Chapter 9

Sequence Analysis and GIS

Chapter Outline

What You Will Learn in This Chapter

- How are patterns over time or sequences of data or actions analyzed?
- What types of business problems are suited for sequence analysis?
- What are clusters of sequences?
- How is data organized for sequence analysis?
- How is sequence clustering used to analyze Web site traffic?
- What features are provided by other sequence mining tools?

Houston Pawn Shops

A geographic information system (GIS) examines data related to location. Spatial Insights is a GIS consulting company and it used GIS tools to evaluate the market for check cashing and pawn shops in Houston, TX. The analysis begins with data on the locations of existing pawn shops and check cashing companies. It then uses Census demographic data to identify the areas with the highest potential demand for the services. Factors include population, median household income, percentage of renter occupied dwelling units, average household size, and population density. To combine the various factors, each was divided into quintiles (20 percent ranges), and each range was given a value from 1 through 5. The individual ratings were summed to provide a single score for each Census block. This index value can then be plotted on a color-coded trend map to show the areas with the highest potential demand. The supply side was mapped using the locations of existing stores by color coding each block based on the total square footage of existing retail stores. The supply and demand side values can be combined by reverse coding the supply and adding the two numbers. A color-coded heat map of the result highlights the areas that have the highest potential demand with the lowest number of existing stores. [Spatial Insights 2008]

Many business decisions and data are related to location. Government demographic data is useful in evaluating many of these problems, and several tools exist to display data geographically.

Spatial Insights, Inc., *Market Potential GIS Case Study*, 2008. [The PDF file contains the charts.] http://www.spatialinsights.com/company/papers/downloads/MarketPotentialStudies.pdf

Introduction

What tools exist for special purpose analyses for sequences of data and geographic problems? Several specialized problems exist in business and other disciplines and interesting tools have been developed to provide useful information in these situations. The problems and tools are a little different from other forms of analysis but they can be useful in specific situations.

This chapter examines two specific types of problems: (1) Sequence data, and (2) Geographic or location-based data. The two topics are not related to each other, but both topics are relatively straightforward and can be examined in just part of a chapter. Sequence data has an interesting history, with most of the tools derived from genetic research into DNA and protein sequences. However, some interesting business problems are based on sequences—notably tracking user interactions on Web sites.

Many business problems are related to location—almost anything related to customers, suppliers, and competitors. Geographic analysis goes beyond simple mapping—it examines how data is correlated through location. For example, how are customer income and sales related through location—how much difference does a wealthier neighborhood make to sales? Many of the geographical tools are visual—emphasizing the ability to display the data on maps to make it easier to see the correlations.

The sequence data and tools are considered first in this chapter because the analysis is relatively complex. Microsoft SQL Server Analysis supports some types of sequence analysis so the examples focus on its capabilities. The geographical analysis follows in separate sections. The challenge to the geographical analysis is to get access to a tool. Microsoft MapPoint is used to demonstrate the basic mapping capabilities, but other tools exist including Google for Web-based charts and ESRI ArcInfo on the high-end.

Sequence Analysis

How are patterns over time or sequences of data or actions analyzed? Several prominent examples of sequences raised important questions and led to the development of tools for analyzing the specific types of data involved. Two of the classic examples are: (1) DNA (and protein) sequences, and (2) Evaluating Web site logs to identify usage patterns. In the context of these tools, a **sequence** is a set of data or actions in a specific order. More importantly, the tools are designed to find patterns or groups of similar sequences. For instance, are there groups of customers who follow a common path when interacting with a Web site? Or, in the biology world, are some patterns of DNA (genes) correlated with specific outcomes or diseases?

The DNA example is interesting and many powerful tools have been designed specifically to attack the many biology questions involved. The results and background are not particularly useful in a general business context, but some business problems echo the same characteristics, so a couple of basic concepts are useful. DNA, the biological building block of life, consists of sequences of four molecules. The four nucleobases are cytosine, guanine, adenine, and thymine; commonly abbreviated as CGAT. Consequently, portions of DNA sequences are written using the four letters, such as: CCGATCGGTA, but the sequences contain millions or even billions of these characters. Researchers often need to compare multiple DNA sequences; for example, comparing sequences from two different

people. Or, if a group of people have a specific disease (such as a rare cancer), finding similarities among their DNA sequences that might have caused a susceptibility to the disease. A key element of analyzing DNA sequences is the need to begin at a specific point within the overall sequence, such as when comparing an individual gene. Many of the DNA tools are highly optimized for these specific tasks and are not directly applicable to business problems. However, some of the base concepts have been generalized into tools that can be used for other problems. Further, the DNA notation is convenient to illustrate concepts because each item consists of a single letter, so sequences are easy to write down.

The most common business example of sequences is analyzing patterns in Web site usage. A Web site consists of a set of pages and user browsing consists of moving through the pages in some relatively random order. Marketers and Web site developers are interested in learning if certain types of people follow similar paths through the Web site. For example, do people who end up making a purchase follow some path that leads them to make a purchase; or do those who do not make a purchase somehow miss a critical page? Stop for a minute and see how this problem is similar to the DNA situation. Each person has (or follows) a sequence of items and researchers want to find similarities in those sequences for groups of people.

Sequence mining tools follow two general topics: (1) **classification** and (2) **clustering**. Technically, a third version of tools combines both topics into one method. Classification is used to help identify a sequence. It can be based on the entire sequence but more likely uses a subset. For example, a researcher might want to know if a specific subset exists within the sequence; such as searching for a specific gene pattern in DNA or identifying whether a Web site visitor worked through a set of checkout pages. Detailed classification can involve identifying or specifying the starting location of a subset. Beyond simple pattern matching, sequences raise difficult questions about how close a pattern must be to constitute a match. In particular, gaps are an important topic. For example, does the sequence CAGTC contain the subset CC? The answer depends on the length of the **gap**, or spacing between items, that the researcher is willing to consider. With a zero-length gap (or no gap), the answer is "no" the sequence does not contain the CC subset. But, if a gap of at least 4 units is allowed, then the sequence does match. Longer patterns also create issues with partial matches. If a pattern contains 50 items and a match is found for all but one entry, should it count as a success, or does the researcher need an exact match?

