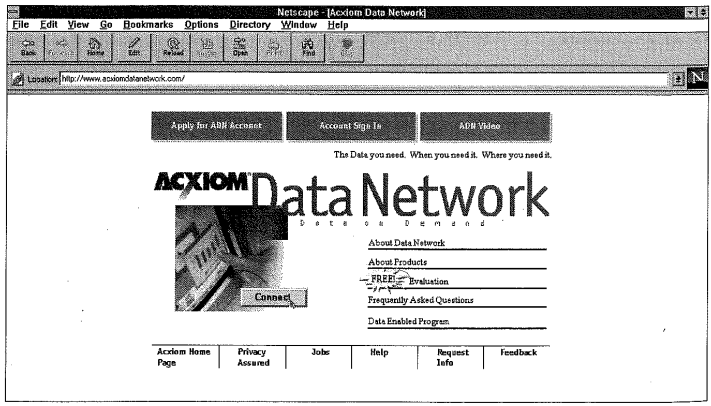


Figure 7–2 *ADN provides consumer data on demand on the Web.*

Amazon/AlmondNet
Exhibit 1009
Page 245

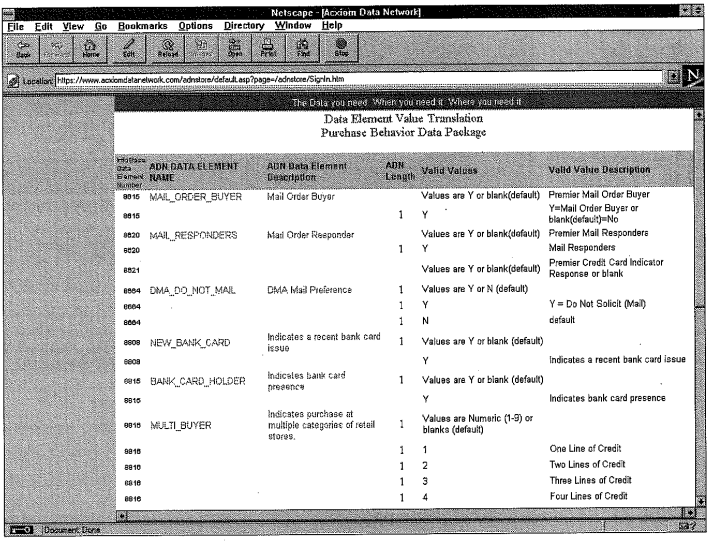


Figure 7–3 *Various types of behavior indicators are available for profiling your web customers.*

information such as details on vehicles, some retail buying habits, and home rental and ownership (Figure 7–3). For example, real property information can be used by home improvement, real estate, and insurance websites to more accurately market their services based on their visitors' and customers' home's age, size, and value. Financial online lenders, like e-loan.com or e-card.com, can also benefit from this type of real estate information for making timely and focused messages based on home equity and loan-to-value information based on a website visitor's length of residence and purchase date.

Auto information from Acxiom can also be used to identify potential prospects for the sale and leasing of cars and trucks for such national sites as autotel.com, carpoint.com, auto.com, and local auto dealership websites. This same type of information can be used by websites selling related products and services, such as auto alarms, stereo systems, painting services, tires or anything else that relates to cars, trucks, and sport utility vehicles. By knowing their website visitors' vehicles, whether they buy used or new vehicles, and what type of vehicles they purchase (compact, convertible, SUV, etc.), they can position the appropriate offers and banners and serve and deliver the appropriate message, product, and service to their site's visitors.

Amazon/AlmondNet
Exhibit 1009
Page 246

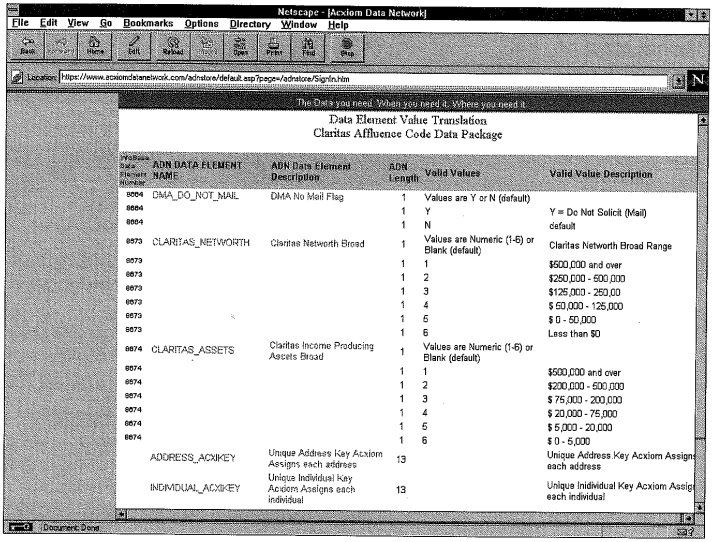


Figure 7–4 *Detailed consumer income information is available from the Acxiom system.*

Extensive consumer buying information is available from Acxiom and their ADN system including the indication of prior mail-order purchases. A propensity to buy from catalogs can be a signal to an electronic retailer that a specific visitor to their site is a consumer who is pressed for time and may favor buying during their spare time, rather than trekking to the neighborhood mall. ADN also contains the Claritas Affluence Code. Claritas is one of the largest providers of marketing data. The affluence code analyzes a variety of economic factors to provide an accepted means of determining relative wealth. This information can be used to identify upscale website visitors who are likely to be interested in luxury items, or for an environmental organization, a possible contributor or supporter.

ADN provides online access to marketing information on U.S. consumers on the basis of demographic, geo-demographic, and personal interest information (Figure 7–4). Axiom does not provide confidential data on individuals, such as credit or medical information. Credit consumer information is protected by federal law and is generally available from the three major credit companies: Equifax, Experian, and TransUnion.

www.acxiom.com

Amazon/AlmondNet
Exhibit 1009
Page 247

CACI (ACORN)

CACI's ACORN neighborhood segmentation system is one of the oldest consumer classification systems. ACORN segments consumers according to ZIP code; thus, if your site is capturing this information in your forms you can use it. For example, suppose you're an electronic retailer of snow boards, and you are looking to expand and grow your online sales. One of the best ways to find new customers is to determine the makeup of your current website visitors and customers and then find others that are similar. Using ACORN to analyze your current website customer base, you may find that a large percentage of your online customers are "Twentysomethings" and "Young Frequent Movers," two of ACORN's 43 clusters, or neighborhood segments.

With this information in hand, you can now seek out other web visitors with ZIP codes with a high percentage concentration from these two clusters. By narrowing your selection of possible customers in this manner, you can now concentrate and focus your incentives or offers to these visitors and shoppers. You can position your ads or banner in websites known to be frequented by these two market clusters. For example, "Twentysomething" consumers are the second ranked market for fast food, while "Young Frequent Movers" are the top ranked market for pet ownership. Armed with this intelligence you may want to position your banner with search engines when keywords such as "dog training" or "pizza" are entered.

ACORN's neighborhood segments are broken up into 43 clusters, which are collected in 9 general categories (groups). The following are some of their key traits:

GROUP 1: AFFLUENT FAMILIES

1A Top One Percent	These are the wealthiest neighborhoods in the United States. This highly educated type is the top-ranked group for investments, drinking wine, and ordering by phone, a key indicator they are also likely to buy via the Web from such sites like Virtual Vineyards *virtualvin.com.*
1B Wealthy Seaboard Suburbs	This type is located along the eastern seaboard and in California. They rank among the highest for auto club memberships and travel extensively, a key indicator for travel-related websites like *travelocity.com.*

Amazon/AlmondNet
Exhibit 1009
Page 248

1C Upper Income Empty Nesters	These are business owners and managers who are distributed throughout the country. Home furnishings and improvements rank high among their expenditures. Office supplier retailers like *officemax.com* and *officedepot.com* may want to identify, attract, and retain these visitors.
1D Successful Suburbanites	Essentially suburban, they drive new cars and spend money on sports, exercise equipment, and personal computers. These consumers are seasoned online shoppers and will likely continue to buy products and services on the Web. They have a particularly high lifetime value LTV ranking. These are savvy consumers and likely to shop at *compare.net.*
1E Prosperous Baby Boomers	They participate in sports, but do not invest or save in proportion to their income. *espn.com* and sport equipment retailers will want to recognize these visitors.
1F Semirural Lifestyle	They maintain their livelihood by self-employment—many of them working at home. They are interested in reading, needlework, and cooking. Home office workers, these consumers have come to use the Web for browsing and shopping. They are likely to be the most loyal and profitable customers of *amazon.com.*

GROUP 2: UPSCALE HOUSEHOLDS

2A Urban Professional Couples	Taking domestic vacations, they spend on theater and concerts and dining out rather than cooking in. These are *geocities.com* browsers and shoppers.
2B Baby Boomers with Children	This is a homogeneous type who spends their time and money on home and family-oriented activities and goods. These are the *disney.com, etoy.com,* and *ty.com* consumers.
2C Thriving Immigrants	Owning 2+ cars, they neither invest nor save. Activities include casino gambling and ordering take-out while watching the NFL on TV. This is the *WebTv.com* crowd.

Amazon/AlmondNet
Exhibit 1009
Page 249

2D Pacific Heights	Living in high-rent districts, located primarily in California and Hawaii, these residents are predominantly Asians and Pacific Islanders who are slightly older, urban, and upscale. These visitors are location and ethnicity-specific and may warrant ads and incentives geared to their area and interests, such as *sfgate.com* the San Francisco local newspaper website.
2E Older Settled Married Couples	Spending is home-oriented; almost 20 percent draw retirement incomes. These visitors are age- and income-driven and may be primarily interested in bargains such as those found in *ebay.com.*

GROUP 3: UP & COMING SINGLES

3A High Rise Renters	They use credit cards extensively and are top-ranked for ATM use. These consumers should be of interest to card issuers and banks such as *wellsfargo.com* and *e-card.com.*
3B Enterprising Young Singles	They tend to rent videos and use PCs at work and at home. These consumers are savvy users of the Web and would be of interest to such sites as *reel.com* and *movies.go.com.*

GROUP 4: RETIREMENT STYLES

4A Retirement Communities	Luxury cars, outdoor gardening, attending sporting events, using vitamins, and ordering by mail and phone are popular. Increasingly consumers over 50 are beginning to make online purchases; as such, these consumers may be of interest to such sites as *gardenweb.com* and *classicar.com.*
4B Active Senior Singles	They tend to use coupons and spend money on specialty food items, romance, mystery books, and lottery tickets. These coupon hunters are likely to be of interest to *www.qpons.com,* a site that not only dispenses coupons online but also e-mails them to visitors when new coupons become available in their area.

Amazon/AlmondNet
Exhibit 1009
Page 250

4C Prosperous Older Couples	Home-oriented, they do take domestic trips and participate in sports such as fitness walking and playing golf. A site of interest to these consumers might well be *golfweb.com.*
4D Wealthiest Seniors	They rank first in domestic/foreign travel, fitness walking, coffee drinking, subscribing to cable, and mail/phone order. These consumers are likely to step up to *walking.org* and *cnn.com.*
4E Rural Resort Dwellers	Four-wheel drive vehicles, camping equipment, and domestic trips reflect their lifestyle. Top-ranked for gardening equipment, owning large dogs, and taking vitamins. These consumers will be of interest to *camping.org* and *4wd.com.*
4F Senior Sun Seekers	They are the top-ranked market for mail/phone orders of audio tapes and the purchase of lottery tickets, and are the second-ranked market for book club membership. Sites, which could benefit from these consumers, include *audiosource.com* and *audiobookclub.com.*

GROUP 5: YOUNG MOBILE ADULTS

5A Twentysomethings	This is the second-ranked market for fast food and they are also dieters. What else but *mtv.com* and *pizzaclub.com.*
5B College Campuses	Top-ranked for take-out fast food, having ATM cards, and making long-distance calls using calling cards, they rank highest for watching sports on TV: *abc.com, nfl.com,* and *nba.com.*
5C Military Proximity	They are top-ranked for Bible purchases, disposable diapers, and watching rented videos: *gospelcom.net.*

GROUP 6: CITY DWELLERS

6A East Coast Immigrants	They spend on foreign travel, furniture, and household goods: *furniturefind.com.*
6B Working Class Families	They buy apparel, are top-ranked for owning 4+ televisions and for drinking wine coolers and cola sodas: *foxnetwork.com.*

Amazon/AlmondNet
Exhibit 1009
Page 251

6C Newly Formed Households	Most drive older cars and use coupons for groceries: *hotcoupons.com.*
6D Southwestern Families	Their lifestyle is both urban and rural with spending being family-oriented purchases: *hispanic.com.*
6E West Coast Immigrants	Their budgets include heavy expenditures on groceries, baby products, and movie attendance: *fox.com.*
6F Low Income: Young & Old	They purchase infant toys, toy cars and trucks, jewelry, and household items such as flatware and washing machines: *toysrus.com.*

GROUP 7: FACTORY & FARM COMMUNITIES

7A Middle America	Their budget priorities are home-oriented (improvement, children's apparel, lawn equipment) with leisure activities including hunting, fishing, and needlework: *build.com, hometime.com,* and *yarntree.com.*
7B Young Frequent Movers	This is the top-ranked market for pet ownership and sites like *petquarters.com* and *petco.com.*
7C Rural Industrial Workers	Most own 3+ vehicles: *ford.com* and *chevy.com.*
7D Prairie Farmers	Their top-ranked activity is hunting and fishing and they are likely to own pets: *fbnews.com* and *sportsmensweb.com.*
7E Small Town Working Families	Active in their communities, they are top-ranked for buying preschool toys and reading *Country Home* magazine: *countryliving.com.*
7F Rustbelt Neighborhoods	They tend to drive used cars, splurge on lottery tickets, and are top-ranked for using coupons for beauty and household cleaning products: *ausedcar.com* and *valupage.com.*
7G Heartland Communities	They are top-ranked for outdoor vegetable gardening, flower gardening, and buying videos at discount stores: *wal-mart.com.*

Amazon/AlmondNet
Exhibit 1009
Page 252

GROUP 8: DOWNTOWN RESIDENTS

8A Young Immigrant Families	They tend to spend their money on major household appliances, athletic and dress shoes, and baby products: *nike.com.*
8B Social Security Dependents	They spend money on mail and phone order items and are the number one market for having personal loans: *hfc.com.*
8C Distressed Neighborhoods	They are top-ranked for shopping at malls and buying children's items. Admittedly, these consumers will not likely be online shoppers, as they prefer the physical malls.
8D Hard Times	They are top-ranked for buying major household appliances and baby products and they tend to purchase fast food and takeout food: *sears.com.*
8E Urban Working Families	They are top-ranked for purchases of grills and climate control machines, and buying educational toys: *grategrills.com.*

GROUP 9: NONRESIDENTIAL NEIGHBORHOODS

9A Business Districts	Business Districts have a residential population of less than 100 and a significant daytime business population.
9B Institutional Populations	These neighborhoods are home to institutional group quarters, including prisons, juvenile detention homes, and mental hospitals.
9C Unpopulated Areas	These areas are unpopulated and include developed acreage such as parks, cemeteries, golf courses, or undeveloped land.

www.caci.com

Claritas

Claritas introduced the original lifestyle segmentation system in the early 1970s called PRIZM. The system segments all the neighborhoods in the United States in terms of 62 demographically and behaviorally distinct types, or "clusters." In doing so Claritas developed a marketing tool that can be used to identify, understand, and target consumers based on their locale and lifestyles. Currently PRIZM links with virtually every major marketing database, consumer survey, and direct mail list in the United States.

Amazon/AlmondNet
Exhibit 1009
Page 253

Claritas also provides information on consumer expenditure databases, which allow the user to segment markets by age and income, and isolate consumers by revenue potential. Claritas uses a network of over 1,600 local government and private sources to create this information. Claritas' Dimensions database incorporates household- or individual-level data from file enhancement vendors (Experian, Infobase, etc.) or actual client files with neighborhood clusters and demographic distributions to produce industry- or company-specific segmentations.

Claritas also has a database for financial services targeting called P$YCLE, a system created expressly to differentiate households in terms of financial behavior. It predicts which types of households will use which types of financial products and services. It is used by leading financial institutions for prospective cross-selling, branch evaluation, product planning and development, advertising, and media planning. Potentially P$YCLE could be used by stock trading sites and other investment and financial related sites, such as *quote.com, etrade.com, ameritrade.com,* etc.

LifeP$YCLE is a household-level segmentation system that groups customers in terms of 57 distinct types of life insurance customers. LifeP$YCLE is optimized to predict who is most likely to buy insurance products based on actual usage patterns of asset accumulation and income protection products. Claritas' Market Audit is the largest syndicated database of consumer financial behavior. Data is compiled annually using comprehensive, random-dialed phone interviews with over 100,000 households. The Market Audit covers household use of specific financial products, accounts and account balances, and where accounts are held. This database will likely interest insurance-related sites, such as *insweb.com.*

www.claritas.com

Equifax

The San Diego-based Equifax National Decision Systems (ENDS) is another consumer segmentation system. To see how the ENDS segments a neighborhood go to the "LifeQuiz" application on their website at *http://www.natdecsys.com* and input your ZIP code (Figures 7–5 and 7–6).

Equifax provides MicroVision reports, which uses demographic data and aggregated consumer demand data at the ZIP+4 level to classify every U.S. household into unique market segments. Each segment consists of households that share similar interests, purchasing

Amazon/AlmondNet
Exhibit 1009
Page 254

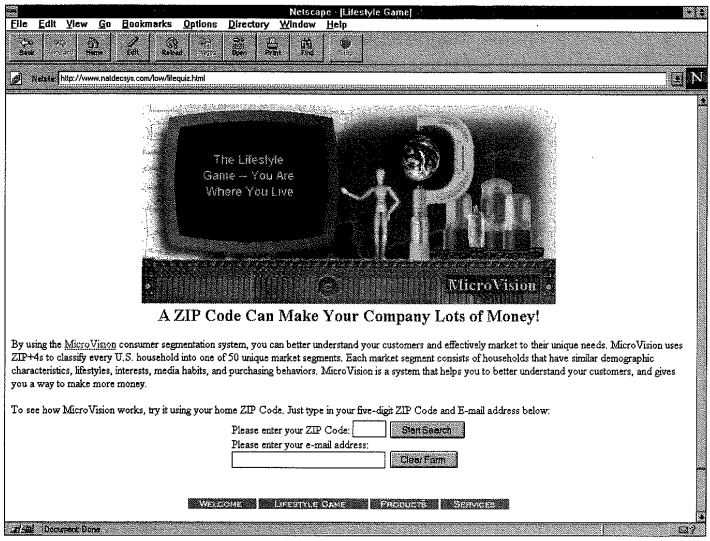


Figure 7–5
Input your ZIP code here.

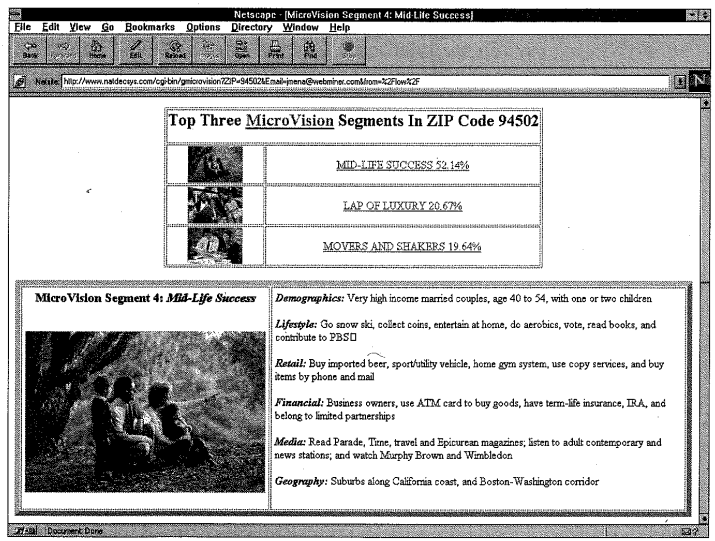


Figure 7–6
And see the demographic profile of your neighborhood here.

Amazon/AlmondNet
Exhibit 1009
Page 255

patterns, financial behavior and demand for products and services. The MicroVision reports cluster each U.S. household into one of 50 segments based on census demographics, interests, purchasing patterns, and financial behaviors.