Most of the business sequence mining tools use a version of clustering which is conceptually similar to the basic cluster tools. Simple clustering tools compare items or people based on attribute measures. They use a distance measure to find items that are close to each other and distant from items that belong in a different cluster. In terms of sequences, the tools search for patterns that are similar across a group of observations. With most business mining tools it is also possible to include static dimensions if they are available—such as customer income or location.

Clustering requires a distance measure function and much of the work in developing tools relies on developing better distance measures. Distance is useful as a measure of closeness—instead of relying on exact matches. For example, the **Levenshtein distance** or **edit distance** is an interesting concept for many sequences. Given two sequences, S_1 and S_2, the edit distance is the minimum number of edit operations needed to convert S_1 into S_2. It is easiest to see in terms of

simple strings or words. For example, converting the sequence of letters "dollar" to "holler" requires two steps: change "d" to "h" and "a" to "e" so the edit distance is 2. But converting "dollar" to "pound" requires at least 5 steps. The edit distance is not always the best measure of similarity. More sophisticated measures assign different weights to change, delete, and insert moves; and others are better at comparing from different positions or alignment. Common algorithms include BLAST, BLOSUM, PAM, FASTA, and Smith-Waterman. With most data mining systems, the choice of the distance function is made by the developers and you will rarely be able to change it. However, it can be one of the main differences between systems, so different tools can give different results. If you have specific data, you might need to test various tools to find one that uses a distance function best suited to the data.

Business Situation

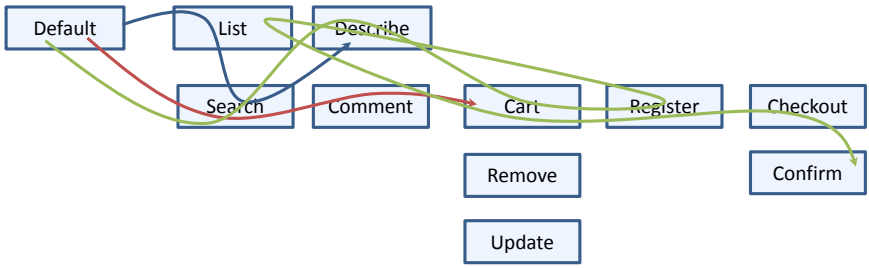
What types of business problems are suited for sequence analysis? Perhaps the most important question in this chapter is to know when sequence analysis can be used. Many concepts in business might appear to be sequence related, but they do not work well for common sequence analysis tools. In particular, much business data consists of time series, such as sales for each month. But sequence analysis on this data will not be very useful. (Instead, time series analysis is more useful.)

At a minimum, most sequence analysis tools require two things: (1) discrete items or events in order, and (2) individual tracks or random paths over that data. The second condition is used for clustering—finding groups that follow similar paths or sequences. Most business-based tools rely on clustering analysis. The alternative is classification analysis which requires defined patterns and outcomes; but these tools are more commonly found in biology applications.

The use of discrete data is relatively important. The data can be numeric or text, but continuous data rarely works. Even discrete items with too many options can cause problems. If tens of thousands of items can appear in a sequence, then it is unlikely that any two sequences will match. In some cases, you can **discretize** the data to create categories or **bins** that reduce the overall number of items. For example, a problem might create sequences by measuring the time between various events. Technically, time is a continuous variable, so it should be converted by defining categories based on intervals (short, medium, long, and so on).

The second condition for clustering is more restrictive—multiple random paths through the data. If the paths are not random, they are not very interesting. For example, a manufacturing company likely has a sequence for producing items, but the base sequence is usually fixed not random—all products go through the same steps in the same order. With a large, complex manufacturing system randomness might be introduced through other variables. For example, multiple machines or workers might exist at each step, so the sequence becomes a question of tracking which machine or person did the work, not the actual processing step. But even then, regression analysis might be more useful than sequence analysis.

Because of the importance of randomness in the sequence, most business use of sequence analysis is based on customer actions. Customers tend to generate random actions; but obtaining the data requires some method of tracking all customer actions. Web sites are one of the few places for accurately tracking customer actions. Two common practices for sequence modeling are (1) analyzing Web page sequences in browsing and purchasing, and (2) evaluating the sequence in which



Possible results/clusters:
Explorers: Default, List, Search, Describe
Interested: Default, Search, Comment, Cart
Purchasers: Default, Search, Describe, Cart, Register, List, Cart, Checkout, Confirm

Figure 9.1

Example of random paths through discrete pages on a Web site. An individual user might browse through the pages in any order. But certain types of users could follow similar paths—creating clusters. The three sample types of users might follow similar paths. A goal of sequence analysis is to identify clusters of similar sequences—such as identifying the three hypothetical categories of users shown here.

products are added to a shopping cart. It might be possible to identify similar actions in bricks-and-mortar stores, but the data collection is difficult and intrusive. (For example, a few grocery stores have experimented with hand-held scanners that shoppers carry through the store to record their own purchases as they load their carts.)

Sequence analysis could be applied to other business problems, just remember to focus on the actions and the desire to compare multiple random sequences. Focus on the potential results: sequence analysis identifies clusters of similar sequences. The clusters can include the sequence data and they can often include other, fixed data. For example, customer income, location, age, or other dimensions can be included in the search for clusters.

Model

What are clusters of sequences? Identifying similar sets of users or customers is a common sequence mining problem in business. Figure 9.1 shows the basic concepts for a Web site project. With a Web site, a sequence consists of a set of pages that are browsed by an individual user—including the order in which they are viewed. As will be shown in the examples, Web servers automatically track this data in a log file—at least identified by user IP address. At a minimum, the data consists of files that show the user IP address, a time stamp, and the name of the page that was delivered to the user. The objective is to identify groups of users or clusters that follow similar paths. From a business perspective, these clusters should indicate different types of customers. To identify clusters, the system searches for multiple people who follow similar sequences. In the example, some customers simply browse the site for a couple of products, a second cluster searches for products and reads reviews, while a third cluster finds a product and makes a purchase.

$$P(S) = P(S_m \mid S_{m-1})\, P(S_{m-1} \mid S_{m-2}) \ldots P(S_2 \mid S_1)\, P(S_1)$$

Figure 9.2

Markov chain probabilities. From probability theory, the probability of any state can be computed as the chained probability of each previous item in the sequence. Starting from the right side is the probability of the starting point; which is multiplied by the probability of the second item given (|) the first item. Most Markov chains are restricted to looking at a specified number of entries, called the order (k).

By identifying the various clusters, it might be possible to understand the customers and perhaps find ways to convert browsers into purchasers. Obviously, any group making a purchase will generate a sequence of pages using the shopping cart that will not exist with non-purchasers. If the only difference between these groups is the shopping cart pages, then the information available will be minimal. From a marketing perspective, the resulting clusters will be more useful if they reveal a difference in paths for the clusters before the purchase steps. Then it might be possible to compare the initial sequences of purchasers and non-purchasers to see which specific page or pages make a difference. For example, perhaps people who make a purchase often read reviews by other customers, while those who do not make a purchase skip (or never find) the reviews. In that case, it would be useful to make the review comments more prominent—such as including them directly on the product page without requiring a second step.

A relatively common method for analyzing sequence data is to estimate the values for a Markov chain. A **Markov chain** is a concept from probability and statistics—where any next state (or sequence item in this case) depends on the current state. The probability of moving from one state to another one is given by the **transition probability**. In the context of the Web site example, a person begins at the default page and works through several pages to get to a product description page. At any point, including this page, there is a probability that the person will move to the shopping cart (purchase the product), along with probabilities of moving to almost any other page. Figure 9.2 shows the basic mathematical concept of a Markov chain using conditional probabilities, which demonstrates how a sequence depends on all of the prior states or events in the sequence.

The transition probabilities estimated in a Markov chain are more useful from a business perspective. As shown in Figure 9.3, picture a user at a specific page in the Web site. The transition probabilities indicate the chance of the user moving to any of the next available pages. A key step in sequence analysis is to estimate these transition probabilities. These values are going to be different within each sequence cluster. People with similar paths will have similar (average) transition probabilities for moving to other items. Comparing these probabilities provides information on the differences across clusters which can help explain how various groups use the Web site differently.

Classification

Classification of sequences or outcomes is a more complex task. Essentially, it requires matching sequences or patterns. As noted in the Model section, matching requires considering gaps and alignment issues. The technical steps are not important in business, but the researcher needs to determine whether gaps and alignment are critical issues for the specific problem.

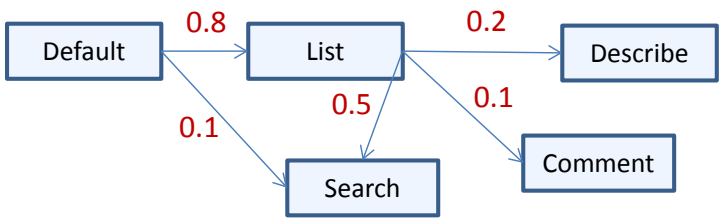


Figure 9.3

Transition probabilities. Think about a user viewing a specific page such as the Product List page. There is some probability of the user moving to any other available page, such as 0.20 to get to a product Description, 0.50 to do a search, and 0.10 to read the product comments. A key step in sequence clustering is to estimate these transition probabilities.

Sequence classification problems in business are similar to other prediction tools. In most situations, the business has a set of outcomes and wants to determine which sequences are more likely to lead to a specified outcome. For example, a bank has sequences of loan payments for many borrowers. These values could be recoded as full payment, partial payment, no payment, or late payment; or perhaps in number of days late. The outcome of interest would be whether the borrower defaults on the loan. The sequence data mining tool would then look for patterns that are likely to lead to borrower defaults. If valid results are produced, the sequence patterns identified could be used to forecast problems with future borrowers.

One complication with sequence classification is that many of the techniques and tools were designed for DNA analysis and might need to be reconfigured to work with business data. For example, sequence classification is not supported in SQL Server Analysis Services. But, over time, more tools are likely to be made available.

Clustering: Business examples

By far, the most common business example for sequence analysis is evaluating Web site logs. The two main tasks are (1) finding clusters of browsing sequences for different classes of users, and (2) evaluating the sequence in which customers add items to a shopping cart. The challenge with business examples is finding the appropriate data sets. The clustering tools need a set of discrete actions or events and a set of random tracks through those events. The requirement for multiple, random events rules out many potential business examples. For instance, most production steps are fixed and not random.

In many organizations, customer behavior is a key random element, so most business applications involve the analysis of customer actions. But collecting detailed data on customers can be a challenge—particularly in terms of maintaining privacy or at least anonymity. Web sites are one of the easiest ways to track detailed customer interactions. With SQL Server Analysis Services, it is also possible to include other dimensions on customers that are not part of the sequence data. For instance, firms might have income, age, gender, or customer category data. These elements can be used as part of the clustering computations to help differentiate sequences. For instance, browsing might be different for males versus females.

Sample DNA Data for individuals:

1. ACAGGTACTGGA…
2. CCAGGAATACG…
3. ACATTACTGAAG…

1. M, 32, …
2. F, 45, …
3. M, 62, …

Each string represents a sequence for a single individual.
The order is contained within the data string (left-to-right).
Additional data might exist for each individual (gender, age, …)

Figure 9.4

Organizing data—DNA example. The DNA example is easy to visualize, partly because only four values exist for each item. The data consists of two components: (1) The sequence of items (shown as a string), and (2) Data about the individuals.

To add more customer dimensions, additional tracking data will be needed. A Web server does not exactly identify a user (or customer). Instead, it tracks usage by IP address. In most cases, an IP address is unique to a person for at least a short period of time. Over time, IP addresses are often recycled and used by other people; they cannot be used to identify individual customers. To track individuals, it is necessary to create accounts and have customers log in, then the associated ID value will have to be recorded along with the various events. From that point, it becomes possible to attach the other dimensions from the database.

It is possible to create other business applications for sequence data mining. However, the challenge is to find random data for the paths. It is also important to use discrete events or observations. For example, a manufacturing firm might be interested in tracking production sequences to see if some of them lead to higher error rates. In general, most production steps are fixed, but if the production events include different interactions by humans, then randomness will be added. However, in the end, the analysis will be evaluating the human (or possibly machine) randomness. The key is to understand exactly what is being measured and evaluated in the clusters.