MicroVision also classifies every U.S. household into one of 95 unique and homogeneous sub-cluster codes. The codes are assigned at the Zip+4 level—it is assumed that all households within that Zip+4 share the same characteristics. These sub-cluster codes are the building blocks for industry-specific systems such as Micro-Vision Insurance. Equifax creates MicroVision reports for businesses in the real estate, retail, restaurant, banking, insurance, investments, cable, telecommunications, utilities, consumer products, publishing, and other industries.

www.equifax.com

Experian

Experian, formerly known as TRW, provides the MOSAIC targeting system, which identifies consumers according to the type of neighborhood in which they live. The MOSAIC system is international and is available in 16 countries; an estimated 779 million people are classified by Experian, which may be of interest to large online directories and major portals who attract visitors from around the world. In an effort to enhance MOSAIC, Experian acquired Direct Marketing Technology Inc. also known as *Direct Tech*, a company that processes billions of direct mailings in the United States.

Experian generates approximately 50 percent of all the prospect names for pre-approved credit offers, using the financial and demographic information held in their large consumer databases. Their geodemographic segmentation system, MOSAIC is one of the world's most widely used systems for classifying consumers. Their Global MOSAIC enables Experian to perform analysis of consumers worldwide. Experian also uses their psychographic systems for developing behavioral analyses for cross-selling among existing customers and risk-based scorecards for minimizing the likelihood of future financial loss when targeting new customers.

Experian has one of the world's largest databases of demographic and financial information. Supported by this core data they provide a wide range of data processing services. These include merging client files with new prospect lists, segmenting their files into more refined prospect groups using Experian's own profiling techniques, and enhancing information through deduplication and the addition of full address details.

Amazon/AlmondNet
Exhibit 1009
Page 256

Experian's lifestyle database in the U.K., called Chorus, is compiled from large-scale surveys to build a comprehensive picture of the way consumers live, together with their demographics and purchasing behavior. In the United States, they have developed a shared database of individual-level transaction information from more than 150 catalog clients. Called Z-24, the database helps contributing members make major improvements in their prospect targeting. Experian also manages customer databases for some major organizations.

www.experian.com

Polk

Polk is the oldest consumer marketing information company in the country. It compiles lifestyle and other demographic data on more than 100 million consumer households across the United States and Canada. It provides annual statistics on nearly 200 million cars and supplies detailed consumer analysis to more than 100 major consumer goods companies. Polk also provides geographic data covering more than 12 million addressed street segments in the United States as well as nationwide census and current postal boundary products. Among the consumer lists they provide are the following:

The Lifestyle Selector®. The Lifestyle Selector is a comprehensive consumer database made up of buyers who complete and mail detailed questionnaires regarding their lifestyle interests and the purchases they have made. It details such products as automotive accessories, consumer electronics, furniture, small and major appliances, outdoor recreation and others. This is the list count for computer hardware, which most likely was culled from warranty cards for each product type:

CD-ROM Owners	2,968,000
Fax Modem Owners	89,000
Laptop Computer Owners	137,000
On-Line/Internet Service Subscribers	92,000
Personal/Home Computers Enthusiasts	14,210,000
Personal Home Computer Owners	7,861,000
Personal Home Computer Users	9,205,000
PC Compatible Owners/Users	6,447,000
Apple/Macintosh Owners/Users	1,525,000

Amazon/AlmondNet
Exhibit 1009
Page 257

Buyer Beware

Although these data resellers are very good at providing timely and accurate demographics, the density of their coverage, or "file penetration," is not always 100 percent. To give you an idea of just how accurate their information is the following table reflects the actual coverage of a household database provider by individual attributes and file density:

Attribute	Inferred	File Density (Accuracy)
Head of Household Age	yes	98% (Exact—YOB)
Income	yes	98%
Marital Status	yes	25–76%
Number of Children	no	22%
Gender	yes	100%
Mail Order Responsive	no	70% (26 categories)
Ethnicity	yes	25%
Education	no	4–100%
Wealth	no	100%
Employment Type	no	11%
Length of Residence	yes	88%
Age of Children	no	15%
Lifestyle	no	12%
Phone	no	80%
Loan to Value	no	48%
Dwelling Type	yes	100%
Available Equity	yes	48%
Home Value	yes	88%
Presence of 1st Mortgage	no	88%
Presence of 2nd Mortgage	no	26%
Credit Cards by type and #	no	6%
Installment Products	no	4%
Savings Accounts	no	2%
Home Age	yes	88%
Lender Type	no	30%
Loan Type	no	28%

Keep in mind that this is a household file, not a consumer credit file, and as such does not contain detailed information on loans and other financial account activity. This type of information is available by simply providing a matching address. You should be aware that the density of the file that you purchase is likely to vary depending on the vendor you select and the type of attributes you desire.

Amazon/AlmondNet
Exhibit 1009
Page 258

Business-to-Business Intelligence

There are also data providers who sell information on companies and corporations. For a business-to-business site doing electronic commerce this type of company intelligence can enable it to improve its sales and evaluate the credit-worthiness of its trading partners and clients.

The data providers of business information are totally different from those providing consumer demographics. These data resellers provide timely and accurate company data gleaned and assembled from regulatory and licensing federal, state, and local agencies. One major provider of public records is Information America, which makes its database available on its site KnowX (Figures 7–7 and 7–8). Through KnowX, Information America provides access to public records including real estate transactions, lawsuits, judgments, and professional license information from states, counties, and cities.

As with the profiling of online consumers, the profiling of companies can enable a website involved in electronic commerce to gain an insight about its clients. Through the same processes of warehousing and mining a business-to-business electronic commerce site can model and improve its sales and its competitive status by constructing composites of trading partners and its most profitable accounts.

`http://knowx.com`

Lexis-Nexis is another firm that offers legal, legislative, regulatory, news, business, and public record information by subscription:

`http://lexis-nexis.com`

Dunn & Bradstreet also provides to U.S.-based companies credit, collections, and marketing reports:

`www.dnb.com`

High Tech Connect®. Of special interest to web administrators and marketers is this Polk list of home computer consumers. The High-Tech database has 45 high-tech data elements, falling into the following five categories:

- Software Users and Buyers (including Mail-Order Buyers)
- High-Tech Product Owners and Users
- Online/Internet Service Subscribers
- Online/Internet Service Intenders
- High-Tech Product Intenders

This is a 36 million-name database self-reported by consumers themselves, which can be segmented by 11 demographic and more than 65 lifestyle characteristics. This is a partial section of this file:

Amazon/AlmondNet
Exhibit 1009
Page 259

Figure 7–7
KnowX provides legal and business information.

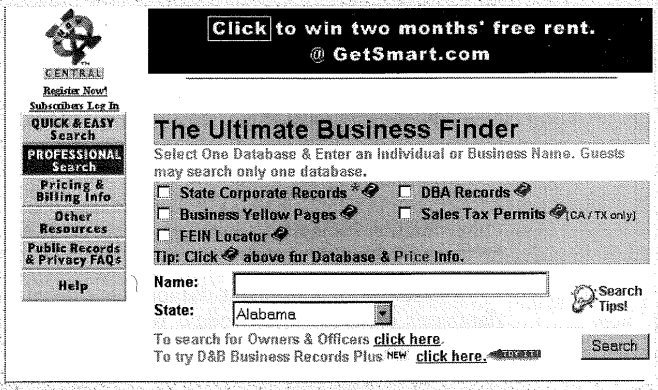


Figure 7–8
Multiple and single searches of company information are available.



Amazon/AlmondNet
Exhibit 1009
Page 260

Online Subscribers	
Any Online Service	266,000
America Online	151,000
CompuServe	41,000
Microsoft Network	25,000
Prodigy	35,000
Other Internet Provider	48,000

TotaList. Polk's TotaList database is one of the largest consumer databases on the market today. With information on more than 95 million U.S. households and over 180 million individuals, the TotaList database delivers a combination of automotive ownership data, lifestyles, interests, and activities. The file is continuously updated with new sources and processed against ZIP+4 from nearly 30 sources of data, and it has more than 100 individual selection features including the following sectors:

- Household Automotive Finance selects General Household characteristics
- Household Member information
- Household Automotive selects
- Specific Motor Vehicle selects Housing/Property information

Other Polk files include the *Vehicle Databases* and the *Auto Purchase Predictor Database.*

www.polk.com

Trans Union

Trans Union is the third credit bureau in the country along with Experian and Equifax, which like the other credit reporting companies also provides several information management products. Among the marketing products lines it offers are the following:

- *Silhouette,* a market segmentation product that groups consumers into one of 25 clusters according to their individual credit lifestyle. Demographics are included. Permissible purpose is required.
- *Facets,* a risk-based clustering system, uses individual-level credit information to segment accounts into one of 13 clusters. Trans Union claims Facets is optimized when used in conjunction with a risk score, because it improves the risk-prediction of these scores.

Amazon/AlmondNet
Exhibit 1009
Page 261

- *Sentry* is an attrition model that predicts the likelihood of an account voluntarily closing within 3–5 months.
- *Revenue Projection Model* rank orders individuals by the relative amount of revenue they are likely to produce on a revolving credit account during the next twelve months.

www.tuc.com

Web Services and Software Providers

The Web is a highly competitive and expanding marketplace that doubles every hundred days. As the general content of the Web gets broader, visitors and shoppers will cease to browse aimlessly and will instead gravitate toward sites that deliver products and services customized to their consumer needs. It is more important than ever to provide customers with an interactive, personal web experience. Retailing sites must plan now to respond to this expectation or risk being left behind as the Web changes to a personal medium. These web vendors concentrate on providing webographic information, which allows sites to tailor content to a visitor's preferences.

The following group of webographics providers is a relatively new breed of Internet companies, who are in the business of selling software or services for collaborative filtering, relational marketing, and customer profiling. They provide detailed information on visitor behavior, preferences, and profiles.

These vendors differ dramatically in how they approach these tasks. Most use proprietary technology on the server-side for capturing some of this customer information, while others concentrate on monitoring the behavior of users and then predicting what type of products or incentives they are likely to purchase and what ads or banners they will respond to. Almost all provide the products and services exclusively on the Internet for electronic retailers and marketers.

It is a very new industry with few standards and clear winners; a highly dynamic market sector with many novel approaches to the main electronic retailing questions: *who are my customers and what kind of products and services are they likely to purchase from me?*

Some of these services and products track and log every web page displayed to every user so you can develop a "clickstream" view of what they saw, when they saw it, and for how long. This type of data can be stored and mined for a complete understanding of their usage, behavior, and preference. A few of these firms provide

Amazon/AlmondNet
Exhibit 1009
Page 262

audience measurements and advertising services, which may not provide additional data for mining, but they may provide additional insight about your visitors and customers.

Art Technology Group

The Art Technology Group (ATG) started as a web design group and is now providing a suite of Java-based electronic commerce products, collectively referred to as *Dynamo Product Suite,* which is designed to provide highly scalable, multi-platform, business-to-consumer capabilities. The suite consists of the following discrete products and two components, the Personalization Server and Commerce Application:

The first part of the Dynamo Personalization Server is made up of the Dynamo Application Server, a web application server (WAS) that provides the runtime engine for all of the three electronic commerce products. It supplies native interfaces for Oracle, Sybase, and Microsoft databases, and runs on Java Virtual Machines of the major variants of Unix as well as Windows NT. The other part is the Dynamo Personalization Station, which was designed to dynamically track user behavior. It also provides user survey capabilities that predict user behavior. These usage tracking and survey tools are often referred to as "personalization" features. This is the key customization component: it combines information from demographic databases and user behavior to dynamically display personalized content. Open Content Adapters are used for integration to financial data systems and other system repositories. Business rules can also be specified regarding the presentation of content to visitors.

The Dynamo Commerce Application is also made up of two parts. The first is the Ad Station, an advertising management product that targets delivery of the entire media range of Web-based ads based on user profiles and behavioral information. Ad Station includes ad response measurement and reporting and audience analysis features, in addition to core ad server functions. The other part is the Retail Station, which provides catalog, storefront, site design, and shopping-cart build and management capabilities for commerce site developers. Retail Station also offers ongoing administration applications for store management, inventory, and sales personnel.

Although buyers may purchase any of the products individually, ATG designed the products to work in concert as a total business-to-consumer "relationship commerce" solution. In addition, ATG offers design and related services in support of its relationship commerce

Amazon/AlmondNet
Exhibit 1009
Page 263

product line. ATG provides turnkey customization of this Java-based electronic commerce system.

www.atg.com

Aptex

Aptex is a subsidiary of HNC, a neural network company and a world leader in monitoring for credit card fraud. Aptex is a company using both proprietary real-time content analysis techniques coupled with neural networks to predict web user behavior; they call it Content Mining technology. Aptex's two core products are *SelectCast* and *SelectResponse*. Both products are designed to enhance customer relationships, personalize one-to-one marketing, and attract potential customers by engaging their interests and motivating them to make online purchases.

SelectCast is designed to assist online publishers and retailers by attracting and retaining customers, delivering targeted one-to-one product recommendations, promotions, content, and advertising based on real-time behavior. SelectCast customers include Advanta, Excite, Infoseek, MatchLogic, and Cendant. Aptex claims SelectCast for Ad Servers is able to boost ad clickthrough over other targeting methods by making each ad an independent, intelligent agent that continuously seeks out interested users. The "agents" make decisions based on real-time user behavior, reacting instantly to serve more impressions to groups of users that demonstrate interest based on the content of current and recent page views, queries, clicks, and ad impressions.

SelectResponse is designed to enhance and personalize customer interactions through automating online e-mail, web, chat, and call-center responses to real-time customer inquiries. SelectResponse reads the content of customer inquiries and either responds, using FAQ documents and corporate knowledgebases, or it can route messages to the appropriate customer service agent with a suggested response.

One of the unique features of the Aptex profiling technology is that it doesn't store personal customer information. It instead uses a neural network to profile users based exclusively on their real-time actions and observed user behavior. Aptex claims their system can learn continuously a user's interests based on the content they browse on, the promotions they click on, and the products they eventually purchase. Among the features claimed by Aptex are the ability to target specific behavioral audiences and the delivery of designated promotional content. The ability to identify new audiences automatically

Amazon/AlmondNet
Exhibit 1009
Page 264

as they emerge and to act as an "intelligent observer"—mining the context and content of all actions—including queries, page views, promotional impressions and clicks, and purchases.

Aptex's Content Mining technology uses a patented mathematical modeling technique it developed under a four-year U.S.-government contract and proprietary neural network algorithms derived from their parent company HNC. Aptex's Context Vector data modeling technique is supposed to "read" and "understand" unstructured text. This learning provides the basis for both interactive and automated decision-making based on the text and its internal relationships and for predictive modeling of future patterns, trends, and behaviors.

www.aptex.com

Autonomy

Autonomy sells a suite of Agent software products for commercial websites, ISPs, and information providers. The Agentware technology provides the ability to categorize, hyperlink, and present large volumes of text-based content automatically, without the need for manual input. Autonomy also claims the unique ability to automatically create "user profiles," which monitor and encode each individual's specific areas of interest and to deliver personalized information in real time. Autonomy is supposed to monitor the activities of site visitors and recognize their areas of interest, from which it develops individual interest-profiles.

Agentware Content Server is the nucleus of the Autonomy's products. Autonomy's Adaptive Probabilistic Concept Modeling (APCM) technology enables the Agentware software to identify and encode the unique "signature" of key concepts within text documents. APCM then uses this signature of the concept, a Concept Agent, to seek out and uncover the presence of similar concepts in volumes of content such as a set of similar products, pages, or websites. Autonomy's technology is based in part on neural networks and Bayesian statistics. Autonomy technology allows for collaborative filtering, eliminating the use of forms at websites—it instead makes recommendations based on the preferences of users with similar tastes and allows for the creation of user profiles.

www.agentware.com

BroadVision

BroadVision One-To-One is a server software application system. It incorporates a suite of management tools that allow a web adminis-

Amazon/AlmondNet
Exhibit 1009
Page 265

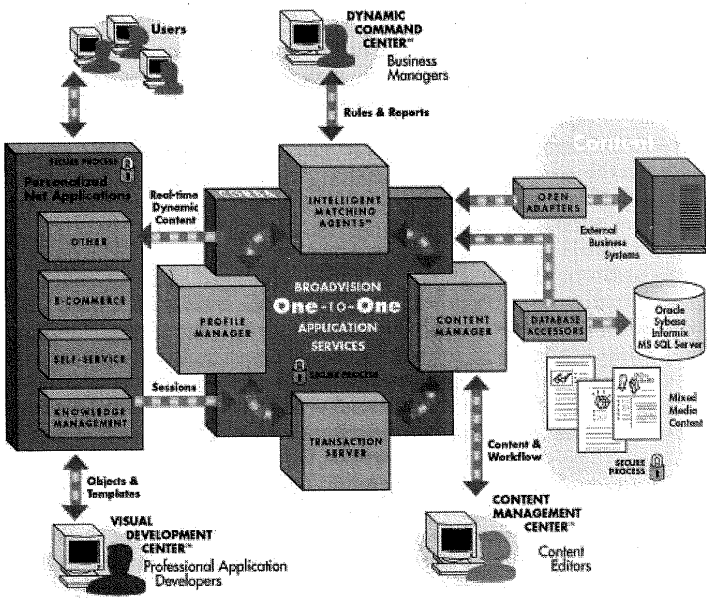


Figure 7–9 *The BroadVision website management system.*

trator to dynamically control multiple server applications. The suite of tools includes:

- Intelligent Matching Agents, which are tools that dynamically match user profiles with content. (The intelligent matching feature can be driven by business managers, users, or agents.)
- Content/Data, which includes BroadVision content that is stored in relational databases plus content/data from any existing business system, host, or legacy database. (See Figure 7–9.)
- Transaction Server provides end-to-end, secure management of transactions that can be customized to meet business needs.
- Profile Manager, which manages user-profiles including information supplied by the user, observed during sessions, or derived from existing customer databases.

Personalized Net Applications deliver individualized content in real time and provide secure interactions and transactions for consumers and business users.

BroadVision was issued a patent on a database system, which creates a "customer monitoring object" for each visitor. Referencing

Amazon/AlmondNet
Exhibit 1009
Page 266

information relating to that visitor creates the object, which is stored in the customer information database and accessed when the visitor selects a product, makes a purchase, or requests information. The customer information database, which the BroadVision system creates, can be mined to uncover important hidden patterns about activity in that website. BroadVision stresses the marketing concepts of one-to-one marketing espoused by Rogers and Pepper and has designed their software to provide the type of retailing model their popular marketing books describe.

www.broadvision.com

DoubleClick

DoubleClick is a provider of targeted ads over a group of large networked commercial websites. The DoubleClick system tracks user movements between various member websites run by clients on their advertising networks. DoubleClick is able to serve up a unique ad for each user, depending upon a user's interests as expressed via their selections in the websites that are part of the DoubleClick network (Figure 7–10). The DoubleClick Network consists of a collection of over 60 highly trafficked sites that commingle their cookies in order to place the appropriate ad to visitors. The Network of sites use

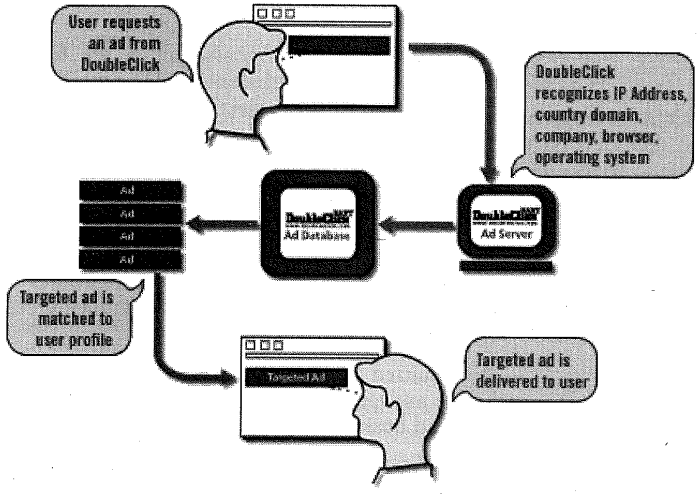


Figure 7–10 *DoubleClick is in the business of putting the "right" ad in front of the "right" visitor.*

Amazon/AlmondNet
Exhibit 1009
Page 267

DoubleClick's proprietary DART targeting technology that allows advertisers to target their best prospects based on the most precise profiling criteria available.

DoubleClick uses its network of sites to collect information on where users have gone and what they are likely to respond to positively. DoubleClick also uses a feature that allows clients to deliver demographically targeted ads based on the site's own database of user registration information. DART, which is DoubleClick's proprietary ad targeting and reporting technology, was once available only to affiliate sites of the DoubleClick Network. However, the company recently began providing its DART Service to other sites. The newly enhanced DART technology aggregates information, voluntarily offered by the user, in a website's database against DoubleClick's other targeting criteria, such as company name and geographic location. The DART data that is being captured may be used in conjunction with a website's own data. This aggregated data may also be mined to extract customer profiles and additional insight.

www.doubleclick.com

Engage Technologies

Engage Technologies sells software that is designed to enable sites to distinguish anonymous and registered visitors and customers. Engage Technologies products include the Engage Suite, which is made up of:

- Engage.Journal, for website data collection and visitor individualization.
- Engage.Portrait, for customizable website registration management system.
- Engage.Discover, for customizable usage analysis, reporting, and iterative querying against website data.
- Engage.Link, for high-performance API leveraging user profile data for personalized content generation.
- Engage.Knowledge services allow website members who are "Engage-Enabled" to access their proprietary database of millions of Internet user profiles, which include historic and multi-site views of anonymous, individual users' interests, experiences and usage patterns.
- Engage.Knowledge Real-Time Visitor Intelligence is a usage-based subscription service that provides immediate access to the world's

Amazon/AlmondNet
Exhibit 1009
Page 268

largest database of behavior-based profiles of anonymous Website visitors. By accessing this database, members can use multi-site interest profiles of their visitors for use in real-time personalization applications and ad targeting.

- Engage.Journal, which offers organizations a standardized method of automating the capture, integration, and management of website visitor data.

Engage.Journal identifies individual website visitors using global and local "certificates." These certificates link visitor identifications to anonymous log statistics, including page requests, the domain from which they originate, and interest classifications based on the content they view. Certificate information can also be transmitted to a global server, housed either at Engage Technologies or at an organization's site, which allows visitor information to be captured and correlated across multiple websites and Internet domains.

Engage.Journal works with Netscape's NSAPITM and Microsoft's ISAPITM to log user actions and issue certificates. When a customer or prospect first visits your website, Engage.Journal automatically issues that user a unique local certificate that identifies the user anonymously during the current session and all future sessions. This process provides a standardized method for identifying and observing unique visitors within a single website or across a network of websites and allows for profiling, analysis, and personalized marketing activities to anonymous individuals.

Engage.Journal can be used separately or in partnership with the Engage.Portrait, Engage.Link, and Engage.Discover products. Engage.Portrait is a complete, customizable website registration system that records and leverages demographic and psychographic information about visitors to an organization's website. Engage.Portrait includes a wide array of demographic and psychographic categories to choose from when designing registration forms. These categories range from personal information, such as name, address, and marital status, to professional information including business interests, job function, industry, and size of company. Engage.Portrait includes affinity information that allows you to build profiles of your website customers and prospects based on general and specific areas of interest. The Engage system uses declared data from registration forms or observed behavior either globally or site-specific, as well as overlaid demographic and lifestyle data to suggest propensity to action. This data of course can also be mined for clustering, segmentation, classification, and prediction of customer behavior.

www.engagetech.com

Amazon/AlmondNet
Exhibit 1009
Page 269

Firefly

Firefly started as an application based on "agent" technology, a collaborative filtering engine that runs on a server and provides advice to website visitors based on their own tastes and on others with similar preferences. Pattie Maes from the MIT Media Lab developed Firefly as an intelligent, active, and personalized agent collaborator. The idea behind collaborative filtering or agents is that instead of user-initiated interaction via commands of direct queries, the visitor is engaged in a cooperative process in which the software initiates communication, monitors events, and performs tasks. The concept of the original Firefly "agent" was that it would become gradually more effective as it learned a visitor's interests, habits, and preferences. Firefly was one of the first collaborative filtering firms, having started at the MIT Media Lab. It was subsequently purchased by Microsoft.

In the core technology of Firefly, an agent memorizes which books, CDs, or other consumer items a visitor or customer has evaluated and purchased. Then it compares these preferences with other agents, correlating them with similar items and values—correlating them and assigning a value of "similarity" which it uses to recommend future books, CDs, etc. The important concept hidden in the application is that the agent gradually "learns" how to better assist the client, much like the corner store clerk who has observed your preferences. An agent learns by observing and imitating the customer, receiving feedback and instructions from the customer, and asking advice from other agents.

Firefly's flagship product, called Firefly Passport, is used to collect user preferences anonymously, recommend Internet content, and send appropriate advertising. Passport, makes it possible for sites to register visitors and exchange user profiles with other Firefly-enabled sites. The company's one-time goal was to build a network of such sites and deliver personalized content and community services to registered users.

The Firefly Navigator Catalog lets sites steer users toward products by creating personalized lists of recommendations and connecting users with similar interests. The Firefly Network Hub allows for the network administration with affiliated sites. The data that Firefly aggregates can also be merged with your own transactional data for subsequent mining, giving you the ability to tailor your offers and presentations to fit your visitors' and customers' needs.

www.firefly.net

Amazon/AlmondNet
Exhibit 1009
Page 270

GuestTrack

GuestTrack is another personalization and tracking software system designed to gather and store information about each visitor and personalize content "on the fly." GuestTrack uses a profile database to ask users questions on practically any Web page, and store the information in their personal profile. The GuestTrack ID is stored in the user's bookmark file, and most GuestTrack personalization is done with special tags in the HTML file. GuestTrack logs every web page displayed to every user so you can develop a "clickstream" view of what they saw, when they saw it, and for how long. Included in the GuestTrack system is an export module that creates a standard comma-delimited file that is compatible with most major database and spreadsheet programs. This provides the ability to sort, select, tabulate, report, and mine the data. The log contains the date and time in both traditional and Unix formats, as well as the individual's GuestTrack ID and the web page displayed. This data is exportable to data mining software for in-depth analysis.

With GuestTrack, companies can customize a guest book that allows users to store profile information and "tag" each user with a unique ID that is recalled each time they visit their Website. A feature of GuestTrack is the "include" function that allows sites to modify one file and have it instantly affect all HTML files that included that changed file. One of the frequent uses for the "include" function involves the navigation links that are repeated on practically every file within the website. Since an include file can use any GuestTrack personalization function, the included file can display profile information, modify links, or decide which piece of information to display. The feature allows for the customization of a site for every unique visitor. This is also very rich information, which can be culled and mined for insightful analysis via data mining.

GuestTrack also sells GT/Catalog, an interactive catalog system that builds upon the GuestTrack tracking and personalization software. GT/Catalog allows organizations to create web catalogs from scratch or use existing catalog product database. In addition to providing a full-featured catalog-shopping environment, the tracking and personalization features of GuestTrack allow businesses to display customized catalog pages by marrying product information with a customer's profile.

www.guestTrack.com

Amazon/AlmondNet
Exhibit 1009
Page 271

Gustos

Gustos Software develops and markets collaborative filtering Java software products and tool kits. The products are designed to integrate into existing websites to enable presentation of product recommendations or content that is personalized to each individual user. Gustos is designed to allow users to find the things they like and avoid those they don't. As with most collaborative "agent" software, Gustos is designed to be trained on visitor tastes and preferences.

The core of the system is the ability to find relationships between users, items, and ratings using collaborative filtering technology for making recommendations or predictions that can be derived by comparing a specific user's ratings to those of "like-minded" individuals. To benefit from Gustos, the user must entrust Gustos with certain information or it is impossible to get accurate results. To preserve anonymity, Gustos simply assigns a unique "cookie" ID to each visitor that chooses to enter ratings for items. Subsequent user ratings are tracked via this "cookie" so no user login is required.

The Gusto flagship product is StarRater. StarRater is a Java applet that a site can use to capture user ratings and give personalized recommendations to visitors. All data collected is stored and processed on central Gustos servers. StarRater provides several automatic functions such as "Recommended" and "Most Popular" item lists. "Gustos" is Spanish for preferences.

http://www.gustos.com

Internet Profiles Corporation (I/PRO)

I/PRO offers several auditing products to allow Websites and ad networks to deliver audited reports to their advertisers. Their flagship product is NetLine, an interactive measurement and analysis system. It automatically delivers a set of daily, weekly, and monthly reports designed to provide insight into the profile of a website's visitors and how they are using the site. Custom reports enable analysis of ad banner performance, activity of particular pages and files, and reports on domains that visit that site. NetLine is an outsourced solution and includes data management and storage, system support, and training. NetLine delivers pre-formatted presentation style reports through a user-friendly interface. Although this provider does not generate any content for mining, its reports may be of interest and provide additional insight to a large website.

www.ipro.com

Amazon/AlmondNet
Exhibit 1009
Page 272

IMGIS

The AdForce service from IMGIS is a full-service, end-to-end advertising management solution for ad networks and large websites (Figure 7–11). Like DoubleClick, IMGIS also uses cookies dispensed from member sites to positioned targeted banner ads.

AdForce is a service for the management and delivery of web advertising. Servicing the needs of agencies, websites, and ad sales networks, AdForce requires no hardware or software; it is an outsource service. AdForce provides advertisers the ability to target ads to all visitors by the type of content they are viewing, the time of day and day of the week, by keyword, by frequency (the number of times a visitor sees an ad), and by sequence (the order in which a series of ads are shown to a visitor).

AdForce Targeting Parameters	
Content	Time of Day
Day of Week	Keywords
Frequency	Sequence Operating System
Browser Type	Internet Service
Provider Domain Name	Telephone Area Code
SIC Code	



Figure 7–11 *The AdForce system from start to finish.*

Amazon/AlmondNet
Exhibit 1009
Page 273

In addition, whenever an Internet visitor requests an advertisement (from AdForce or any other ad server), the request specifies the visitor's operating system (Windows, Macintosh, etc.), browser type (Microsoft Internet Explorer or Netscape Navigator), and IP address (e.g., 205.11.55.1). Using its database of IP addresses, AdForce is able to infer additional information about the visitor. For dial-up visitors, AdForce determines the visitor's Internet Service Provider. For corporate visitors, AdForce determines the visitor's domain name (e.g., .com or imgis.com), telephone area code, and SIC code.

www.imgis.com

InText

One of the products supplied by InText is WebInterests!, a software product that automatically watches for new information as it becomes available in a website, such as a news provider, then dynamically compares it to user interest profiles and delivers live via e-mails. WebInterests! allows for "Active Notification" providing website visitors with new information specifically of interest to them. It also allows users to describe their own interests by a few sentences or a set of concepts. Users can set relevance thresholds describing whether to route documents whose content closely matches their needs, or only touches on it. For lengthy documents, WebInterests! sends users automatically generated summaries. Of interest to data miners is the content of the profiles this product is able to create.

www.intext.com

LikeMinds

LikeMinds is another collaborative filtering tool able to assist websites to make personal recommendations and to offer direct marketing based on visitors' past behavior. Its *Preference Server* delivers personalized recommendations based on either preferences explicitly stated by the visitor or customer or implicitly determined by sales records, click throughs, or other user interactions within that site.

As with other collaborative filtering system, LikeMinds works by asking users to state their preferences about multiple topics. The software compares one user's preferences to others', finds people of similar tastes, then makes recommendations based on what similar users like. The software also can factor in preferences based on what a visitor buys or looks at on a site. LikeMinds tools also generate consumer data, which it calls "cybergraphics," that

Amazon/AlmondNet
Exhibit 1009
Page 274

Websites can use for direct marketing. Cybergraphics, LikeMinds claims, allows greater accuracy for targeting interested customers because it takes into account not just demographic data or psychographic data (based on what people say they like) but also the activity of site visitors.

LikeMinds is positioning itself against Firefly, which has the first collaborative filtering technology to hit the market. Firefly issues a "Firefly passport" that registrants can use on any site that licenses its tools. The data from all those sites is compiled in Firefly's database and used to make recommendations. In contrast, LikeMinds does not require its licensees to share their data, and it does not promote its own brand. Rather than offering a collaborative filtering service, LikeMinds delivers applications websites can use themselves.

LikeMinds recently merged with Andromedia, a website analysis software company. Demographic predictions can be incorporated into LikeMinds applications as "archetypes," allowing marketers to exploit existing databases of demographic information. LikeMinds generates data, which can be mined for additional insight and for profiling visitors. LikeMinds collaborative filtering technology can be used for personalizing based on consumer taste and behavior, while additional demographics and psychographics data coupled with machine learning and neural networks can enhance the overall performance of relational marketing.

www.likeminds.com

MatchLogic

MatchLogic is an online ad firm that integrates Internet advertising, database, and direct marketing into a single service company. Exclusively a service company, MatchLogic provides the management of ads. MatchLogic's core service is *TrueCount*, which tracks and manages banner ad impressions even on pages served from a cache on a proxy server (normally one of the chief stumbling blocks to effective online advertising management). MatchLogic developed the TrueCount technology at the request of General Motors in an effort to increase measurement standards as they relate to marketing on the Web. TrueCount is a very accurate method of counting and controlling all impression views that occur on a website, page, or advertising banner—even those normally hidden from counting by proxy servers and web browser caches. TrueCount not only counts, it also permits the optimization of the ads caught in proxy caches. It does this by controlling the rotation of the ads.

Amazon/AlmondNet
Exhibit 1009
Page 275

MatchLogic has five data centers in the United States and overseas, and currently serves ads on more than 2,500 sites. MatchLogic also has expanded to develop a marketing service. The new service, called DeliverE, currently has half a million users in its database, and MatchLogic expects that number to grow to five million. The basic premise behind DeliverE is one of sending e-mail to users who have opted in and soliciting information about their lifestyles and interests. MatchLogic offers a suite of services that enable advertisers and agencies to outsource their Internet advertising campaigns. It performs as an ad insertion network for advertisers and agencies, which pay up to $4 per thousand impressions for the company to place ads. It also measures their effectiveness and delivers real-time reports. This service may be of use in the data mining of your website since it can provide a more accurate view of what is actually happening at your server.

www.matchlogic.com

MicroMass

MicroMass sells software designed to develop and sustain unique individual customer relationships. MicroMass's proprietary IntelliWeb software assembles video, audio, text, and graphic images into individually tailored communication products at mass production levels. These communication products are based on individual profiles generated from customer database information. IntelliWeb personalizes messages by dynamically creating single web pages or entire websites based on information provided by each website visitor.

MicroMass links database information to individual content elements through the use of simple "if-then" statements. At runtime, the software pulls the appropriate content together based on an individual's database profile. Using this approach, messages can be sufficiently refined to provide each reader with a personal perspective. The IntelliWeb system can tie content databases to expert system-based rules/facts databases that are triggered when visitor's information is entered via a standard HTML form provided from a customer database or supplied by an individual's personal IntelliWeb Profile (Figure 7–12).

MicroMass claims it can create each web presence in real-time, based on the current *individual* profile of each website visitor. With the IntelliWeb Plug-In and IntelliWeb Profile Administrator, end-users can construct confidential profiles that automatically create personalized versions of an IntelliWebsite when the user enters. As more information is passed to the IntelliWeb Server, each message becomes increasingly more relevant and timely. MicroMass uses proprietary

Amazon/AlmondNet
Exhibit 1009
Page 276

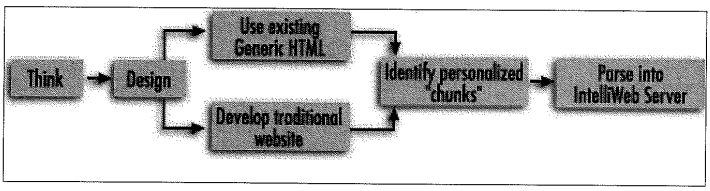


Figure 7–12 *MicroMass's IntelliWeb links web content to web profiles.*

HTML tags to identify which areas to personalize, adding content as desired. This process breaks up a website into "chunks." When a user hits the site, IntelliWeb pulls together appropriate chunks in real-time based on that user's individual profile. The resulting content can also be mined in order to build lifetime value for each of these customers.

www.micromass.com

NetPerceptions

NetPerceptions is another software provider that allows for the profiling of website visitors based on real-time interaction and dynamically changes the contents of the website accordingly. NetPerceptions use strategies in which they ask customers to name favorite albums, for example, then predict other selections users might like. It is mining in real-time, based on purchase data customers provide and using that data to make recommendations.

The core technology of NetPerceptions' system is their GroupLens and Learning Intelligent software, which learns personal preferences about an individual. Working behind the scenes, GroupLens software observes what visitors actually do at a Website, rather than relying solely on what they may say about themselves in psychographic or demographic profiles. Preference information can be collected explicitly by directly asking individuals about their opinions on a particular subject. Information may be gathered implicitly, by inferring interest from specific actions and behaviors of visitors at a website using NetPerceptions software.

Learned personal information is used to predict how desirable a customer will find particular items. The information is matched up in real-time against a community of other individuals who have demonstrated similar purchase or preference behavior. This allows certain predictions to be formed about likes and dislikes. In addition, other prediction tools (e.g., expert reviews, market summaries) can be included in the GroupLens database and applied to the individuals for whom they add value.

Amazon/AlmondNet
Exhibit 1009
Page 277

Predictions are used to recommend products or services for consideration or promotion. The knowledge gathered by the GroupLens Recommendation Engine is presented to an online business for their further action. The business is then given the final opportunity to select only what it wishes to present or recommend to its customer. Obviously, the GroupLens database presents an excellent source of information that is ripe for data mining.

www.netperceptions.com

Vignette

Vignette Corporation provides software for sites who need to provide dynamic content-driven service applications, such as promotion, publishing, community, and customer support. Vignette flagship product is StoryServer4, which consists of four components to attract, engage, and retain website customers.

- *Lifecycle Personalization Services* is based on observed behavior and explicitly stated preferences. Three agents enable lifecycle personalization—Presentation Agent, Matching Agent, and Recommendation Agent. Each is useful at a particular stage of the customer relationship lifecycle, from arrival to the stage where the visitor becomes a trusting customer. The Recommendation Agent uses a collaborative-filtering technology based on Net Perceptions' GroupLens Express.
- *Open Profiling Services* has three key capabilities: content catalog, visitor registry, and the observation manager. These features manage and populate a centralized repository of visitor profile and content information. The observation manager gathers and stores information about visitor activity and behavior.
- *Advanced Content Management* provides advanced support for XML-based content, dynamic versioning of websites, and broad platform support for databases and Web servers.
- *Business Center.* This component provides an environment for managing and reporting on visitor profile and behavior data collected with the Open Profiling Services. A variety of reports can be created to gain a deep understanding of customer behavior and to fine tune the personalization capability of the website.

StoryServer 4 provides open access to a datamart of customer information for integration with other enterprise applications, including data mining analysis. The features of this product can be integrated with additional demographics and best of breed data mining tools.

www. vignette.com

Amazon/AlmondNet
Exhibit 1009
Page 278

Your Data Warehouse

Your company's data warehouse represents an additional source of information, which although it is not an external source, it can be linked to your website for data mining. Your data warehouse supports informational processing by providing a solid platform of integrated, historical data from which to do analysis, including of course customer profiling of online visitors and shoppers. One of the most valuable sources of customer information is that which is already in your own data warehouse. For this reason you should consider this as an additional resource that can be utilized for insight into the identity of your website visitors. For example, you may find certain inherent demographics in your existing clients which match those of certain visitors to your company's site.

A data warehouse provides the facility for integration of disparate application systems, from marketing, sales, operations, shipping—and now a company's website. A functional definition of data warehouse is that it must be a subject-oriented, integrated, and nonvolatile collection of data for decision support. One of the most important features of a data warehouse is that it is oriented toward the major subjects of the enterprise, such as its customers. A data warehouse is thus not only data-driven but subject orientated—in contrast to the classical data applications, which are organized around processes. For example, a bank's processing systems are built around such functions as loans, savings, bankcards, ATMs, marketing, real estate, sales, and investments. On the other hand, a bank's data warehouse is organized around a single major subject: its *customers.*

The differences between processing applications and subject applications show up quite differently in the content and detail of the data. A data warehouse contains very rich information about who its customers are and for this reason it can be used to match the demographics of website visitors. Data warehouse data is a long series of customer snapshots, which can be compared to those of your website visitors for similarities. For example, a bank can merge the data from its website with data from its warehouse to develop probability models and propensity profiles for one-to-one marketing of specific financial products, such as credit cards, bank loans, and CDs.