Some trickier business examples can use timing as the random element. For example, perhaps customers go through only four or five steps; but those steps might be stretched out over time. That is, the events are fixed, but the time between them is random. For instance, an engineering firm might go through the same basic steps with each client, such as: define project, detail design, start construction, make changes, complete project. Because the steps are similar for each project, they do not provide a useful sequence for analysis. However, the time between the steps is more likely to be random so that data would provide more interesting results for sequence analysis. The time would have to be discretized to convert to a finite and smaller number of possible sequences.

Data

How is data organized for sequence analysis? The main issues with the type of data have already been described: Discrete events or items where the order is important. Analysis also requires a collection of observations of these sequences that have a random element—essentially random tracks through the events or items. If the tracks are fixed instead of random there is nothing to gain from analyzing them because the goal of the analysis is to find patterns or clusters of similar tracks.

Collecting and organizing the data for sequential analysis is important because the tools are picky about how the data must be structured. Also, understanding the data requirements makes it easier to think about business applications that can benefit from sequence analysis. In some ways it is easier to think about the data by starting with the DNA tasks—partly because it is easier to write down and visualize the data. Figure 9.4 provides a simple example. A sequence for one individual is written as a string of letters where the order is defined by the position within the string from left-to-right. Additional data might exist for each individual and could be stored in a second table, and include things such as gender and age.

Attributes and Observations

Business data is likely to be more complex than the DNA examples—each item usually consists of considerably more than four possible values, although the sequences will rarely be as long as a DNA chain. Business sequence data is also often stored in different formats—such as different rows in a table. For instance, in the common problem of analyzing a Web log, each page browse is stored as a single row in a table. Figure 9.5 shows some sample Web log data as it would be stored in a database table. Note that the Web logs are initially sorted by date and time. But, most importantly, each event is stored as a separate row within the table; and the rows from multiple users are intermixed. Without a customer login step, the only way to identify a sequence is to rely on the UserIP address. Within a period of time, a given user generally maintains a single IP address. The sample data in the table comes from a real Web log and shows three different IP addresses, so it contains the start of three sequences.

Figure 9.5

Sample Web log data. Some of the available data for each Web log event. Note that each event is stored as a separate row. Without a customer login step, the only way to create a sequence is the UserIP address. But the sequences are intermixed.

ID	Date	Status	Time	UserIP	Method	Bytes	URIStem	Bytes	Referer
1	01:28.0	200	250	207.46.199.39	GET	29247	/DBMS/...	344	NULL
2	03:11.0	200	46	173.192.34.91	GET	0	/	162	NULL
3	08:11.0	200	46	173.192.34.91	GET	0	/	162	NULL
4	11:16.0	200	140	62.212.73.211	GET	318	/robots.txt	327	NULL
5	11:17.0	200	140	62.212.73.211	GET	715	/petstore/...	364	NULL

PersonID	Seq	Item
1	1	A
1	2	C
2	1	C
1	3	A
2	2	C
3	1	A

1. ACAGGTACTGGA…
2. CCAGGAATACGG…
3. ACATTACTGAAG…

Figure 9.6

Organizing data—DNA example as a table. The table requires two keys: Person and the sequence order. Note that the data might be intermixed. Some tools will require the data to be sorted correctly before beginning the analysis.

For comparison, how would the DNA data be stored in a similar table? Figure 9.6 provides a partial answer, but the actually table would be considerably larger. Note that the table requires two key columns: (1) The identifier for the individual, and (2) a sequence order. In database terms, the data can be intermixed within the table. Most tools will require the data to be sorted correctly (PersonID, Seq) before beginning the analysis. The point of the example is that business data is typically stored in the data table format so business tools are designed to handle data in that format. However, it is more difficult to visualize the sequences in a table.

Additional dimensions on individuals would be stored in a separate table. It would have a single key column (PersonID) and columns for the dimensional data (age, gender, and so on). Figure 9.7 shows a sample table for the DNA data example. This table needs only one column in the key (PersonID) which is why it needs to be a separate table from the sequence data that used two columns (PersonID, and SequenceOrder). The two tables are linked by the PersonID column.

In the case of pure Web log data, the additional data regarding the individuals does not exist. However, many sequence analysis tools still require a base table

Figure 9.7

Data table for individuals. Only one column is needed for this key (PersonID) so it needs to be a separate table from the sequence data.

PersonID	Gender	Age	…
1	M	32	
2	F	45	
3	M	62	

UserIP
62.212.73.211
173.192.34.191
207.46.199.39

```
INSERT INTO IPTable
SELECT DISTINCT UserIP FROM Weblog;
```

Figure 9.8

Data table created by a query on the Web log data to obtain distinct individual key values (IP address).

for individuals. Essentially, this table needs to list each person one time. In this case, it is necessary to create the table by using a query on the Web log table:

```
SELECT DISTINCT UserIP FROM Weblog;
```

Note the use of the "DISTINCT" keyword to eliminate duplicate entries. This query essentially creates a table with a single column (UserIP) that contains each IP address a single time. Figure 9.8 shows the sample data. Ideally, the analysis tool will allow the use of the simple SQL query. However, if it requires a separate table, Figure 9.8 shows the addition of the INSERT INTO phrase that takes the selected data rows and inserts them into a table that was created with a single column.

Continuous and Discrete Data

Sequence mining requires discrete data. In business terms, the sequences are usually events, such as browsing a Web page or selecting a product. The data is already discrete, and the researcher simply has to code the data so that the items will be recognizable in the evaluation of the results.

It is possible for business problems to use continuous data—such as using time intervals as a sequence. In these situations, the data will need to be converted to discrete bins; such as short, medium, and long time frames. A SQL Server CASE function can be used to define the categories. In more complex cases, a new table can be created with the category definitions and inequality-joins can be used to assign the category values to the data.

Missing Data

Obviously, key columns cannot contain missing data. For the most part, sequences do not contain missing data—particularly in business exercises. For example, a clickstream generated from browsing a Web site simply moves from one page to the next. There is no requirement on the number of pages to be visited; so sequences can be varying lengths. For other problems that require a defined number of sequence items or that rely on position, missing data is handled by defining it as a new state. In the DNA example items would be allowed to take on five values: C, G, A, T, and missing; instead of just the four base values. However, no real rules exist for handling the missing entries. Instead, it is up to the specific tool,

and in some cases control over parameters to determine how to handle the missing values. For instance, they might be treated as gaps—that match any value. Or, they might be treated as a separate item. The difference matters most when using classification tools instead of clustering tools. For example, does the partial DNA string: CCA_T match the string: CCAGT or CCATT or neither one? The answer has to depend on the specific problem and the researcher, so it should be controllable through parameters in the tools.

Web Log Data Files

Note: This section is only necessary if you want to read your own Web log files. It is not needed to load the sample databases for the book.

Web logs are a common business problem analyzed with sequence mining tools. However, they have some special properties that need to be considered before they can be analyzed. The most important issue is that some standards exist for storing Web log files. In particular, the Web log files are almost always stored as plain text files. Each row represents one event, such as a page retrieval, and the detailed data are stored in a specific order on the row. The data is generally spread across multiple text files—where one file commonly holds data for a single day. The data mining tools, particularly SQL Server, require extracting the data from the text files and storing it in tables in a relational database. Fortunately, Web logs are used for several purposes so some tools exist to help with the conversion. Different tools exist for different servers, so additional research might be necessary to find a tool for a specific server.

Microsoft supports a parser tool that reads, extracts, and aggregates data from common Microsoft log files. The current version of the tool is "Log Parser 2.2" which has been around since 2005. It is available as a free download from Microsoft—just search for it by name. The tool is relatively powerful and can read common Microsoft files including Web logs, event logs, registry entries, and XML and CSV files. However, it runs as a command line tool without a graphical-user interface. This feature actually makes it more useful because it is easier to automate, but it can make it harder to learn the various options. More recently, Microsoft has added the "Log Parser Studio" which runs as a graphical interface on top of Log Parser. Only the Log Parser 2.2 will be used for the examples in this chapter.

The data files provided for this book have already been converted into database format so it is not necessary to deal with the original log files to follow the examples in the book. However, to perform a similar task on other log files, it will be necessary to convert them, so it is worth understanding the basic steps. Figure 9.9 shows the Log Parser command used to create a CSV file from text files generated by Microsoft IIS. The command needs to be entered on a single row but is split up to make it easier to read. The basic format is similar to a simple SQL SELECT command. The goal is to select the desired columns, but the data can be pulled from multiple log files and combined into a single output file. It is possible to add simple WHERE conditions, but they are not needed in this example. The example contains several special functions to clean up the data. First, the OUT_ROW_NUMBER() function is used to generate a unique ID value for each row. Second, the log files treat date and time as separate columns, but they will work better in the database as a single column. So the TO_STRING and ADD functions are used to combine the two columns into one. Finally, three columns can contain long character strings: User-Agent (browser), cs-uri-stem (page request), and cs(Referer). These values need to be limited to a set number of characters (512

```
"c:\Program Files (x86)\Log Parser 2.2\LogParser"
"SELECT OUT_ROW_NUMBER() AS ID,                    Generated ID #
ADD(TO_STRING(date,'MM/dd/yyyy'), TO_STRING(time,' hh:mm:ss')) AS dt,
sc-status,                                  Combine date and time into one.
time-taken,
REPLACE_CHR(SUBSTR(cs(User-Agent),0,512), ',',';') AS User-Agent,
c-ip, cs-method, sc-bytes, cs-version,
REPLACE_CHR(SUBSTR(cs-uri-stem,0,512), ',',';') AS CS_URI_Stem,
cs-bytes,
REPLACE_CHR(SUBSTR(cs(Referer),0,512), ',',';') AS CS_Referrer
INTO temp.csv
FROM *.log
ORDER BY dt“
-i:IISW3C -o:CSV        Specify Web log format as input and CSV as output
```

Limit length and change commas to semicolons

Figure 9.9

Log Parser command to read IIS Web logs and convert to a CSV file for loading into a database table. Only common dimensions are selected. In the log, date and time are separate columns that need to be combined into a single column for the database table. Also, three of the strings are limited in size (512) and examined to convert commas into semicolons to prevent loading problems. The command must be entered as a single-row.

in this case) because the corresponding table in the database will be given a finite amount of storage. Additionally, the data will be loaded from an intermediate **comma-separated-values (CSV)** file and the data columns cannot contain extra commas. So the REPLACE function is used to convert commas to semicolons. In some cases, it might be more useful to simply truncate the data at the point a comma is encountered, which relies on the use of the INDEX_OF function.

As a side question: How do you find the names of the columns (such as cs-uri-stem)? Usually, these names are header lines in the log files. However, it is often easier to issue a test statement of the form: SELECT * FROM *.log and then break the command after a couple of rows have been displayed; which will display the column names as a heading.

The remaining elements of the Log Parser command specify the input files (*.log), the output file (INTO temp.csv), and the sort order (which is optional). The –i and –o options specify the format of the input and output files. The choices can be found by typing the Log Parser command without any parameters. Using the command-line version of Log Parser, you open a new command window (Start: cmd), navigate to the folder holding the copies of the log file (cd …) , and type the full command. With many large log files, it might take several minutes or more for the command to run and generate the CSV file.

Once the CSV file has been created, it can be examined inside Excel—although large files will be truncated. More importantly, it can be imported into SQL Server using the Bulk Load command. First create a table with the matching columns. Then issue the EXECUTE (N'Bulk INSERT …) command. Figure 9.10 shows an example of the CREATE TABLE and EXECUTE commands that match the data from the Log Parser command. The columns in the table need to exactly match

```
DECLARE @data_path nvarchar(256);
SELECT @data_path = '<PATH>';
--
CREATE TABLE WebLogs
(
  ID int identity(1,1) NOT NULL,
  WebDateTime datetime NULL,
  SCStatus int NULL,
  TimeTaken int NULL,
  UserAgent nvarchar(512) NULL,
  UserIP nvarchar(250) NULL,
  CSMethod nvarchar(250) NULL,
  Bytes bigint NULL,
  CSVersion nvarchar(250) NULL,
  URIStem nvarchar(512) NULL,
  CSBytes bigint NULL,
  Referer nvarchar(512) NULL,
    CONSTRAINT pk_WebLogs PRIMARY KEY (ID)
)
;
EXECUTE (N'BULK INSERT Weblogs FROM '' + @data_path + N'temp.csv''
WITH (
    CHECK_CONSTRAINTS,
    CODEPAGE=''ACP'',
    DATAFILETYPE = ''char'',
    FIELDTERMINATOR= '','',
    ROWTERMINATOR = ''\n'',
    KEEPIDENTITY,
    TABLOCK
);');
```

Figure 9.10

SQL Server commands to create a table to hold the log data and the BULK INSERT command to load the table from the CSV file created from the text log files. Change the <PATH> entry to match the folder holding the CSV file.

the columns in the CSV file. Change the <PATH> command to be the name of the folder holding the CSV file, such as C:\Temp\ or wherever the file is stored. The BULK INSERT command then loads all of the CSV data into the database table. A similar process is used to transfer data files for this textbook because it is relatively fast and works with all recent versions of SQL Server.