You should consider importing your web data into your data warehouse for enhancement with other customer files and mining. The linking of your data warehouse with your website represents the merging of the ultimate data depository with the ultimate marketing tool and delivery channel. On the Web your ability to target con-

Amazon/AlmondNet
Exhibit 1009
Page 279

sumers leads to great efficiencies. You not only have the ability to reach only specific target audiences (avoiding waste), but also to learn more about your consumers so that you can service them more efficiently in the future. It is this process—the building of a one-to-one relationship—that can truly make your website a success.

Closing the Loop

The building of customer profiles comes not only from data mining but also from the data warehousing technique of constructing the object of your analysis from disparate components and sources. These customer attributes may well come from demographic providers as well as components created by collaborative software and finally your own warehouse. Your advertising, marketing, retailing, and business needs, as well as the products and services that you are providing, drive the components that you use in the profiling of your website visitors. For example, some information is slanted toward predicting financial behavior, such as Claritas' P$YCLE, while some is geared toward predicting online behavior, such as the content provided by Engage Technologies. Ideally you may want to assemble components from both types of providers to gain the optimum quality in your data for mining.

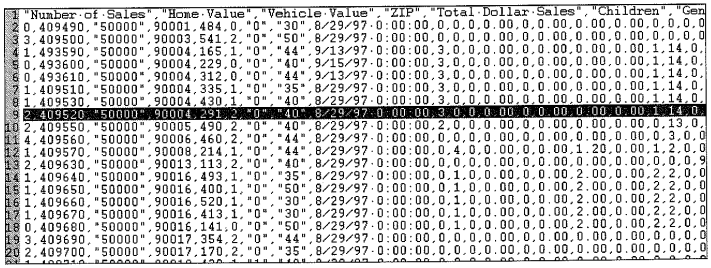
Amazon/AlmondNet
Exhibit 1009
Page 280

8

E-Retailing

Through data mining a company can synthesize website patterns into meaningful information, enabling it to individually distinguish, understand, and engage customers and prospects over the Web (Figure 8–1). The mining of transactional data and the vast flow of business intelligence it represents is the key to creating a lasting relation with online customers and establishing a productive website. The instant interactivity of web transactions breeds accelerated change in product design, delivery, and customer service. Make no mistake, electronic retailing is here to stay, and buying and selling will never be the same—for every transaction and "look" can now be digitized, aggregated, visualized, averaged, and (most importantly) mined.

In a networked environment with instantaneous connections, there is a huge premium on instant response and the ability to learn from and adapt to the marketplace in real-time. Winning electronic retailers accept this current culture of constant change and are willing to constantly break down and reconstruct their products and processes to evolve to their customers' expectations. Remember that data mining is about perception and learning; it is about artificial intelligence technologies being used to look for patterns in this fast changing business environment. Data mining is about using artificial intelligence technologies for business competitive intelligence. It is



	"Number of Sales"	"Home Value"	"Vehicle Value"	"ZIP"	"Total Dollar Sales"	"Children"	"Gen
1							
2	0.409490	"50000"	90001.484.0	"0"	"30"	8/29/97 0:00:00	0.0.0.0.00.0.00.0.00.0.00.0.0.0.0
3	3.409500	"50000"	90003.541.2	"0"	"50"	8/29/97 0:00:00	0.0.0.0.00.0.00.0.00.0.00.0.0.0.0
4	1.493590	"50000"	90004.165.1	"0"	"44"	9/13/97 0:00:00	3.0.0.0.00.0.00.0.00.0.00.1.14.0.
5	0.493600	"50000"	90004.229.0	"0"	"40"	9/15/97 0:00:00	3.0.0.0.00.0.00.0.00.0.00.1.14.0.
6	0.493610	"50000"	90004.312.0	"0"	"44"	9/13/97 0:00:00	3.0.0.0.00.0.00.0.00.0.00.1.14.0.
7	1.409510	"50000"	90004.335.1	"0"	"35"	8/29/97 0:00:00	3.0.0.0.00.0.00.0.00.0.00.1.14.0.
8	1.409530	"50000"	90004.430.1	"0"	"40"	8/29/97 0:00:00	3.0.0.0.00.0.00.0.00.0.00.1.14.0.
9	2.409520	"50000"	90004.291.2	"0"	"40"	8/29/97 0:00:00	3.0.0.0.00.0.00.0.00.0.00.1.14.0.
10	2.409550	"50000"	90005.490.2	"0"	"40"	8/29/97 0:00:00	2.0.0.0.00.0.00.0.00.0.00.0.13.0.
11	4.409560	"50000"	90006.460.2	"0"	"44"	8/29/97 0:00:00	0.0.0.0.00.0.00.0.00.0.00.0.3.0.0
12	1.409570	"50000"	90008.214.1	"0"	"44"	8/29/97 0:00:00	0.4.0.0.00.0.00.0.00.1.20.0.00.1.2.0.0
13	2.409630	"50000"	90013.113.2	"0"	"40"	8/29/97 0:00:00	0.0.0.0.00.0.00.0.00.0.00.0.0.0.9
14	1.409640	"50000"	90016.493.1	"0"	"35"	8/29/97 0:00:00	0.1.0.0.00.0.00.0.00.2.00.0.00.2.2.0.0
15	1.409650	"50000"	90016.400.1	"0"	"50"	8/29/97 0:00:00	0.1.0.0.00.0.00.0.00.2.00.0.00.2.2.0.0
16	1.409660	"50000"	90016.520.1	"0"	"30"	8/29/97 0:00:00	0.1.0.0.00.0.00.0.00.2.00.0.00.2.2.0.0
17	1.409670	"50000"	90016.413.1	"0"	"30"	8/29/97 0:00:00	0.1.0.0.00.0.00.0.00.2.00.0.00.2.2.0.0
18	0.409680	"50000"	90016.141.0	"0"	"50"	8/29/97 0:00:00	0.1.0.0.00.0.00.0.00.2.00.0.00.2.2.0.0
19	3.409690	"50000"	90017.354.2	"0"	"44"	8/29/97 0:00:00	0.0.0.0.00.0.00.0.00.0.00.0.0.0.0
20	2.409700	"50000"	90017.170.2	"0"	"35"	8/29/97 0:00:00	0.0.0.0.00.0.00.0.00.0.00.0.0.0.0

Figure 8–1
Data mining is about recognizing the signature of sales and the customer who made it.

Amazon/AlmondNet
Exhibit 1009
Page 282

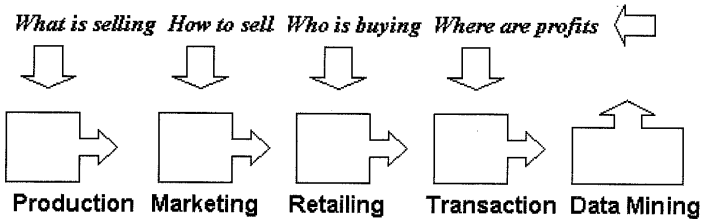


Figure 8–2 *The cyclical relationship between data mining and electronic retailing.*

also about evolving, and leveraging technologies and tools to quickly adapt to the flow of change of a digital networked marketplace that keeps expanding and accelerating.

The Web represents a unique retailing channel that by its very nature establishes a new type of foundation and structure for a close relation between the seller and the buyer. Electronic retailing enables the retailer to collect information about the consumer's tastes, rate of consumption, and preferences for products by quality, price, and volume—as they make their purchases. Unlike other retailing channels, the Web facilitates a one-to-one selling environment, allowing for tight interactions between retailers and consumers. Electronic retailing allows for two-way interaction, which feeds directly into a cycle of production, marketing, retailing, transaction, and mining (Figure 8–2). Data mining is an integral part of this retailing cycle, for it provides important insight into the consuming process, tendencies and attributes of customers. Data mining can answer the *what, how, who,* and *where* of electronic retailing:

The typical retailing distribution networks involving distributors, agents, and other middlemen are eroding. Already changes are starting to take affect in the buying and selling of airline tickets, stocks and mutual funds, car rentals, hardware and software, books, CDs, and a large range of digital products—everything from market reports to pornography. Electronic retailing changes not only the way products and services get distributed, it also impacts pricing—due in part to the customization of products and services—and the ability to use data mining technology in the analysis of consumer preferences. As the Web blurs the traditional concepts of retailing, the conventional relationships between consumers and retailers are gradually changing.

For example, I recently I purchased a sport utility vehicle, for which I had done some preliminary window shopping on the Web—some virtual tire kicking. By the time I walked into the physical auto

Amazon/AlmondNet
Exhibit 1009
Page 283

dealership I knew more about that truck than the salesman did. I knew what engine options, colors, power features, and configurations the SUV came in, far more than the salesman did. I had already done some comparisons of several SUVs and knew exactly what I was buying. The Web makes consumers very informed buyers—more knowledgeable at times about product and services features, options, selections, and pricing than the traditional sales personnel.

The Product Is the Message

Unlike traditional retail channels, selling and buying on the Web truly take on the Marshall McLuhan adage, *"the medium is the message."* On the Web a digital product such as an e-mail message can contain as its content some digitized intelligence such as a company report, some market information or a stock quote, etc., and may be sold directly as a commodity or conversely used as a marketing component. On the Web "the message" can also be "the product" of the transaction or the medium for "the marketing offer." In the electronic retailing of digital products, the value of the transaction may well be in the content of the message. With the voluminous number of interactions that are occurring daily on the Web, McLuhan may well have seen it coming when he predicted that the only recourse to survival in our modern world would be pattern recognition.

Electronic retailing changes not only the distribution and marketing of products but more importantly it also alters the process of consumption and the related transactions of buying and selling. The data, which is an aftermath of every product and service purchased on the Web, is the ore that can be mined to develop customized products, forecast demand, profile customers, and improve relational marketing. Because of the interactive nature of electronic retailing, consumers not only order and buy products online, they also search for product specifications (thus revealing their preferences and criteria) and in some venues their willingness to pay.

The act of retailing on the Web is an interactive one in which the consumer can negotiate, exchange information, specify, and customize the product and services he or she wants from the retailer. For the electronic retailer it is of paramount importance that it analyze what consumers are doing and saying. Data mining can serve retailers by providing them the technology to understand how to sell more, learn what's working, and learn what's not. It can assist them in

Amazon/AlmondNet
Exhibit 1009
Page 284

knowing how to personalize their storefront for repeat customers. As we discussed priorly, a data mining technology such as case-based reasoning (CBR) is similar to "collaborative filtering" software in that it can serve as a "recommendation engine" to suggests items that might interest a shopper, based on what similar shoppers are buying.

The potential for electronic commerce is immense due to the fact that a large number of services and products can be digitized. For example, all newspapers, magazines, articles, journals, and audio, video, and multimedia products can be sold online. This means you can use the Web to get stock quotes and buy movies, stamps, white papers, currency, concert tickets, hotel and airline reservations, and checks and other financial products. In fact products that require a minimum amount of descriptive information to advertise and of which there is a high number of different types and are cheap to ship are ideal for selling on the Web—books are the prime example.

In the traditional model of retailing, middlemen or bottlenecks controlled access to information between sellers and buyers. For example, insurance and travel agents typically control the access to information about claims, policies, and flights; they in turn processed inquiries and controlled pricing. However, this is changing today: with the advent of electronic commerce, the consumer is in control and can dictate the product they want to purchase and with what terms and conditions.

Data mining can be used to model patterns of behavior based on both the attributes of past web sessions and demographic information merged with the preference and customer data you gather at your website via your log files and registration forms. Propensity scores can be generated from the analysis of this data in order to predict the likelihood a visitor or customer will select an ad, banner, or a particular page in your website. Propensity to purchase the products and services on your website can also be modeled and scores can be generated in order to prioritize potential prospects for e-mail and special incentives and offers.

Figures 8–3 and 8–4 illustrate a sample session involving the development of a model designed to predict the selection of a specific page based on an analysis of thousands of sessions with demographic data matched against the website visitors' ZIP code as captured at a registration form. The data mining tool incorporates a neural network to construct a propensity to purchase model.

C code can then be generated from the data mining analysis for the scoring of propensity to buy for new unseen visitors and customers.

Amazon/AlmondNet
Exhibit 1009
Page 285

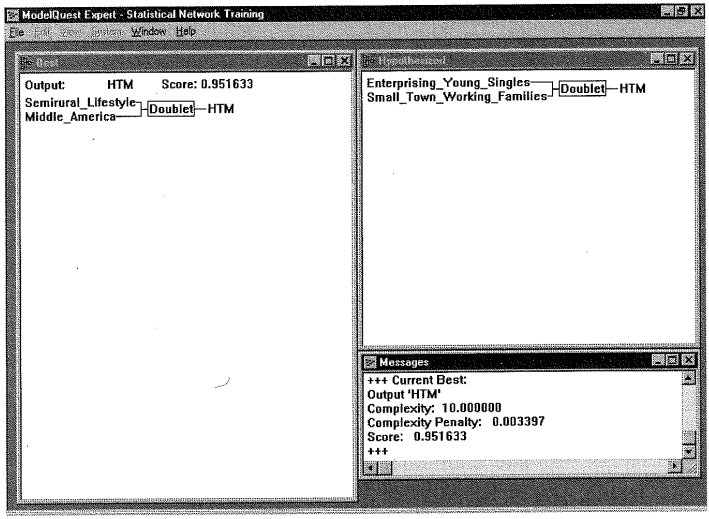


Figure 8–3
The network begins to slowly evaluate several input variables.

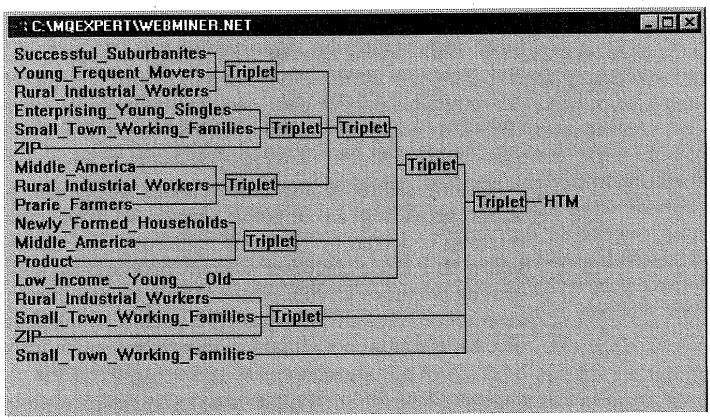


Figure 8–4
Eventually the data mining tool converges on a model.

```
double *webminer(double inarr[]){
double Successful_Suburbanites = inarr[4]; /* rename input for clarity */
double Product = inarr[45]; /* rename input for clarity */
static double outarr[1]; /* declare output array */
double HTM    ; /* rename output for clarity */
double node5  ; /* working variable */
/* node5-Successful_Suburbanites */
```


Amazon/AlmondNet
Exhibit 1009
Page 286

```
   node5  = -1.19493 + 0.032543*LIMIT(Successful_Suburbanites,0,82.9);
    /* node5-Triplet */
   node5  = 0 + 5.19459 + 15.1832*node14 - 61.5623*pow2(node14)
          + 15.62*pow3(node14) - 4.98188*node34
          - 13.6275*node14*node34 + 60.8895*pow2(node14)*node34
          + 0.0592666*pow2(node34) + 0.00802485*pow3(node34)
          - 4.45746*node44 - 25.4388*node14*node44
          - 20.8693*pow2(node14)*node44 + 0.519106*node34*node44
+ 0.105443*pow2(node34)*node44 - 3.17381*pow2(node44)
     - 18.6362*node14*pow2(node44) + 0.408554*node34*pow2(node44)
          + 0.295813*pow3(node44) ;
HTM = 0 + 75.1651 + 2.7993*node48 ;  /* perform output limiting on HTM */
HTM = LIMIT( HTM, 69, 79 );  outarr[0] = HTM ;  /* output value */
return( outarr ); /* output array */}
```

Sale Segmentation

To establish a competitive presence on the Web today and in the future you need to design your site to interact with your visitors. Your website is an antenna for receiving, not just a medium for pushing products and data. Your retailing site needs to establish a dialogue with your clients to better position appropriate messages, products, and services. Your website increasingly assumes a mission critical role for your company—it provides not only a new marketing channel, but more importantly an opportunity for discovering who your current and potential new customers are. The combination of data mining technology in a networked environment totally changes the traditional concepts of marketing and commerce. This is because electronic retailing changes the fundamental buyer and seller relationship.

Online selling is not simply about dispensing products. It is also about providing an ongoing service to your customers. Electronic retailing is driven by relations marketing and data mining; it is about selling to a single customer as many products as possible over their consuming lifetime. Online commerce is also about learning about your customers' likes and dislikes with each sale. Subsequently, the content of online transactions matters more than in traditional retailing. Each interaction and purchase is an opportunity for the electronic retailer to capture data about its customers and learn about their preferences. Electronic retailing is about encouraging continuous contact with your customers and incrementally improving your relations with them. In a digitized networked environment anything that tightens your relation with a customer increases your revenue potential.

The size of firms doing electronic retailing is not a significant factor in this type of marketplace. What is most important is the ability of the

Amazon/AlmondNet
Exhibit 1009
Page 287

smart retailer to personalize their responses to each of their customers' needs and desires. An electronic retailer of any size can be competitive in this marketplace as long as it is capturing important transactional and customer information and leveraging data mining to enhance the service and quality of communications it provides to its clients.

On the Web shoppers can search for product information and do side-by-side comparison in a completely open marketplace. Online buyers have a dramatic new power to barter, bid, and compare for the best prices on products and services. Consumers today are assisted by new services, websites, and software that can quite easily retrieve the best available prices. In this type of marketplace your competitor is always a mouse-click away.

To stay competitive and attract and retain customers, it is important to provide an extra-added service along with low costs. You must personalize your products and services, making it difficult for customers to leave. Those firms that rely on physical barriers or traditional brand advertising and retailing channels to combat competition will eventually fail in this fast evolving environment. The winners will be those merchants who recognize the value of transactional data and mine it for all it's worth.

Monitoring the relationships with your customers in a digitized environment involves the mining of your data. This entails an analysis of all repeat visitors as well as those who have made purchases but have stopped buying from you. Periodically you may need to do a data mining analysis of your dormant customer accounts to investigate why they have stopped buying from you. In order to ascertain what kind of product they may be interested in, do a clustering analysis of products purchased by visitors with similar demographics. Make a special offer to encourage dormant customers to take the time to answer your "please come back" e-mail. Send a "we missed you" e-mail message, and ask them why they have not returned since their last purchase of "Product Z." Tell them of a special offer for a similar product they may be interested in based on your data mining analysis.

Cultivating existing customers is the most cost-effective strategy you can pursue in retailing on the Web. They represent visitors who appreciate your products and services, and you need to establish a strategy for retaining them. Do this via e-mail informing them of products they are likely to be interested in based on prior purchases and the interests they have indicated on their registration forms or through their prior interactions. Of course, data mining can also reveal what product they are likely to be interested in based on affinity analysis using association and segmentation tools. Cultivate your

Amazon/AlmondNet
Exhibit 1009
Page 288

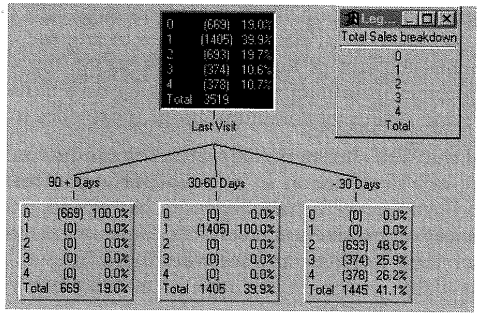


Figure 8–5
This decision tree analysis reveals no sales for visitors with Last Visit over 90 days.

current customers, prize them, and reward them for buying from you again and again. A data mining analysis of your interaction with your customers can reveal those clients who represent most of your sales as well as those who are effectively dormant and who you may want to e-mail for purposes of making them profitable again (Figure 8–5).

This data mining analysis used a segmentation tool to generate a decision tree view of customers by the number of sales they have made in relation to the number of days since they last visited this website. For example, this decision tree found that:

IF Last Visit = –30 Days

THEN

Total Sales = 2 48.0%

Total Sales = 3 25.9%

Total Sales = 4 26.2%

The propensity to buy deteriorates quite quickly on the Web when interactivity drops, which can be remedied via an e-mail campaign. Effective e-mail interaction, however, should not be perceived as intrusive and you should ask for permission from your customers prior to sending them any e-mail. Offer legitimate incentives and send only quick personal messages. Ensure that you allow customers the option to discontinue receiving messages and include your address in the body of the message for single click access to your website. Lastly, don't keep sending messages to nonrespondents and don't send the same message more than once. There are several mass mail systems that can help you automate some of this process.

Amazon/AlmondNet
Exhibit 1009
Page 289

The Digital Marketplace

In the digital marketplace the buyers come to the retailer, and advertising is iterative with the consumer interacting with the ad and participating in the content of the message. Because of the digital nature of the banner or ad on the Web, selection of the audience and measurement of the participants of the advertisement are relatively easy. One of the unique abilities for shopping in this marketplace is the ability to automate the indexing and cataloging of product information. Queries for product information generate a trail of information about consumer demand and tastes, which can be mined. Data mining of this product and preferences information can result in the refining of your inventory and product development. It enables electronic retailers to eliminate waste and provide targeted products designed to meet specific consumer preferences. Measurement of clickthroughs or the requests for additional product information via a form can be stored for data mining analysis.

The mode of shopping in this marketplace is driven by "keyword" searches, which the retailer may use to position itself and its advertisements at the major directories like Lycos. Search engines such as Infoseek are positioned to not only guide consumers toward retailers, but more importantly they can monitor visitor activity and preferences. These engines can monitor consumer preferences based on keyword searches they make and associate them with the visitor information they provide in the personalization of their news, weather, stocks, etc., especially at such popular websites as my.yahoo.com, my.excite.com, or my.whatever.com (Figures 8–6 and 8–7). These search engines can become great sources of cross-reference information about consumers. The engines not only can customize products for the consumer, but by revealing their preferences consumers can drive the design of products and pricing.

The essence of targeted marketing on the Web lies in the ability for consumers to interact with retailers and the search engines that they use to browse the marketplace. Advertising on the Web can be driven by the consumers' choice of search keywords, allowing for real-time targeting which can in turn be linked to data mining, market research, distribution, and sales. Data mining can be used to analyze the consumer information gathered by these engines in order to customize the services and content they provided to their visitors. Retailers are better able to provide customized products and services based on specific preferences, rather than average tastes. As the Web grows in size and complexity, consumers will increasingly reveal their preferences

Amazon/AlmondNet
Exhibit 1009
Page 290

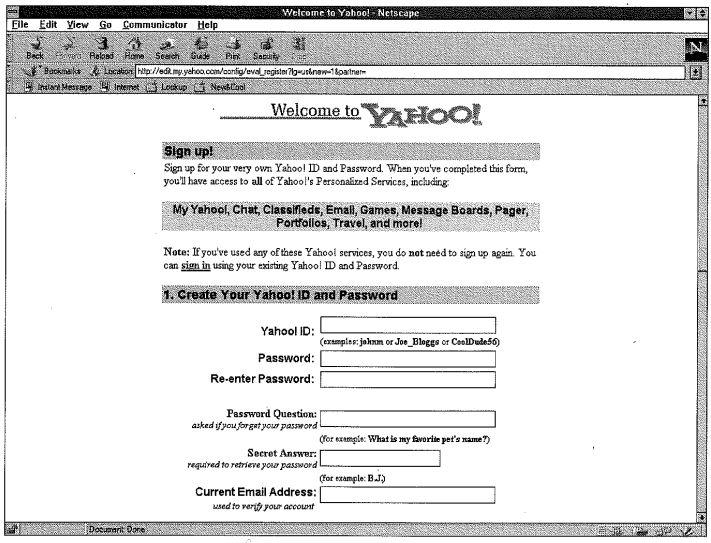


Figure 8–6
The personalization of the Web can target the content you receive.

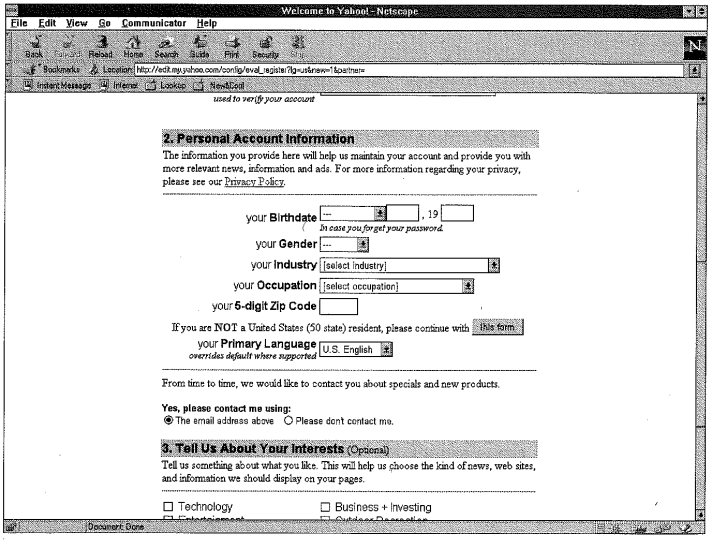


Figure 8–7
The information you provide about yourself can be mined.

Amazon/AlmondNet
Exhibit 1009
Page 291

in order to get above-average quality in their products and services quickly and efficiently and in order to dictate selection and pricing.

This type of marketplace enables consumers to deal directly with manufacturers of products and services, eliminating middlemen. The Web reduces the number of intermediary steps necessary for doing business and trade. The Web hardwires the topography in which the buyers and retailers exchange goods directly instead of through intermediaries, creating a more efficient market. In this type of marketplace the creator and seller of products and services is able to deal directly with consumers and thus know who they are and what their preferences are. Electronic retailing radically changes the manner in which products are distributed, sold, and priced. The costs associated with the production of personalized digital products, such as for reports, news, or quotes, can vary for each customer.

This type of networked market in which buyers can deal directly with sellers seriously threatens traditional distributors and agents. But a new brand of middleman software or search agents may be created as the Web continues to expand and electronic retailing grows. As the amount of information clutter increases, new forms of agent software or specialists using special data mining technology will evolve to sift and sort through massive amounts of data into usable consumer information, or to negotiate pricing. These new agents or rovers will offer aggregated services, or intelligent customer assistance using data mining technology to provide product-based searches based on prearranged prices or consumer conditions. These new middlemen will sift through the ever-expanding digital marketplace in search for the products and services their clients have commissioned them to locate at the price they are willing to pay.

Product Customization

Because products and services can be customized to meet the specific needs of consumers a high degree of satisfaction can be obtained in this type of marketplace. In turn, the consumers obtain a high amount of value from customized products than traditional off-the-shelf products. Data mining can be used to converge on consumer preferences and assist in the customization of product design. Products can be manufactured to match the specific demands of specific consumers, which in turn leads to a high degree of customer retention and loyalty.

A high degree of product customization discourages customer attrition or churn. Customization of products and services discour-

Amazon/AlmondNet
Exhibit 1009
Page 292

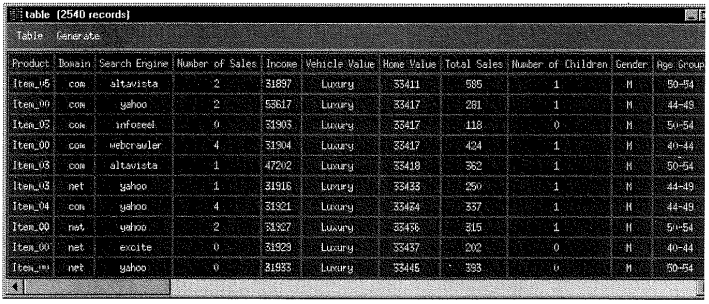


table [2540 records]

Table Generate

Product	Domain	Search Engine	Number of Sales	Income	Vehicle Value	Home Value	Total Sales	Number of Children	Gender	Age Group
Item_05	com	altavista	2	31897	Luxury	33411	585	1	M	50-54
Item_00	com	yahoo	2	53617	Luxury	33417	281	1	M	44-49
Item_03	com	infoseek	0	31903	Luxury	33417	118	0	M	50-54
Item_00	com	webcrawler	4	31904	Luxury	33417	424	1	M	40-44
Item_03	com	altavista	1	47202	Luxury	33418	362	1	M	50-54
Item_03	net	yahoo	1	31918	Luxury	33433	250	1	M	44-49
Item_04	com	yahoo	4	31921	Luxury	33424	357	1	M	44-49
Item_00	net	yahoo	2	31927	Luxury	33436	315	1	M	50-54
Item_00	net	excite	0	31929	Luxury	33437	202	0	M	40-44
Item_[illegible]	net	yahoo	0	31933	Luxury	33446	393	0	M	50-54

Figure 8–8
Your web server data can be merged with demographic information for mining.

ages consumers from reselling or sharing of products, which they have developed, based on their unique needs. For example, my interests in music or books is not likely to be the same as yours. This is especially advantageous for real-time digital products, such as stock reports, which can be produced easily by the retailer—allowing for the pricing to be set on the basis of the consumers' willingness to pay. However, in order to effectively customize products and charge individualized prices, the retailer is required to have detailed knowledge about the consumers' preferences. This type of consumer knowledge and customer profiling can be obtained through the strategic use of data mining technology and techniques.

The economic impact of personalization and customization on the Web is accelerated when data mining technology is used. It's a one-on-one game. Information is easier to customize than hard goods. The information portion of any good or service is becoming a larger part of its total value. Thus, suppliers and retailers will find it easier and more profitable to customize products. In this type of electronic marketplace consumers will gradually begin to demand this type of customized tailoring from the products they consume. With increased customization of products and services, homogeneous pricing will no longer be practical or applicable. It will also mean that consumers will come to accept the fact that in order to receive premiere service and custom products, they will need to share personal preference information with their trusted retailers—personal customer information which can be mined in order to maximize the efficiency and value of the products and services being sold online.

In a traditional marketplace where products and services are sold by various competitors, they tend to be similar in functionality with the only difference between them being their price. However, if products and services are customized then the retailer can negotiate a

Amazon/AlmondNet
Exhibit 1009
Page 293

price with each consumer. Customization also means product differentiation is based on the degree of dissimilarity. Customization means retailers differentiate products and services by adding extra value and thus reducing customer attrition. Time, for example, can be a function by which a product is priced. A product's timeliness in a networked environment and its detail level or response time can be used to determine its value to the consumer. All of these pricing and customization factors can best be achieved by the use of data mining to calibrate their degrees of detail, value, and time.

In a digitized marketplace retailers are able to gain detailed information about their customers' attributes and preferences in order to customize products and services without much added cost. In addition, since products are custom built, customers can be billed independently, which in the traditional marketplace would be difficult if not impossible to do, due to product variation and elaborate billing costs. However, in a digitized marketplace, transaction costs are low and the retailer has information about each consumer valuation and is easily able to negotiate prices independently on a one-to-one basis. In a digitized marketplace it is easy for the retailer to gather information, which can be mined in order to learn about the consumers' purchasing behavior. In the end, marketing, retailing, and data mining in a networked and digitized environment mean that the emphasis is on the interaction between the retailer and the consumer.

Consumer Profiling

Database marketing has always relied on the use of consumer information to target offers via mail. Consumers are often required to disclose and provide some form of personal data in most transactions involving the extension of credit or the use of noncash payment, both of which leave a trail of personal information that is often used and then re-sold to third parties who dissect, cross reference, and aggregate the data evolving into secondary consumer information often sold again by such companies as Experian and Polk. Usually this information is aggregated at the household level, which can vary in detail, but may contain the value of your home, the type of auto you drive, and the number of children that you have.

Yet another type of profiling involves the granting of credit, in which case a financial transaction is involved which legally allows the retailer, such as a renter or auto dealership, the right to obtain an individual's credit consumer report. There are several types of reports

Amazon/AlmondNet
Exhibit 1009
Page 294

assembled by Trans Union, Equifax, and Experian (formerly TRW), which aggregate and assemble this information in different formats:

- *An individual identity profile:* This includes names and aliases used, as well as current and associated addresses (multiple), date of birth, marital status, and social security numbers used.
- *An employment profile:* Present position description, length of employment, and previous positions.
- *A public record profile:* A record of every civil suit, judgment, tax lien, bankruptcy, and legal proceeding recorded in a local, state, and federal court involving a monetary obligation.
- *A credit history profile:* This will list every grantor account you have including credit cards, debit cards, revolving accounts, auto loans, mortgages, etc. This section will detail when the accounts were established, how much you owe, and how much you pay on a monthly basis. It will contain account activity information detailing when you were late for any payment. It will also contain a trail of inquiries made by grantors and government agencies who have requested and paid to see your credit report.

These three national credit bureaus gather extensive credit and financial data on virtually every adult in the nation, for use by banks, mortgage companies, and credit-card issuers that need the information when deciding whether to make a loan or issue a credit card. The law prohibits the release of credit data in most circumstances, unless a consumer authorizes it—explicitly by signing a release, or implicitly by applying for a loan or credit.

The electronic marketplace allows for a different type of profiling. This type of profiling is based on user preferences and online behavior, much of which is captured from log files, cookies, registration forms, collaborative software, and server networks. This type of information is used to position the appropriate ad and messages based on observed behavior and explicit selections and preferences. Electronic retailing and consumer profiling deal with digital products, which are fundamentally different from other type of consumer items. Retailers and consumers in a digitized marketplace actively interact to influence each other to determine the product specifications, quality, volume, and price.

One of the functions of digitized products and services is that of their filtering features. This would be in the form of customizing news, reports, and information feeds, or the monitoring of markets, events, or stock activity, all based on personal profiles. The ability to effectively customize products and services has always depended on the ability of the retailer or marketer to know something about the

Amazon/AlmondNet
Exhibit 1009
Page 295

likes and dislikes of the targeted consumer. If consumer tastes vary greatly, as can be the case with such products as software or books, then these products benefit from being highly differentiated. In traditional marketplaces this used to be done in part through warranty cards and market surveys. However, in today's digitized marketplace, the consumer information is hardwired to the retailer since not only does the buyer reveal the methods of distribution they require but also the type of products they want, the rate they want them at, and how much they are willing to pay for them. All are values that can be mined for the refinement and delivery of products and services by retailer.

Consider your customers and the data, which you are mining as the basis for the evolution and development of your website. Solicit feedback via e-mail from first-time buyers. Did they find what they wanted? How long did it take? What else should you be doing at your website in order to get your first-time buyers to return and buy again and again. How do you transform first-time buyers into loyal customers? Asking them is one way; another is doing an analysis of your website data. Use data mining tools such as decision trees and rule generators to segment your visitors and develop profiles of your existing customers. Use these analyses to find out what your visitor information requirements are. How did they arrive at your site? Are they looking for a leisurely browsing experience or the fastest path to a specific product or service at your site?

Your data mining analysis may identify several different market segments that are attracted to your website, which may impact how you organize your site. For example, a travel site may find that it is attracting travel agent professionals, people who do not travel much, die-hard travelers, and investors. This may point out the need to design different website sections in the home page targeted to these four different types of visitors. Data mining may also uncover what visitors are browsing for by analyzing the keywords they are using in their quest for product information. Poring over log files, registration forms, and e-mail can uncover some hints; however, using a data mining tool can accelerate this process and point out hidden patterns manual inspection may not detect, especially for a heavily visited site with thousands of daily visitors.

Electronic retailing is really about linking your back-end inventory database to the browsers of your visitors and customers. In the process, however, you have the opportunity to observe the interactions and patterns which can evolve over time. Data mining can assist you in uncovering these patterns and assist you in improving your retailing efforts and your relationship with your customers. With so

Amazon/AlmondNet
Exhibit 1009
Page 296

much choice on the Web, the key differentiation is personalization. The mining of your customer preference information is the key to improving your relationships and personalizing your products, services, and communications with your visitors. Your data mining analysis can assist you in developing new insights into your visitors' behavior and your website's performance in meeting their needs.

Real-Time Buying

A large percentage of Americans have purchased something online, so that despite the fact that people see security concerns as a drawback to shopping on the Web, these concerns are evidently not enough to halt the wheels of commerce or hinder the growth of electronic retailing. For those who do buy online, benefits like not having to go to the store and around-the-clock shopping outweigh the disadvantages or concerns about privacy and security. There is also a new business model, which is prompting online transactions: online bidding auctions. The stimulating forces of greed, cheap prices, and getting something for less are driving this model of online retailing.

Single-click convenience and choice make up yet another business model driving the increasing prominence of small retailing products, such as book and music purveyors like Amazon.com. Electronic retailers continue to address security issues, since consumers who haven't bought through the Web say the primary reason is concern about giving out their credit card number. However, consumers are beginning to realize the advantages of electronic commerce, that is, that every product is available everywhere. The gap between desire and purchase has closed. The shelf space of the Web is unlike any other in that it has no bounds.

Artificial constraints on choice are replaced by the ability to purchase the precise product you desire, anytime, anywhere. The impulse to buy and the purchase itself used to be separated by a combination of physical and mental barriers. No more with the Web. Today you can discover the "right" product you desire and hit the "buy" button and the product will be FedExed to you that day. The entire retailing processes for marketing, sales, and fulfillment are merging into a single digitized transaction, which is recorded, stored, and made ready for data mining.

The ability to greet former customers with a personalized message or offer is possible with the utilization of a multitude of techniques and tools. These components include your website log files, server,

Amazon/AlmondNet
Exhibit 1009
Page 297

and browser site cookies, and registration forms. The data in your website can further be enhanced by merging or appending it with household and demographic databases, or by the use of collaborative software and the networking with other websites. Of course, to gain the most value from all of this customer information the use of powerful pattern-recognition tools can allow you to know your client at a one-to-one level.

The more customized and personalized your website is, the more products and services you are likely to sell. For example, a bank may mine the data from its website and data warehouse and discover a correlation between high account balance customers and specific pages for products such as investment services or other types of bank accounts. This analysis may also reveal higher-than-average click-throughs for certain ads and banners with such messages as "Make your money work for you." Through the analysis of current and recent user behavior—that is, the content visitors are looking for as well as the content they use in their e-mail or registration forms—you can make some subtle adjustments in your website. For example, are visitors using certain keywords indicating an interest in investments, such as "investment for retirement," "good stocks," or "mutual funds"? The web designer of this banking site may want to make changes or provide the type of information on planning for retirement and investments—courting to the specific interests of these visitors.

Data mining this type of information can lead you to select the appropriate advertising, product, offer, section, incentive, coupon, banner, or wording. For example, an analysis using a data mining tool incorporating a rule-generating algorithm may uncover the following pattern:

IF KEYWORD	"camping"	
OR KEYWORD	"hiking"	
OR KEYWORD	"mountain"	
AND AGE	42	
AND ZIP	94502	
AND GENDER	MALE	
THEN	Sports Utility Vehicle Ad	89%
	Sedan Ad	11%

Amazon/AlmondNet
Exhibit 1009
Page 298

The demographics breakdown for this ZIP code reveals that the following type of consumers populate it:

52.14% are very high-income married couples, age 40 to 54, with one or two children, who snow ski, collect coins, entertain at home, do aerobics, vote, read books, and contribute to PBS. They tend to buy imported beer, sport/utility vehicles, home gym systems, use copy services, and buy items by phone, mail, and the Web. They are business owners, use ATM cards to buy goods, have term-life insurance, IRAs, and belong to limited partnerships. They live in suburbs along the California coast, and the Boston-Washington corridor.

20.67% are very high-income married couples, age 40 to 54, with two or more children, who go to movies, visit zoos, go swimming and to the beach, engage in fundraising, and eat at fast-food restaurants. They buy "big ticket" furniture items, tennis equipment, video games, toys, jewelry, and boats. They have money market funds, savings bonds, passbook accounts, overdraft protection, group life insurance, and bank near work. They read business, computer, and sports magazines; listen to soft contemporary radio during commute hours; and watch NCAA basketball tournaments.