SQL Server Sequence Clustering

How is sequence clustering used to analyze Web site traffic? Microsoft SQL Server Analysis includes a sequence clustering tool that is useful for business analyses (as opposed to DNA). It can be used to analyze Web server logs to identify common patterns in usage. Remember that clustering works by finding groups of sequences that are similar to each other. SQL Server also supports the use of non-sequence dimensions in the clusters—behaving similar to traditional clustering. However, this data is usually only available if the Web site requires people to log in and then collects personal data on the users. The example used in this section does not include any data that can be traced to individuals.

To illustrate the high variability in real data, this section uses data collected from a live Web server. The Web server is used by the author to handle educational books and class content. The IP data was randomized so that it is not possible to track addresses back to individual users (even if it were possible to obtain network-address translation logs). Despite the relatively small number of total users (19,600 unique IP addresses), the log files generate a large amount of data—several hundred megabytes for five months of usage. This section shows the basic process of configuring the sequence analysis and briefly explains how to evaluate the various result screens.

Goals

The primary goal of using sequence clustering on Web logs is to identify the main groups of users. Most organizations and Web sites develop traffic from various groups and it helps to understand the differences and similarities in how these groups use the Web site. For example, in a business site, some users will simply be browsers—perhaps searching for products or comparing prices. Hopefully, one group will be purchasers—people who buy products online. Possibly one group will be investors—looking for company background information. It is not necessary to know about these groups ahead of time—the clustering tool will automatically identify different groups and show how they behave differently in working through Web pages.

It will help to understand the results if the researcher knows the Web site—and the various pages. It will also be useful to have some initial concepts of the people who use the Web site. The clustering tool does not need this information, but the results can be difficult to follow if the page names are meaningless and the researcher has no background in the industry.

Data

As explained in the prior section, the data is pulled from Web server logs. Web servers automatically record pages, date/time, and user IP address along with other tidbits such as the size of the pages (in bytes). Converting these text log files into a database table can be a problem—but the Log Parser simplifies and automates the task. Be sure to add a key column (ID) to the data and ensure that it is sorted by date, time, and IP address. Date/time by itself is not going to work as a key because Web servers are multitasking and can deliver multiple pages at essentially the same time. Also, the IP address cannot be used as a key in this sequence table.

Web analysis today has an additional catch. Large organizations often use relatively automated systems to manage their Web servers. These content-management systems consist of shell pages that retrieve data from a database to be displayed on a single page. For instance, a user might search and request data on a specific product. This data will always be displayed on a page that might be called "products." The server log only records the "product" page—it does not record the data displayed on that page. Some systems deliver most content through these systems—including images and text. Likewise, as companies add more programming code to Web pages, it becomes more difficult for simple text log files to track the total user interaction. The log files will provide a broad overview of user actions, but perhaps not the detail wanted by everyone. Detailed Web site tracking requires the installation of tracking code directly on the pages. For example, Google Analytics provides several tools to track and display detailed usage on Web pages using this approach. However, these tools do not usually perform sequence clustering so both methods are often useful.

Tools

Begin by loading the sample data into a SQL Server database (WebLogs). Because the data does not contain other information about customers, only one table is created. Then create a new project in Visual Studio using the Analysis projects. Add a connection to the WebLogs database and create a Data Source View that uses the WebLogs table.

SQL Server sequence clustering always requires two tables: (1) a Case table that contains a key value for each sequence—such as a CustomerID; and (2) the actual sequence item data in a "Nested" table. But, the sample data loads only the single (nested) table where each row contains a sequence item. The solution is to create a "Named Query" in the data source view that holds a list of unique UserIP addresses. The IP address will be used as an identifier for the each user. It is not perfect—IP addresses get reassigned over time, but it is the only identifier available.

In the data source view, create a named query and call it UserIPView. Add the WebLogs table and select the UserIP column. Modify the SQL so that it includes the DISTINCT keyword:

```
SELECT DISTINCT UserIP FROM WebLogs;
```

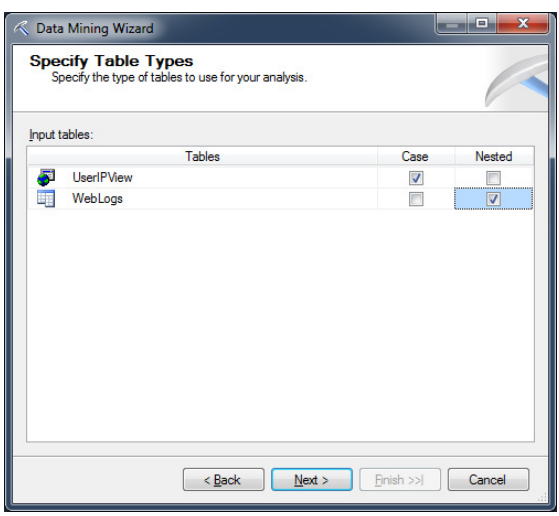
Because the WebLogs table contains the actual sequence data, the UserIP repeats. This new view extracts just the unique values of the IP address and treats the data as a separate table. After the view is created, you will have to right-click the new UserIP column and assign it as the logical primary key. Finally, drag the UserIP column from the WebLogs table and drop it on the UserIP column in the new UserIPView query to establish a relationship.