19.64% also have very high income and are between 25 and 54, with no children. They recycle products, phone radio stations, do fund-raisers, participate in environmental causes, and lift weights. They also order flowers by wire, purchase greeting cards, soft contact lenses, laptop PCs, and light beer. They have asset management accounts, VISA cards, accidental D&D, use financial planners, and have publicly held stock. They read Money, Consumer Reports, *and* Epicurean *magazines; and listen to album-oriented rock stations.*

The above is a description from one consumer psychographic company; the following is a description from another demographic database provider. Note the similarities:

37.4% live in suburbs located along the eastern seaboard and in California. They are married, middle-aged professionals at the peak of their lifetime earnings. They are more likely to belong to a health club than a country club. They pursue physical fitness through court and winter sports and golf and are heavy users of vitamins. The proportion of householders between 45 and 64 is 40 percent higher than the national average. They are predominantly white but include foreign-born residents, particularly Asians. They are well educated and have high incomes. They live in neighborhoods with single-family homes with values twice the national average.

Amazon/AlmondNet
Exhibit 1009
Page 299

62.6% have achieved success and are homeowners. They drive SUVs and are sports active. They have loans and use credit cards to buy home furnishings, apparel, toys, and travel. The average household is 3.1, most householders are between 35 and 54, and have school age children. They have dual incomes and investments. Almost a third commute across county or state lines to work.

Given the demographics of this ZIP code, think of what other ads or banners a visitor might be shown. These consumers would probably be interested in financial products such as IRAs, travel services, and business supplies, as well as sports equipment and toys. All of these ads and banners would be driven by the combined "keywords," age, and other information available to the Web analyst professional, which as we have discussed can be gleamed from external data sources, as well as internal customer files and a company's own data warehouse.

Customer Service

Data mining can also be used to provide after-sale service, to ensure your customers are satisfied and become long-term loyal clients. An AI technology known as case-based reasoning (CBR), for example, can be used to provide customer service support in order to answer trouble report questions your customer may have on a product they have purchased from you. The basic idea in CBR is really quite simple: *adapt solutions that were used to solve old problems to solve new problems.*

A CBR software engine finds those cases in memory that solved problems similar to the current problem, and adapts the previous solution or solutions to fit the current problem, taking into account any difference between the current and previous situation. Finding relevant cases, such as trouble tickets, involves characterizing the input problem by assigning the appropriate features to it, retrieving the cases from memory with those features, and picking the case or cases that match the input best. CBR software engines are learning systems that use experience to adapt their behavior to the requirements of a particular situation or a particular user. The detailed behavior of such a system, such as a help desk or customer support site, need not be completely defined in advance by a programmer or engineer.

A CBR system can be trained through the presentation of examples to adapt to the particular cases that are essential for an application, such as a product support help desk. The idea is that as a learning system, a CBR engine is much more flexible and can dynamically react to changing problems. Any new solutions not contained in the index of a CBR engine are added so that, it gradually *learns* these new solutions. So that, for example, if your customers encounter a problem with a printer they recently purchased from you, an interactive area at your website can be provided so that a "trouble ticket" session can be had with a CBR expert engine on your website. In the session shown in Figures 8–9 through 8–11, the CBR engine would prompt the visitor for certain trouble symptoms, in order to narrow down the possible solutions.

Amazon/AlmondNet
Exhibit 1009
Page 300

Customer Service *(continued)*

CBR is an index-based memory technology able to provide comprehensive knowledge management capabilities, coupled with an intuitive dialogue-based interaction. It allows your customers to solve technical problems as if they were talking with an expert. The technology provides after-sales customer care solutions, which you can use to significantly reduce your support costs and increase customer satisfaction. This type of automated customer support can translate into increased customer loyalty.

The CBR engine provides an extra-added benefit in that it can point out glitches and problems with products and services. Recurring problems that the CBR engine is time and again asked to solve can be incorporated into product enhancements and improvements. Not unlike reverse engineering, gradual product improvements can occur from the CBR process, impacting the help desk and customer service processes the technology supports.

One of the key aspects of a CBR software engine is that if it does not find a solution to a problem, it adds the new solution to its index, thus evolving and improving over time, as other data mining technologies from the field of AI tend to do. Of course, one of the beauties of the Web is that you can gradually accumulate a database of carefully written responses to common problems and gradually become more efficient in providing online customer support.

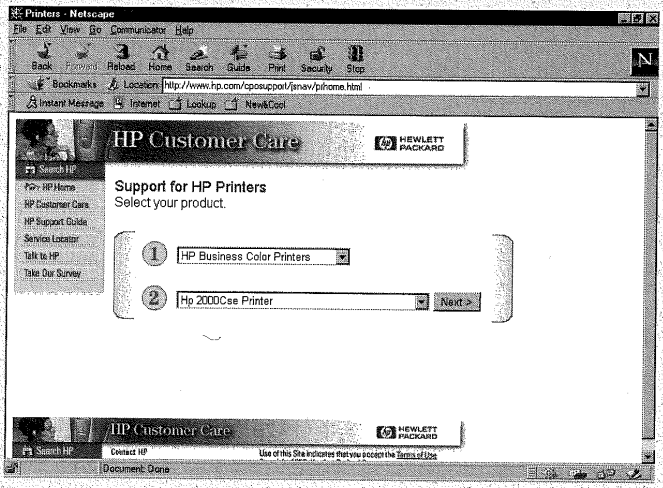


Figure 8–9
A customer service section in your website provides your clients automated support.

Amazon/AlmondNet
Exhibit 1009
Page 301

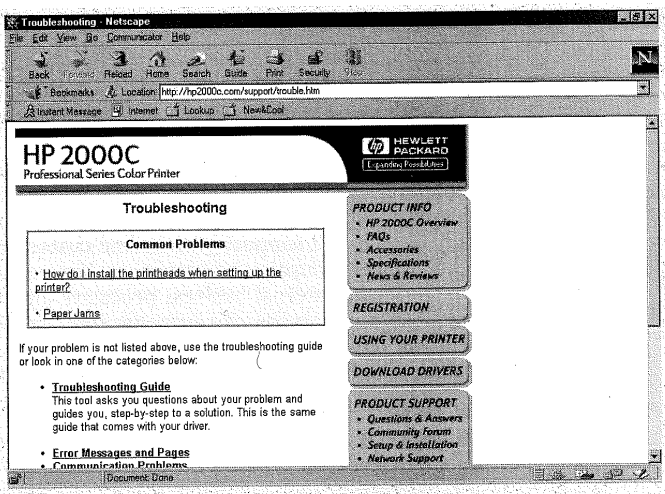


Figure 8–10 *Common problems are dealt with quickly and inexpensively.*

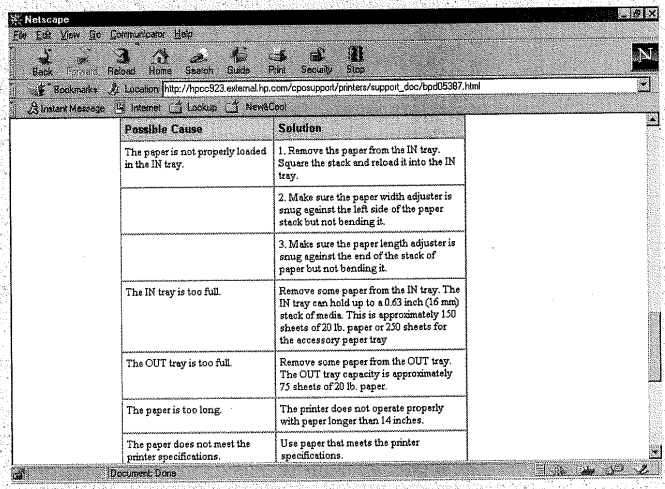


Figure 8–11 *A CBR system learns common problems and solutions.*

Amazon/AlmondNet
Exhibit 1009
Page 302

Organize Your Storefront

An electronic commerce strategy should include a clear view on your company's plans on entering and using the electronic marketplace. Since the electronic channel is destined to become the primary vehicle to conduct business in the future, your firm must learn how to attract and engage customers in it. It must plan on how it will process online orders and payments, how it will distribute products and services, and how it will support your online customers, suppliers, and partners in this new business environment. You need to consider how you will transform internal business processes into the requirements imposed by the new types of electronic interactions. As with your datamart and data warehouse, you need to plan ahead and consider what kind of decision support tools you will use to analyze your data and for what end products and processes.

Your company needs to plan and develop a strategy for doing business in a new type of neighborhood: the nonphysical marketplace. Consider how electronic commerce is going to impact your industry, your relationship with your distributors, your partners and customers, and eventually, your business. Consider what business opportunities are being created by electronic commerce in your market sector. Consider how your firm can take advantage of new electronic access to current and new customers and suppliers. What distributors and other intermediaries can be eliminated by electronic commerce? Decide what and how you want to conduct your electronic retailing, how you plan to organize your products in a database for quick browsing by your visitors, how shopping carts will be positioned, how payments will be processed, how you will facilitate one-to-one, one-stop shopping. Above all, how and what visitor information will you be capturing for subsequent data mining in order to foster a stronger relationship with them and a better shopping experience on their return?

To establish stronger relationships, offer special incentives to your online visitors. Offer them your latest product or service developments or some other unique method of benefiting your visitors that cannot be replicated via other retailing channels. How is your company prepared for processing and analyzing more and more customers as you migrate toward electronic retailing? Are you prepared to link your website to your other company databases, such as your customer information file and data warehouse? Consider what benefits such linkage can provide.

Amazon/AlmondNet
Exhibit 1009
Page 303

Microsoft's Intelligent Cross-Sell

Microsoft is incorporating data mining technology into its electronic-commerce server software, Microsoft's Site Server 3.0 Commerce edition, with a component called Intelligent Cross-Sell. ICS incorporates a data mining algorithm that can be used to analyze the activity of shoppers on a website and automatically adapt the site to that user's preferences. It's the first time Microsoft has made data mining a standard feature in one of its products. The vision for this product is that it will observe visitor patterns and use the information to completely optimize a website and reorganize the appearance of the online store to serve the customer better.

The Intelligent Cross-Sell component is based on historical sales baskets in stores, the browsing behavior of shoppers, and the contents of current shopper baskets.

The ICS module is designed to try and predict the ranking of products in electronic retailer sites most likely to be of interest to shoppers. The data mining component can be installed from a wizard in the NT Site Server. To keep online customers coming back time and time again, Site Server offers personalization and promotion features that help companies create targeted advertising campaigns and promotions according to preferences and past purchase patterns.

If you already have an inventory database, a call site, automated ordering process, financial back end, or lead tracker, consider how you can integrate it with your website. For example, if you are committed already to an IBM, Oracle, or Microsoft database for some of these operations, consider their web retailing products for faster integration within your company. Ensure the database software you plan to implement for your electronic retailing site is compatible with your central company database or data warehouse as well. A discussion with your database administrator over the compatibility benefits versus overall cost for linking your existing company databases and the planned website should take place before selecting your system. Include in your discussion the existing procedures for decision support of your existing legacy systems, such as your datamarts and customer database.

As your company moves toward electronic commerce, how will it change the nature of your products and services? Just as you monitor your marketing efforts and ad campaigns, you likewise need to plan how you will measure the efficiency of your website. Make sure you are prepared to mine your online data as you currently are mining your offline data. If you are not mining your

Amazon/AlmondNet
Exhibit 1009
Page 304

website data, or planning on doing so, you will lose both time and money, especially when retailing your products and services on the Web. In this retailing channel, marketing ads, banners, offers, incentives and other methods of selling, provide you *immediate* feedback. The type of instantaneous feedback—in a digitized format which can be data mined, which is not possible with other media such as mail, radio, or television. Finally, consider how your company plans to manage and measure the evolution of your electronic commerce strategy.

Retailing and Mining

With the rapid advent of electronic commerce, retailing transactions are rapidly migrating to the Web. As the numbers of web transactions continue to increase, electronic retailers will be faced with the same dilemma as physical retailers—to recognize the patterns of opportunity and threats to their survival. As browsers turn into shoppers on the websites, companies will realize that customers can be lost with a mouse click; the ability to decipher purchasing patterns will increasingly become a critical factor to their survival. In the daily ebb and flow of online interactions, patterns are evolving about shopper tendencies, preferences, features, and behavior. By taking advantage of both server and client side programs to monitor, identify, and track the activity of your website visitors, you can gain the ability to capture important customer information for mining.

The primary objective of a data mining analysis of your website is to identify the characteristics of those visitors and customers who make purchases. To mine your website data you will need something much more powerful than the existing server traffic analysis tools. You will require software incorporating multiple pattern-recognition technologies, such as neural networks and machine-learning algorithms. The mining of your website data should provide you with an insight into the identity of your online customers in order that you may improve your relationship with them. Through e-mining, you may discover that a relatively small percentage of your customers account for most of your sites' profits.

The key to electronic retailing revolves around providing premiere services, customized products, and convenience. For the customer, convenience is not having to tell the retailer what it is you want or need in the products and services you are purchasing. For the retailer, this means being able to figure out what the customer wants

Amazon/AlmondNet
Exhibit 1009
Page 305

before they ask. This is accomplished through to use of cookies, CGI forms, a customer database, and data mining analyses. It also involves using the preference information provided by the visitors as they register or make their purchases, and sending targeted reminders via e-mail or providing them special lists of services or products on their return to your site. Data mining is the missing link in providing the connection and relationship between consumers and the Web—a relationship that can be leveraged to serve them in a timely, efficient, and personal manner.

Although some consumers are concerned about privacy and security, many continue to register at websites and fill out surveys and sign up for one-click credit card ordering. The fact is that personal information is more secure online than offline. Consider how many people look at your credit card number when it is taken away by a waiter you have never seen before. In addition, personal preferences of products are already routinely captured every time you use your credit card—to the level that these companies know whether you use room service or not. Supermarket retailers, who encourage you to use their discount cards, already know the number of apples you consume in a week.

Customer preference information about what you like and consume is already out there and is being used to position the right products, services, and messages in front of you. Convenience drives the sales on the Web for individuals who on the average have higher-than-average incomes and education levels and tend to be very busy. On the whole, more and more consumers will come to accept that by providing some personal information, services will improve and products will be better suited to their unique needs. By providing personal preferences, consumers will find that the "right" products, services, and information will flow to them, rather than their having to go out and search for them in the ever-expanding Web.

To demonstrate how data mining can be used to improve online sales, an analysis was performed on a small sample of visitors who had made purchases of computer hardware. The objective is to discover what attributes constitute good prospects for this website. We begin the analysis by selecting our desired output, in this case "total number of sales" (Figure 8–12).

The total number of customer account records for the analysis was 1200, and the total number of errors was 66. The success rate for predicting when more than one sale was going to be made was .943. The following are some of the rules generated from the analysis:

Amazon/AlmondNet
Exhibit 1009
Page 306

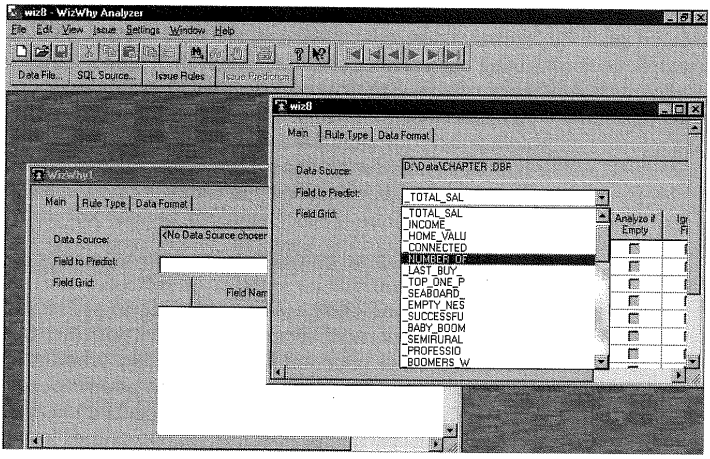


Figure 8–12
We select the field in the data we want to predict—the total number of sales.

```
If      Number of Children is 2.00
Then    Predicted # of Sales is more than 1.00
        Rule's probability: 1.000
        The rule exists in 391 records.
        Significance Level: Error probability <     0.01
```

This rule, or condition, is indicating that visitors to this website who have more than two children have a very high probability of making more than one purchase. This is a very important pattern since it occurs in nearly one-third of the sample data set: 0.32583. It may indicate the fact that a family will likely buy a PC for the kids along with a second purchase such as a printer for those school reports.

On the other hand another rule was generated, which found that customer income level coupled with the number of children indicated the propensity to make no more than one purchase:

```
If      Customer Income_ is 47,807.00 ± 205.00
and     Number of Children is 1.00
Then    Predicted # of Sales is not more than 1.00
        Rule's probability: 1.000
        The rule exists in 109 records.
        Significance Level: Error probability <     0.1
```

Still another rule found a relationship between the number of children and the search engine used to find this website: it found it limited the number of sales to no more than one:

Amazon/AlmondNet
Exhibit 1009
Page 307

```
If      Number of Children is 1.00
and     Search Engine_ is Excite
Then    Predicted # of Sales is not more than 1.00
        Rule's probability: 0.908
        The rule exists in 99 records.
        Significance Level: Error probability <     0.1
```

This rule found a relationship between the keyword used and gender and the propensity to buy more than one item:

```
If      KEYWORD_ is desktop
and     Gender is MALE
Then    Predicted # of Sales is more than 2.00
        Rule's probability: 0.930
        The rule exists in 66 records.
        Significance Level: Error probability <     0.1
```

This rule found a relationship between the visitor's age and their ZIP code, which in this case are those of a consumer class classified as Young Frequent Movers with a percentage between 71 and 92 percent:

```
If      AGE_ is 33.00
and     YOUNG_FREQUENT_MOVERS_ZIP is 71.00-92%
Then    Predicted # of Sales is more than 3.00 Rule's probability: 1.000
        The rule exists in 22 records. Significance Level:  Error
        probability <     0.1
```

The Young Frequent Movers consumers are young families with children. Typical of this population is their mobility: they tend to live in the South, with some concentrations in Nevada, New Mexico, and Wyoming. The data mining analysis found that when the visitors come from a ZIP code with these demographics, with a concentration between 71 and 92 percent and their age is 33, then their propensity to purchase multiple products is very, very good.

An analysis such as this can provide executive management and marketing and Internet teams the insight they need to know who their visitors and consumers are. It can tell them what consumer class they represent, their median age, how they found the website, what portals they use, their income level, and a bit about their family makeup. All of this intelligence can assist this retailer in how and where to advertise and where to position links to their website. It can also assist them in knowing what types of incentives they can use to increase their sales, what banners and wording to use. In the end it will help them in improving their relationships with their customers. For example, this website seems to make sales to visitors with children. This fact may indicate that marketing campaigns geared to these consumers, such as discounted kids' software, may be a good strategy.

Amazon/AlmondNet
Exhibit 1009
Page 308

Electronic Commerce

Despite modest growth, electronic commerce revenue is expected to hit $400 billion by 2002, according to a study by the research firm International Data Corporation. The figure was estimated in a study predicting global Internet and e-commerce growth from 1997 to 2002. An overall increase in Web users around the world will be the main stimulus for the robust electronic commerce revenues expected over the next few years. The IDC predicts that the global number of Web users will jump to 320 million by 2002. During this same period, the percentage of Web users making purchases online will grow to 40 percent by 2002. The same study predicts that actual number of Web users who will purchase goods online will shoot to 128 million in 2002. Another research firm, Forrester Research, Inc., places much higher estimates for e-commerce; they project $3.2 trillion by 2003! These are promising figures for electronic retailers.

Electronic commerce is changing the way all companies do business. Electronic commerce impacts all aspects of retailing: business-to-business as well as business-to-consumers. However, the core of electronic retailing means more than the buying and paying for products and services over the Web. Electronic retailing also ushers in a new method of doing online transactions and market research and data mining. Electronic commerce is pervasive, encompassing such tasks as electronic mail, peer networking, order processing and tracking, customer support, inventory management, and distribution. It impacts the monitoring of manufacturing processes, marketing, and pricing. At the core, however, is the ability to digitize, store, and analyze all aspects of marketing, retailing, delivery, and after-sales servicing.

Products and services sold and purchased over the Web can be both the traditional (physical) type as well as virtual (digital), with the latter being produced and delivered exclusively via this new retailing channel. Retailers of digital products and services have an inherent advantage over traditional ones, in that manufacturing, marketing, ordering, advertising, customer service, and all other aspects of doing business leave a digital trail of transactions which can be mined for insight, competitive intelligence, and targeted marketing. Thus the digital retailer is more responsive to the whims and preferences of his or her customers.

Consumers of physical and digital products also behave differently in this type of networked environment. They are instantly converted

Amazon/AlmondNet
Exhibit 1009
Page 309

into experts and given the powerful ability to window shop the worldwide mall that the Web represents. The efficiency of these online shoppers will over time impact product choices, pricing strategies, and competitive efforts, which will include the aggressive use of data mining strategies, tools, and techniques. Companies and their relationships with their customers will also be affected as transactions migrate to the Web, removing the spatial and temporal limitations of the traditional modes of retailing. New costs, network efficiencies, product customization, relational marketing, models of interaction, and data mining on the Web will evolve. Vertical preference-based territories will supersede the traditional markets, which in the past were bound by national or physical regions.

In an electronic retailing environment, consumers have higher control over product specification, quality, and price because of the shorter delivery cycle from design, production, and delivery. In this type of digital economy, physical products are superseded by digital ones, innovations to product design and delivery are accelerated, and there is a convergence in products, markets, and the infrastructure in which they exist. In this type of marketplace, information becomes a product commodity with value coming from knowledge-based processes such as data mining. It is a fluid and hardwired economy where marketing and retailing processes are interactive and digitized.

Consider Priceline.com, which sells airline tickets over the Web. Priceline.com lets consumers post prices they want to pay for trips, and the first airline to accept that offer gets the business. This business model has been so successful that Priceline.com has patented the process and is moving to include other retail items such as rental cars and hotels. Now consumers have control over the pricing for specific trips they want to take, challenging the retailers to meet the best possible product. Priceline.com has introduced one-to-one interaction between consumer and retailers. For the consumers it means they have control over pricing of the trip they want to purchase. For the retailer the challenge is making the sale while still making a profit. For the electronic retailers it means arriving at the right profit margin and having the ability to do some powerful analyses of sale transactions via data mining.

The same type of data analysis is required by other types of electronic retailers in which price is of paramount importance: online auctioneers. As bids and sales occur on specific retail items, patterns on prices will evolve which lend themselves to predictive modeling via data mining tools and technologies. Because every bid amount can be captured and digitized across such factors as wholesale price, how

Amazon/AlmondNet
Exhibit 1009
Page 310

OptiMark: The AI Middleman

OptiMark is a new AI electronic retailing model changing the rules in securities trading at the Pacific Stock Exchange. OptiMark lets investors buy or sell in large volumes without telegraphing their moves to other traders, which helps keep the market from moving against them.

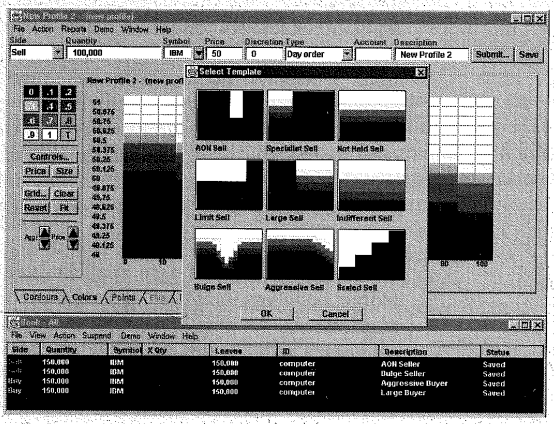
OptiMark automates buying and selling at Internet speed and efficiency. The core technology of OptiMark is that of fuzzy logic, which looks for a matching price between what the retailer wants to sell for and what the consumer wants to buy at. The OptiMark system could very well be one of the electronic retailing models of the future.

The basis of fuzzy logic, if you recall, is that of shades of memberships, with nothing being a crisp black or white. Fuzzy logic is a departure from classical two-valued sets and logic, that uses "soft" linguistic (e.g., large, hot, tall) system variables and a continuous range of truth values in the interval [0,1], rather than strict binary (True or False) decisions and assignments. Fuzzy logic is a structured, model-free estimator that approximates a function through linguistic input/output associations. Fuzzy rule-based systems apply these methods to solve many types of "real-world" problems, especially where a system is difficult to model, or where ambiguity or vagueness is common. A typical fuzzy system consists of a rule base, membership functions, and an inference procedure.

OptiMark uses fuzzy logic technology to play the role of the ultimate AI middleman. First, it allows buyers and sellers to express hypothetical trading preferences along a range of prices and volumes. Secondly, it conducts a search of all buyers' and sellers' interests at that instant. Thirdly, in less than two seconds it returns with a mutually satisfying trade at the optimal available price and volume. The OptiMark system allows for complete anonymity, nondisclosure, and the ability to express "degrees of willingness" to trade over a range of prices and volume. Using a supercomputer and fuzzy logic algorithms, OptiMark conducts an exhaustive search of all possible trades in its network, completing up to several billion calculations per stock in one or two seconds. And rather than simply matching buyers and sellers, OptiMark discovers the "sweet spot," or best possible price.

The OptiMark system is currently being used for trading in the Pacific Stock Exchange, but its developers are already looking at consumer markets on the Web for selling everything from airline and stadium tickets to insurance, energy, and diamonds—at real-time trading marts. With the Web as the conduit, consumers using the OptiMark trading system could specify what they'd be willing to pay and let the AI

Figure 8–13 *OptiMark uses a fuzzy logic engine to facilitate buying and selling on the Web*



Amazon/AlmondNet
Exhibit 1009
Page 311

OptiMark: The AI Middleman *(continued)*

middleman find the right deal. The system would seek to match consumers and retailers wherever they can be matched, creating a compromise with the best possible trade.

The OptiMark vision: all trading interests and requests are processed via an AI system instantly and continuously via the Web, making all the mathematically and logically possible matches between consumers and retailers. No consumer would ever pay more than they wanted and retailers would never undercut themselves unnecessarily.

long the bid has been open, what the starting bid was, the number of prior bids made, how many other retail items this bidder has bought, etc., all of these factors can be analyzed. Online bid websites can use such modeling tools as neural networks to model starting prices for their items. Through the strategic use of data mining technology, the electronic auctioneer can edge their profit margins on consumer items they put up on the electronic block for bid.

The Web represents a seamless marketplace combining pattern-recognition technology, telecommunications, and commerce. In this fluid marketplace, structural changes will impact retailing, transforming marketing methods by hardwiring consumer demands directly to product design, manufacturing, quality, performance, and pricing. Competitive strategies will impact entire industries and consumers will drive the changes to the marketplace. Consumers will select content and interact with producers and retailers in the design of their own products. The voluminous nature of digitized transactions in this type of networked environment leads to the prevalent use of data mining software enabling retailers to interact with customers and to create and process digitized products.

Today, money is nothing but a type of data moving from one system to another, with data mining being nothing more than a technique of trying to recognize and anticipate these evolving patterns. Electronic retailing represents a growing exchange of this type of data over the Internet—where consumers and retailers create an evolving process—much as an organism develops a nervous system. In this type of networked marketplace the processing and analysis of these digitized transactions will allow companies to know who their customers are. It will allow them to leverage human-emulating technologies for recognizing and calibrating their customers' loyalty and value.

Amazon/AlmondNet
Exhibit 1009
Page 312



9

E-Mining

Mining Your Site

E-mining is the iterative process of analyzing the patterns of your online transactions and extracting knowledge about who is buying what, when, and, most importantly, why from your website. It is about extracting previously unknown, actionable intelligence from your website interactions. It is about using pattern-recognition technologies to answer such bottom-line questions as:

- How do I optimize my website inventory and design?
- What strategic partnerships should I be developing?
- How do I increase my share of my customers?
- What cross-selling opportunities do I have?
- Who is visiting and buying at my website?
- What ads and banners should I be using?
- Who are my most profitable customers?
- How do I increase my online sales?

Websites of all sizes are currently experiencing an explosive growth in their capabilities to generate massive amounts of server data. However, few administrators are leveraging data mining technology to transform this web data into the type of competitive intelligence that can improve their e-commerce efforts. E-mining is a powerful method of "recalling" the daily contacts your company's website had with your visitors and customers and "seeing" their behavior. Of course, memory and observation are human functions which data mining tools and techniques, based on AI technologies, are designed to emulate in the processing and analysis of large data sets. This type of inductive data analysis can answer questions you hadn't thought of asking, because it can discover relationships you were not aware existed and in the process uncover hidden patterns in thousands of visitor interactions in your website.

Amazon/AlmondNet
Exhibit 1009
Page 314

E-mining can assist you in figuring out exactly what customers want and deliver just that. It can assist you in improving your profit margins, in gaining control over your back-end inventory, and in the end, in improving the design and sequencing of your web pages. Most of all it can assist you in keeping your online customer satisfied by allowing you to profile them, through the mapping of their online behavior and demographics. E-mining removes a lot of the guesswork from doing business on the Web. Through the mining of your transactional data you can become more responsive to your online customers, making it hard for them to go elsewhere. Above all, e-mining is a business process with a specific goal: *to extract a competitive insight from your online interactions.*

Beyond providing your customers the right product and service options, e-mining can also guide you in creating a strong relationship with your online clients. If your website provides a product or service that succeeds in your customers' eyes, and you are responsive to their needs again and again, then their investment in your company can create a lasting loyalty, helping you increase your market share of that client. In the following example, we will demonstrate how several data mining techniques can be used to gain this type of insight through an analysis of visitor site behavior, which in turn can provide you a composite profile of who they are.

However, as with all other data mining projects, we must first start with the basics: the data and its various components. For these analyses we start with some basic server log files coupled with some customer information culled from a registration form database. We then merged this website data with our own data warehouse and some external demographic and household database sources. Several data mining analyses will then be performed to demonstrate the different insight and output each of the technologies and tools can provide to web, marketing, and business teams involved in e-commerce. The e-mining process breaks down as follows:

Data Analysis	Technology	Goal
Association	Visualization	Discover product and consumer relations.
Clustering	Self-Organizing Map	Discover cross-selling opportunities.
Segmentation	Decision Tree	Discover key market segments.
Classification	Neural Network	Predict propensity to purchase.

Amazon/AlmondNet
Exhibit 1009
Page 315

Data Analysis	Technology	Goal
Optimization	Genetic Algorithm	Maximize overall website design.
Mapping	Geographic Information System	View purchasing patterns.
Integration		Combine clustering, segmentation, and classification to gain an overview.

Figure 9–1 illustrates the entire e-mining process.

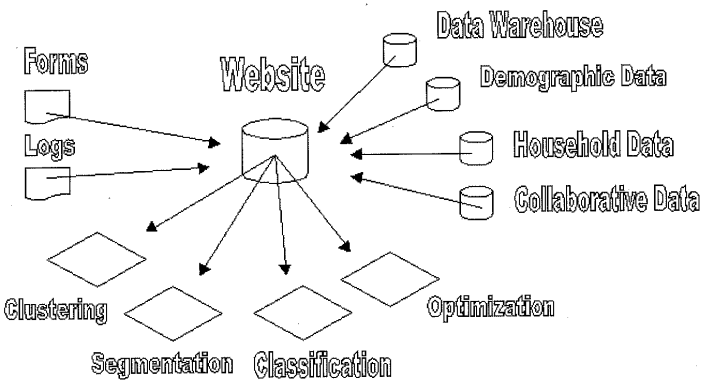


Figure 9–1
The e-mining process.

Capturing the Look

At this very moment your future customers are accessing your website. Do you know who they are, where they live, what they make, how they live, what they buy, or what they like? In the preceding chapters we have discussed a rather diverse number of topics, everything from relational marketing, to neural networks and machine-learning algorithms, to cookies and collaborative filtering. Most importantly we have tried to emphasize the glue which binds them all together: *data mining*. In this chapter we put these concepts in practice through a series of examples using several data mining techniques, technologies, and tools.

Amazon/AlmondNet
Exhibit 1009
Page 316

We begin by reviewing the process involved in capturing visitor and customer data at the server level and then using those components to append additional household and demographic information. The resulting data set is then mined using a variety of tools. The final results of these analyses are varied, although all of them are designed to provide you a new insight into who your website visitors are, where they came from, what they like, and most importantly, how you can gain and retain a relationship with them over their consumer life.

When it comes to your website, the one concept you need to be clear about, especially in a networked environment, is that it's not just a transmitter—more importantly it's a receiver. What it receives is visitor behavior information: Information about who your visitors are, what they are searching for, how and what information, services, and products they want, and which offers they most likely will respond to. Start thinking of your site as a magnifying glass, capable of discovering the features and characteristics of its many visitors and potential new online customers. To gain an insight into the behavior and features of these visitors to your website, multiple techniques can be used to perform such tasks as:

Sequencing: If they buy Product A, will they buy Product C or Product X and when?

Segmentation: What differentiates my most loyal clients from all the others?

Profiling: Who are my most profitable customers and how do I keep them?

Association: What relationship does visitor gender have to sales at my site?

Classification: How do I recognize high propensity to purchase visitors?

Clustering: What attributes describe high return visitors to my site?

Prediction: Who is likely to be my most loyal customer over time?

Optimization: How do I design my site to maximize my online sales?

Be aware that these analyses are not exclusive of each other or need to be done sequentially; your business needs will determine when you execute them. As we have discussed previously, most data mining tools are based on artificial intelligence technologies designed to emulate human perception and learning. Unlike database query programs, web analysis report generators, collaborative filtering software, or statistical packages, data mining tools are different in that they perform their analyses of patterns in an unbiased and

Amazon/AlmondNet
Exhibit 1009
Page 317

autonomous manner; they are driven by the data contained in log files and web-generated databases rather than your personal hypothesis or interpretation of web traffic behavior. Data mining can be used to surmise and simplify the intricacies of thousands of website sessions and visitor patterns in the form of a score, a ratio, a rule, or a decision tree graph.

As with more traditional sources of data mining, the key to predicting the future lies in the analysis of historical transactional data. In the retailing industry, for example, this is usually in the form of bar code data. In the financial services sector, this may be in the analysis of prior loans, credit card accounts, portfolios, CDs, or other types of financial products. In both instances the key to the construction of predictive models starts with the modeling of past transactions and interactions with customers. In the data mining of your website, these building blocks are the components of online transactions captured at the server level. These data components, coupled with a collaborative filtering engine or external demographic and household information, allow you to profile your visitors and potential future customers and to discover their preferences by analyzing their online behavior and purchasing patterns.

A Sample Website Session

1. Website visit generates an entry in the server log file.

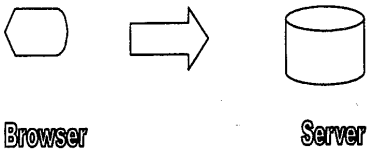


Figure 9–2a *Generated Data Components: access time, referring site, search keywords, visitor's domain, etc.*

2. Server issues a cookie to browser.

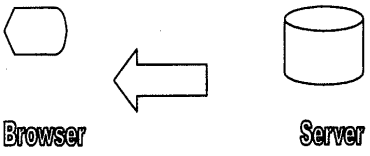


Figure 9–2b *New Data Components: visitor's unique identification name or number value.*

Amazon/AlmondNet
Exhibit 1009
Page 318

3. Website visitor completes a registration form.

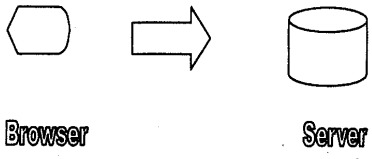


Figure 9–2c *New Components: visitor's age, gender, ZIP code, etc.*

4. Website data components are matched with other databases.

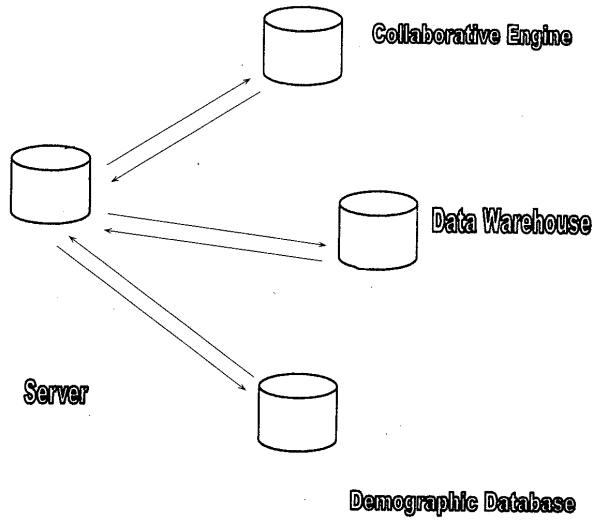


Figure 9–2d *New Components: visitor's consumer behavior, value of home and auto, number of children, etc.*

5. Finally, the enriched website data is mined.

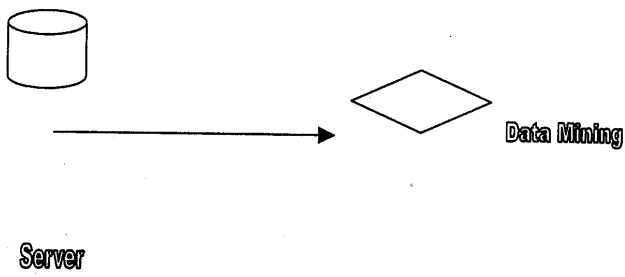


Figure 9–2e *New Component: insight about your visitors/ customers.*

Amazon/AlmondNet
Exhibit 1009
Page 319

Interactive by Design

Prior to mining your website data, you first need to have a strategy for identifying who your visitors and customers are and what kind of information you want to capture through your logs, cookies, forms, and personalized or collaborative filtering software, or combination thereof. With incremental steps you can begin to assemble a composite of who your website visitors and customers are. This commonly starts with your:

Log Files, to find out where they came from, how they found you, and what they saw.

```
ntt.com.jp - - [08/Jul/1999:16:44:14 -0744] "GET /webminer/index.html
    HTTP/1.0" 200 5898
    "http://search.yahoo.com/bin/search?p=data+mining+websites"
    "Mozilla/3.01 (Macintosh; I; PPC)"
```

Cookies, to find out where they went and who returned.

```
Session_ID=1BABV317MEN
```

Forms, to find out who they are and how to contact them.

```
Joe Blow,34,Male,94502,jblow@doc1.com
```

This then is the *start* of a customer record for data mining:

```
ntt.com.jp - - [08/Jul/1999:16:44:14 -0744] "GET /webminer/index.html
    HTTP/1.0" 200 5898
    "http://search.yahoo.com/bin/search?p=data+mining+websites"
    "Mozilla/3.01 (Macintosh; I; PPC)", Session_ID=1BABV317MEN, Joe
    Blow,34,Male,94502,jblow@doc1.com
```

This comma-delimited string can easily be imported into a database for pre-processing and prepared for merging with other data for subsequent data mining. This is the genesis of the data mining analysis, because the true value of your website data is when it's merged with other personal customer information. True value is obtained once this enriched new data set is mined and "golden ore" is excavated. One of the key strategies is to merge this "stateless" Internet data with household, demographic, and your own company databases and in doing so convert it into a "solid state."

Solid State

The web and the data it generates are often referred to as being "stateless" due to the architectural design of TCP/IP. At times this makes the

Amazon/AlmondNet
Exhibit 1009
Page 320

task of mining this type of server data quite difficult. However, certain data captured at the server level can be linked to lifestyle and consumer information in order to convert this data into a *solid state*. Through the strategic use of site-specific data merged with external data, you will be able to "profile" who your customers are. For example, the preparation of the data may involve the merging of your registration form database with a demographic data set using the ZIP code as the joining link. Or, it may involve merging your log files with a collaborative or personalization service provider using shared cookies, pooled over a network of websites. In both instances, a new and much richer data set will be created, which may be mined in order to discover specific features and behavior about your website visitors.

A common methodology for extracting features from customer information files and data warehouses, especially for highly competitive market sectors such as retailing, financial services, insurance, and telecommunications, involves the enrichment of internal customer data with external demographic and psychographic information. These types of commercial databases are usually based on census information, which has been enhanced by database marketing resellers. These data vendors usually sell information culled from such public record sources as:

- real estate deed recordings
- government publications
- motor vehicle records
- public record files
- tax assessor files
- warranty cards
- questionnaires
- birth records
- white pages
- direct mail
- the census

Detailed demographic information from these diverse sources not only helps you to determine the types of products and services your visitors are most receptive to purchasing, it can also assist you in choosing the types of ads, banners, links, and messages you present to them. E-commerce is highly driven by branding and the method by which a product is marketed. The lifestyle and values of your visitors can influence the overall design of your website and the method by which you market your product and services. Your visitors' demo-

Amazon/AlmondNet
Exhibit 1009
Page 321

graphics may also lead you to forge alliances or partnering with other firms and sites. In fact the level by which your website data can be enhanced by merging it with other information sources varies in relation to the activity which occurs with each visitor or customer. You can only link information to data, which is generated by the interactions that take place in your website:

Visitor Action	Data Captured	Data Can Be Linked To
Browse	Referrer log, cookie	Collaborative filtering, ad network
Register	ZIP, gender, age	Demographic database
Purchase	Customer address	Household file, data warehouse

With every interaction, website visitors provide more information about themselves that can be used to merge it with other sources of lifestyle and consumer preference databases. In addition, with every purchase a retailer can determine a consumer's preferences and rate of consumption. This data gathering starts at a very low level, where only server log information is being captured and where the only matching which can occur is that involving the matching with collaborative filters and the commingling of cookies, such as that of the DoubleClick DART ad network. At the next stage, when a visitor completes a registration form at your website, important information can be captured, such as their ZIP code, which can be used to match it against such demographic databases as those from the Claritas, CACI, and Acxiom systems.