Configuring Microsoft Sequence Clustering

With the data defined in both a Case and Nested table, the sequence cluster can be created as a new Mining Structure. Figure 9.11 shows one of the first steps in

Figure 9.11

Case and nested table selection in SQL Server sequence clustering.



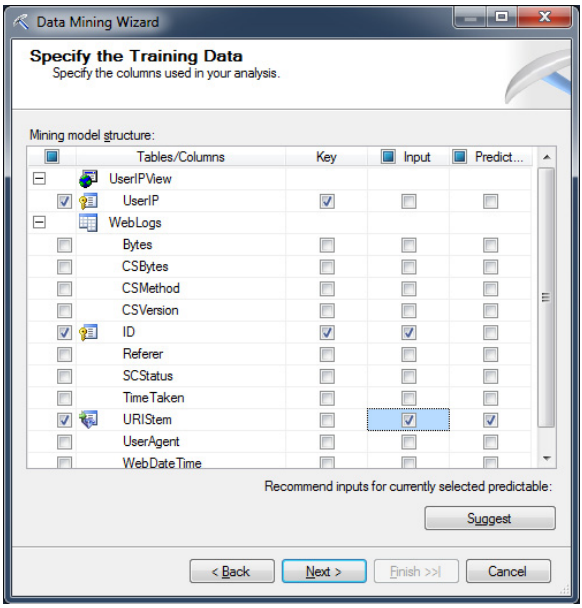


Figure 9.12

Selecting key columns correctly. The UserIP is the only possible key for the UserIPView Case table. The generated ID column is the key (and input) for the Nested WebLogs table. The selection of the sequence items is critical. In this case it is the URIStem column which contains the name of each page that was browsed. The URIStem column is slected as Input and Predicted—never as key.

the process—setting the Case and Nested tables. Remember that the Nested table contains each sequence item on a new row. The Case table holds unique values for the user that generated the sequence. The Case table usually contains individual or customer data. This table or view can hold other non-sequence data such as age and gender if it is available.

One of the most important and trickiest steps is to correctly select the columns for the analysis. Figure 9.12 shows the choice of the columns for analyzing the Web logs. The UserIP column in the UserIPView table is the only possible key. In other situations, a CustomerID might be the key column. The key for the WebLogs data is the ID column which contains generated values. This column must be carefully created—it specifies the sort order for the sequence. Recall that it was created based on date and time. It would be tempting to use the WebDateTime column instead but it will not work because it includes duplicate data—the Web server delivered some pages at the exact same date and time because of multitasking. Also, note that the UserIP in the WebLogs table is not available because it was already selected in the Case table; none of the other columns track the sequence. The next step is to select the URIStem column as the sequence item data. This column holds the name of the page that was requested so it represents the selected value—much like the letters CGAT in the DNA problem. This column is specified as the input and predicted column—but can never be chosen as the key. If additional data on customers is available in the Case table, these columns can also be selected as input and predicted columns in this step.

Parameter	Default	Range
CLUSTER_COUNT	10	[0,…)
MAXIMUM_SEQUENCE_STATES	64	0, [2,65535]
MAXIMUM_STATES	10	0, [2,65535]
MINIMUM_SUPPORT	10	[0,…)

Figure 9.13

Control parameters. These parameters control the processing for the sequence clustering tool. CLUSTER_COUNT is treated as a hint not a fixed value. A value of 0 requests use of a heuristic to choose the number of clusters. The MAXIMUM_SEQUENCE_STATES parameter restricts the number of item values observed to those that are most common.

Those are the main steps for configuring the sequence clustering: (1) Choose the Case and Nested tables, (2) Specify the key columns in those two tables, and (3) Select the column in the Nested table that indicates the item value for the sequence. Processing the model is handled the same way as with other tools. Right-click the new data mining model and choose the option to process it. The processing could take a while to run—the sample data includes several million rows.

Processing the Sequence Cluster Model

In Microsoft Visual Studio, processing is initiated the same as with other tools: Right-click the model name in the explorer window and choose the option to Process the model. However, sequence clustering requires a hefty amount of processing power and time. The sample data on a fast machine with a high-speed drive is reasonably fast and takes only a couple of minutes at most. However, remember that the data is pulled from a relatively small server and consists of only five months of Web logs.

Despite the relatively small size of the server pages, when processed, the sequence cluster tool complains about the number of items (pages) and it generates a warning message that: *Cardinality reduction has been applied*. Look at the parameters for the sequence cluster model (Mining Models tab, right-click the Microsoft_Sequence_Cluster entry). Figure 9.13 shows the four parameters available to control the processing. The CLUSTER_COUNT variable defaults to 10 but this number is treated as a hint not an absolute constraint. Setting the value to zero requests the use of a heuristic to determine the best number of clusters to use. The parameter MAXIMUM_SEQUENCE_STATES is more important in many situations. It limits the number of values to use for each item state. In the Web browsing example, the default value of 64 limits the number of pages to examine to 64. The system tries to choose the most active pages, so it discards pages that have few hits. This value can be increased to bring in more pages. However, Microsoft notes that values larger than 100 "can result in a meaningless model." It is this parameter value that is triggering the warning message about cardinality.

For a first pass model, letting the system choose the most active pages probably makes sense. Researchers first want to examine the most commonly-used pages. However, it is possible that the lesser-used pages are important. For instance, per-