Lastly, when a visitor becomes a customer and actually makes a purchase requiring that a product be mailed to them, their physical address is captured, which can be matched against an assortment of other sources of information. These sources include household information providers like Experian and Polk, which cull consumer information from warranty cards and other public records and consumer sources. All of these third party data providers vary in the amount of detail they can provide about consumers starting from the ZIP code, ZIP+4 and down to the track or single household level. It is recommended that you do some comparison shopping and testing to determine which vendors give you the highest file penetration for your particular customers. Ask for specific statistics on their information content and their percent of file coverage.

Website customer information can also be linked to your own data warehouse enabling you to do some additional data mining analysis,

Amazon/AlmondNet
Exhibit 1009
Page 322

such as a clustering or segmentation analysis. One of the immediate benefits of linking your website and data warehouse is that you can match, compare, and mine the data in order to compare the features of your existing customers with potential new clients. You may uncover some similarities in their regional location, age, gender, or other demographic, income, or lifestyle features. An important side benefit is that you may find a new market segment or a completely new type of customer you did not have before.

The Data Set

As you interact with your visitors and customers you gradually capture more and more information about them. You begin by first finding out how they found your site and what keywords they used to find you, which you capture from your log files. Next, as they complete a form you find out what ZIP code they live in and other important information, such as their preferences and age or gender. Finally, should they make a purchase, you then have their physical address. All of these elements that you have captured can be used to match them against a collaborative engine or demographic and household information files in order to gain an additional understanding of their lifestyle and their preferences. More importantly the combined data set can be mined in order to segment and construct profiles about who they are. This composite can be constructed as a comma-delimited string, which can easily be imported into a series of data mining tools.

The Data Set Breakdown

A single customer record with server data and appended household and demographic values:

```
ntt.com.jp - - [08/Jul/1999:16:44:14 -
    0744] "GET /webminer/index.html
    HTTP/1.0" 200 5898 "http://search.
    yahoo.com/bin/search?p=data+mining+web
    sites" "Mozilla/3.01 (Macintosh; I;
    PPC)", Session_ID=1BABV317MEN, Joe
    Blow,34,Male,94502,jblow@doc1.com,
    4,$81,895, "Luxury",$153,408, $278,
    1, 1998-08-29, 3, 5, 10, 70,0,2,0,0,
    0,0,0,0,0,0,0,0,0,0,0,0,0,0,etc.
```

This is the information breakdown of this single customer record sample:

This contains the domain:	ntt.com.jp
The data and time:	[08/Jul/1999:16:44:14 -0744]
The requested page:	"GET /webminer/ index.html
The version number:	HTTP/1.0
The status results:	200
The number of bytes sent:	5898"
The referring search engine website:	http://search. yahoo.com
The search keywords:	/bin/search?p =data+mining +websites
The browser used:	"Mozilla/3.01 (Macintosh; I; PPC)"

Amazon/AlmondNet
Exhibit 1009
Page 323

The cookie ID:	Session_ID =1BABV317MEN
The visitor name:	Joe Blow
The visitor's age:	34
The visitor's gender:	Male
The visitor's ZIP code:	94502
The visitor's e-mail address:	jblow@doc1.com
Number of prior sales:	4
The visitor's estimated income:	$81,895
The visitor's type of vehicle:	Luxury
The visitor's estimated home value:	$153,408
The total amount of online sales:	$278
The estimated number of children for this household:	1
The last date a purchase was made by this visitor:	1998-08-29
The demographic breakdown by percentage of consumer classes for this visitor's ZIP code:	3,5,10,60,2,20,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,etc.

These last fields are the percentages of consumer classes as provided by a demographic database and can be interpreted as: 3 percent of Top One Percent, 5 percent Seaboard Suburbs, 10 percent Empty Nesters, 60 percent Successful Suburbanites, 2 percent Prosperous Baby Boomers, 20 percent Semirural Lifestyle, etc. There are 40 possible consumer classes, each of which is represented as a percentage value.

For purposes of these analyses we used this single data set containing components from log files and a forms database, appended with customer information from a company data warehouse, and household and demographic information:

Attribute	Source	Data
Product	Form Database	"Item_03"
Domain	Log File	"com."
Search Engine	Log File	"yahoo"
Number of Sales	Data Warehouse	4
ZIP	Form Database	94502
Income	Household Database	$81,895
Vehicle Value	Household Database	"Luxury"
Home Value	Household Database	$153,408
Total Sales	Data Warehouse	$278
Number of Children	Household Database	1
Gender	Form Database	M
Age Group	Form Database	45–49
Last Buy Date	Data Warehouse	1998-08-29
Top One Percent ZIP Code	Demographic Database	3
Seaboard Suburbs ZIP Code	Demographic Database	5
Empty Nesters ZIP Code	Demographic Database	10
+37 other ZIP code percentage breakdowns		

Amazon/AlmondNet
Exhibit 1009
Page 324

Clustering (Association and Sequencing)

In order to discover possible associations and sequencing patterns in the data, the first analysis involves using a data mining tool incorporating a Self-Organizing Map (SOM), also known as a Kohonen neural network. This type of neural network is ideal for exploratory analyses and can be used to search for discrete clusters in the data. A k-nearest neighbor statistical algorithm can also be used to accomplish the same type of task. The purpose for doing this type of clustering analysis on your website data is to discover associations between visitor attributes, such as gender or age and the number of sales they make or the total amount of purchases they have made at your site.

Clustering is a good e-mining start—it is an exploratory method of discovering relationships, which you may find surprising and totally unexpected. An SOM network explores all the data that you have compiled and gradually constructs spatial clusters of data. This type of analysis is commonly done by retailers in "market basket analysis," where they search for patterns or relationships between the sale of certain products for cross-selling and mixed promotion opportunities.

This type of clustering analysis is also known as "unsupervised learning" because the neural network used to do the clustering is not instructed or provided a desired output. A SOM is instead designed to discover—on its own—specific clusters or classes within a given data set. For purposes of this example we will use an SOM to explore and discover clusters within a 10,000-sample data set. Hopefully it will find for us some cross-selling patterns and marketing opportunities. Using a data mining tool with a graphical programming interface we are able to import the comma-separated value (CSV) flat text file and partition a random sample of 50 percent of the records. We connect a table viewer to visually inspect the data and ensure the expected values are correct (Figure 9–3). Next, we connect a "type" node, which ensures the data types are correctly set prior to importing the data into the Kohonen (SOM) modeling node (Figure 9–4). The data mining tool will generate a graphical view from which the user can gradually begin to see the partitioning of the data into clusters identified by different shades of color and density, as shown in Figure 9–5.

After some time (this can vary from a few minutes to a few hours depending on your CPU, the size of the data set, and the complexity of the settings of the network), the SOM will construct some distinct data cubes, or clusters. The parameters for determining the distinctions between these clusters can be set or allowed to default to the software setting. Next, as shown in Figure 9–6, the results of the clus-

Amazon/AlmondNet
Exhibit 1009
Page 325

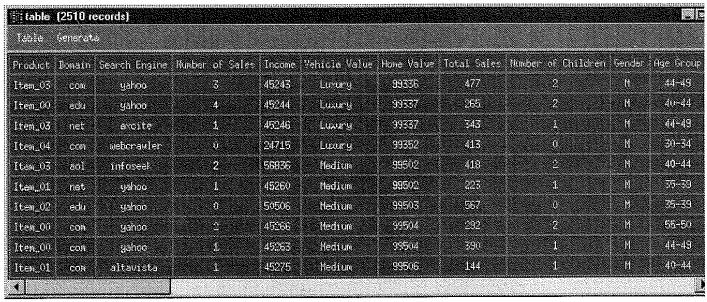


Figure 9–3
It is a good practice to visually inspect your data once imported into a data mining tool.

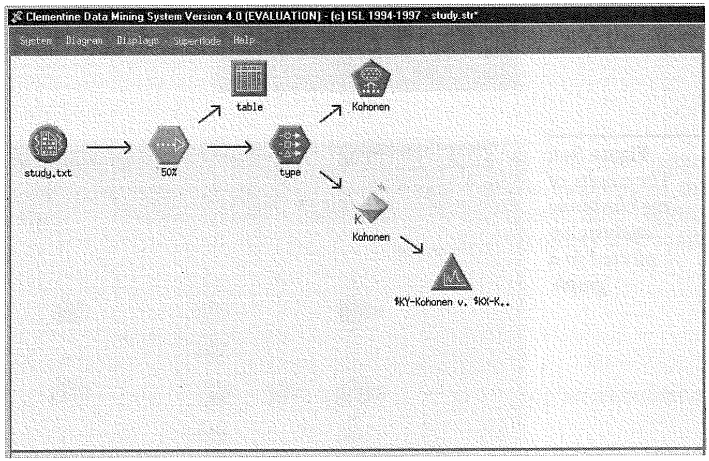


Figure 9–4
A random 50-percent sample of the data is passed through an SOM network to create clusters.

tering analysis can be plotted and the significant distributions of data can be "marked" for further "drilling" by simply marking them on the fly and generating a sub-segment from them for further analysis, using a rule-generating algorithm.

The marked clusters can now be generated as a derived class of the entire database, which represents the five clusters marked in Figure 9–6. This subset of the data is now passed through the rule-generating C5.0 machine-learning algorithm, which is used to extract a set of rules, or conditional statements, that describe the clusters discovered by the Kohonen neural network. This two-step process involves first generating an autonomous analysis of the data (clustering). Next the clusters are diagrammed and areas of interest are identified and

Amazon/AlmondNet
Exhibit 1009
Page 326

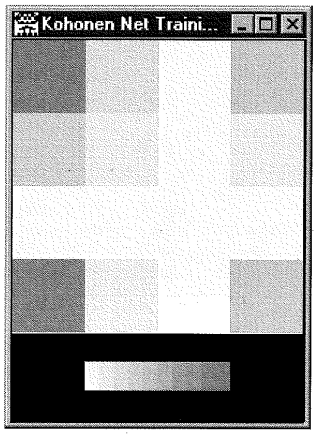


Figure 9–5 *Gradually clusters begin to develop from the website data.*

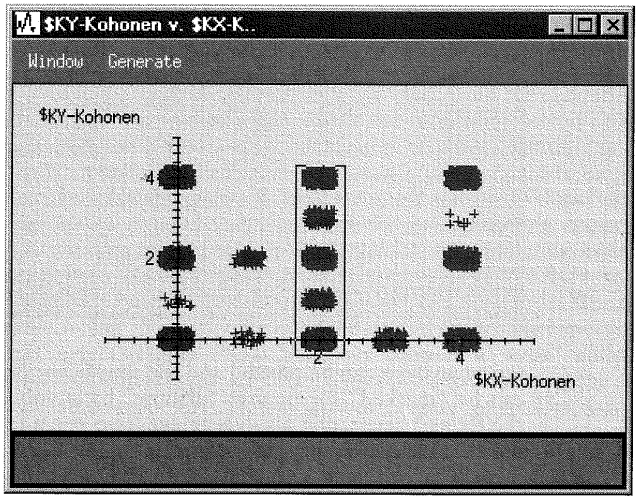


Figure 9–6 *The results of the clustering analysis are plotted to a graph.*

partitioned. These partitioned clusters are then evaluated using an algorithm, which describes the data further via the process of descriptive rules. These are rules that are easily understood by online retailers, webmasters, marketing, and business professionals. Figure 9–7 illustrates the steps taken to partition the data into clusters and link the output to a rule-generating algorithm.

Amazon/AlmondNet
Exhibit 1009
Page 327

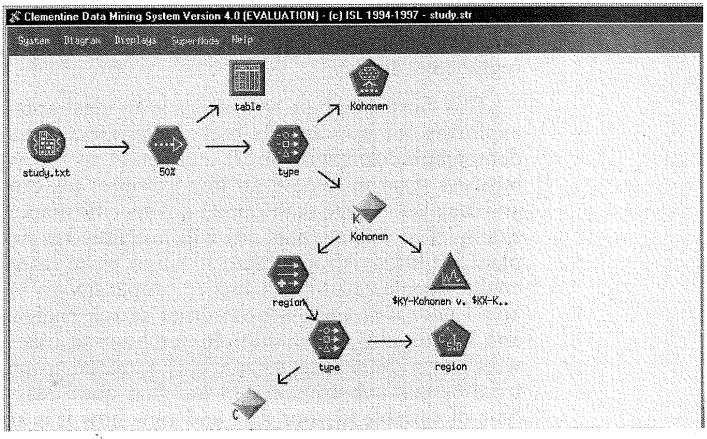


Figure 9–7 *Rules can be generated to describe a region in a cluster.*

The following is a set of rules generated directly from the clusters created by the Kohonen SOM neural network. The mapping of the clusters represents the high concentration of accounts around region1 (the five-cluster area), which represents website customers who made two purchases:

```
Rule 1.
IF Search Engine excite
AND Domain [aol edu] -> THEN region1 (2 Sale Items)
```

Rule 1 relied solely on log file information to develop an association.

```
Rule 2.
IF Search Engine infoseek
AND Vehicle Value High -> THEN region1 (2 Sale Items)
```

Rule 2 uses components from the log file (the referrer section) and information culled from a household database to find an association with the number of sales.

```
Rule 3.
IF Search Engine webcrawler
AND Vehicle Value Luxury
AND Age Group 30-34 -> THEN region1 (2 Sale Items)

Rule 4.
IF Domain com
AND Vehicle Value Luxury
AND Age Group 40-44 -> THEN region1 (2 Sale Items)
```


Amazon/AlmondNet
Exhibit 1009
Page 328

Rules 3 and 4 found associations between log file and household information, and with the visitors' age, which was obtained from the registration form.

This clustering type of analysis is an ideal approach to take when exploring for associations between products and transactional and demographic information. It can reveal to the web marketing and business team undiscovered opportunities for cross- and up-selling, and may lead to the development of specific marketing strategies. For example, in this analysis ads will probably be most cost effective if placed at Infoseek rather than at Yahoo, since visitors from this search engine seemed to have a higher propensity to buy at this website. You'll note that visitors from other search engines, like Webcrawler and Excite, also made multiple purchases at this website, but it was with other conditions, such as a certain age group, vehicle value, or from certain sub-domains. In fact it is quite easy to zero in on any type of variable in your data and view how it is affected by the clus-

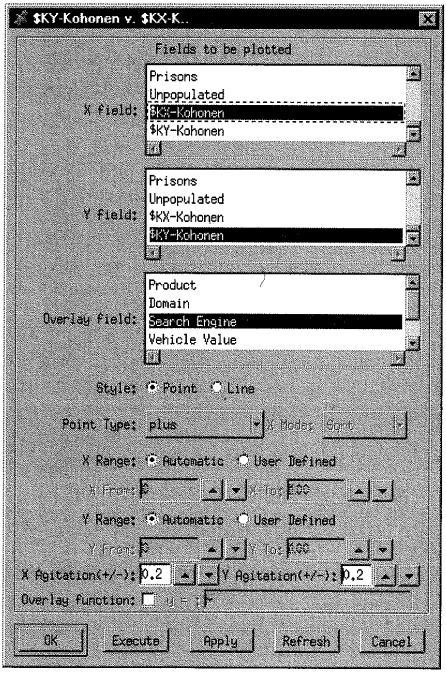


Figure 9–8
We select Search Engine as the variable to plot and view.

Amazon/AlmondNet
Exhibit 1009
Page 329

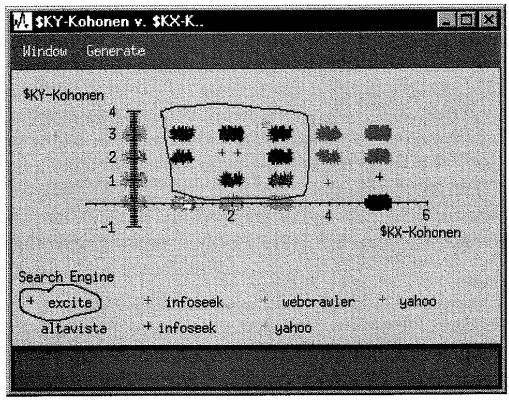


Figure 9–9
The clustering of website visitors based on the search engine they used.

ters uncovered in your data mining analysis. The fields to be plotted can be selected on the fly, as shown in Figures 9–8 and 9–9.

Note in Figures 9–8 and 9–9 that the predominant search engines for this website seem to be Excite and Infoseek. This clustering analysis concentrated on identifying the features of those visitors to this website who made two purchases and what search engine they used. However, similar analyses can be performed to identify other types of associations, such as those customers who have spent more than $50 or $100 in purchases.

Another clustering analysis was performed, this time with the intent of identifying associations between various product lines being sold at this website. For this example, there are five distinct categories of product types: Item_00, Item_01, Item_02, Item_03, Item_04, and Item_05. The graph in Figure 9–10 shows a high concentration on the right, which can be marked again, in order to create a derived data subset, which will be generated as product_region1.

As in the prior analysis, we next derive a subset of the data directly from the clustering graph by simply marking the area we are interested in exploring further, in this case, the clusters on the right. We next connect the derived data set to the rule-generating C5.0 algorithm, which will extract the following rules describing this particular grouping:

```
Domain com
Vehicle Value Luxury
    Product Item_00 (248.0, 1.0) -> product_region1
    Product Item_02
```


Amazon/AlmondNet
Exhibit 1009
Page 330

```
Working Class =< 0.5 (54.0, 0.796) -> product_region1

Product Item_04
Age Group 44-49 (10.0, 1.0) -> product_region1

Product Item_05
Age Group [21-24 60+](0.0, 1.0) ->product_region1
Age Group 25-29 (3.0, 0.667) -> product_region1
Age Group 35-39 (6.0, 0.833) -> product_region1
Age Group 40-44 (11.0, 1.0) -> product_region1
Age Group 44-49 (17.0, 1.0) -> product_region1
Age Group 30-34
Number of Sales > 1.5(3.0, 1.0) -> product_region1
Age Group 50-54
Number of Children > 1.5
Income =< 40518(2.0, 1.0) -> product_region1
Product Item_01
Number of Sales > 1.5
West Coast Immigrants =< 0.5 (6.0, 1.0) ->
product_region1
```

The most important rule generated from this clustering analysis is the first one at the very top:

```
IF
Visitor's domain is .com
AND Their Vehicle Value is Luxury
THEN Product Item_00 (this occurred with 248 customers with a probability
    of 100%)
```

All of the other rules for the other product items had relatively small number of occurrences. For example, the next highest rule is for product Item_02 with 54 observations and a probability of 0.79

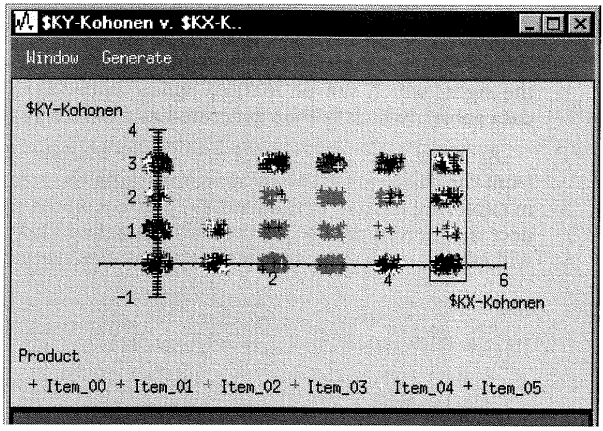


Figure 9–10 *Each sale item is represented by a different color in the graph.*

Amazon/AlmondNet
Exhibit 1009
Page 331