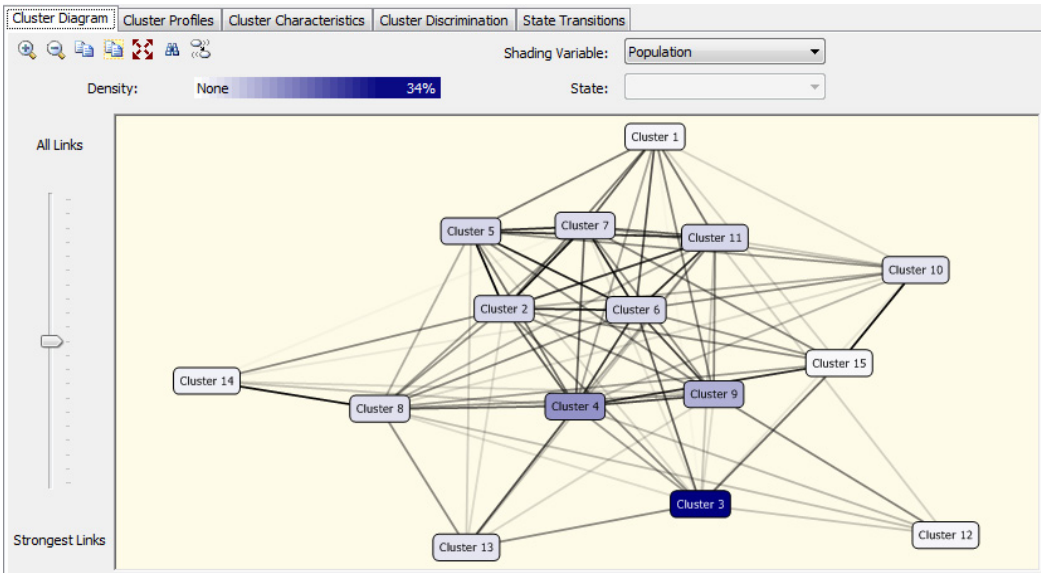


Figure 9.14

Cluster diagram. This diagram highlights the number of observations in each cluster—darker tints represent more observations. The link lines indicate how closely the clusters are related—heavier lines indicate closer relationships.

haps higher-end customers use specific short-cut pages. The solution to examining these other pages is to create a query that selects just the pages you want to investigate. For complex Web sites it can be helpful to build models that cover separate sections of the site.

Results

The results viewer in Microsoft Analysis Services is similar to that for standard clustering. Several graphical views are available along with a few tabular presentations. It is also possible to write queries to extract data directly from the Analysis database. However, the results for sequence clusters are a little more challenging to read than traditional clusters because the sequences can be relatively long. In one sense, sequence clustering needs to display data similar to traditional clustering—such as the choice of the Web pages displayed. But the results also need a way to show the order in which those pages were visited.

Clusters

Figure 9.14 shows an example of the cluster diagram which provides an overview of the results. The main purpose is to show the number of clusters where the number of observations within each cluster are highlighted by the darkness of the tint. In the example, Cluster 3 has the most number of similar responses. The thickness of the link lines also shows how closely the clusters are related. Heavier lines indicate more overlap between clusters. SQL Server uses the expectation maximization algorithm which computes a probability that each item (sequence) belongs to a cluster so an identified sequence might be associated with multiple clusters.

The cluster diagram has an additional feature that is useful for understanding clusters in terms of Web pages. By default, the shading is set for the population

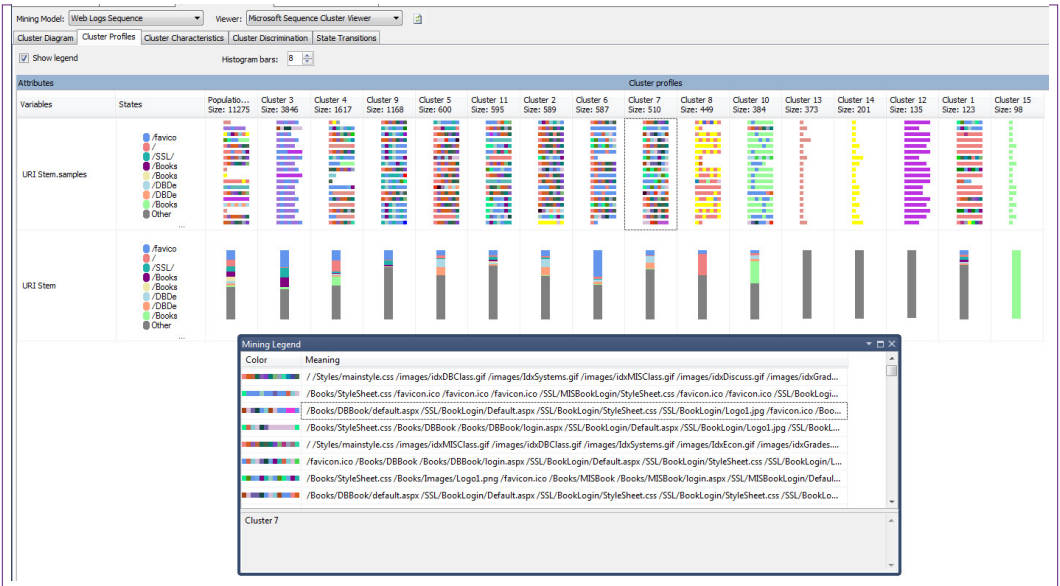


Figure 9.15

Cluster profile The profile highlights differences across clusters with a color-coded display. The .samples row shows the actual sequences (order) and the bottom row shows just the list of pages. Use the Mining Legend box to see the actual names of the pages for a selected cluster.

or number of sequences within the cluster, making it easier to spot the clusters with the most traffic. However, a drop-down box can be used to change the Shading Variable to the item values (URI Stem in this example). In this mode, a second drop-down box enables selection of individual pages (up to the limit of the 64 pages allowed). This trick then highlights the clusters that most heavily use a specific page. For example, choose the /robots.txt page and see which sequences most heavily use that page. The robots.txt file is a specific file aimed at search engines and it contains rules about what the robots should focus on and which pages they should ignore (voluntarily). Any sequences that heavily use that file are visits from search engine robots.

The three main result views focus on the individual clusters: Cluster Profiles, Cluster Characteristics, and Cluster Discrimination. The Profiles tab displays color-coded indicators for the sequences (URI Stem.samples) and the pages (URI Stem). The display shows multiple clusters on the screen to emphasize the differences across the clusters. On the other hand, the Characteristics tab focuses on a single cluster at a time—by showing the detailed list of items (pages).

Figure 9.15 shows the cluster profile for the sample data with the results highlighted for Cluster 7. Notice that the display contains two primary rows: (1) Top row labeled with *.sample* shows the sequences or page ordering, and (2) Bottom row shows just the list of pages. If additional customer dimensions were available and used they would be included in the bottom row analysis. Selecting a cluster in either the top or bottom row brings up the details in the Mining Legend box, which makes it easier to see the actual page names. Sequences are still difficult to see because they are usually too long to fit into the display box. It might be easier to read the results if the original data was recoded with shorter names.

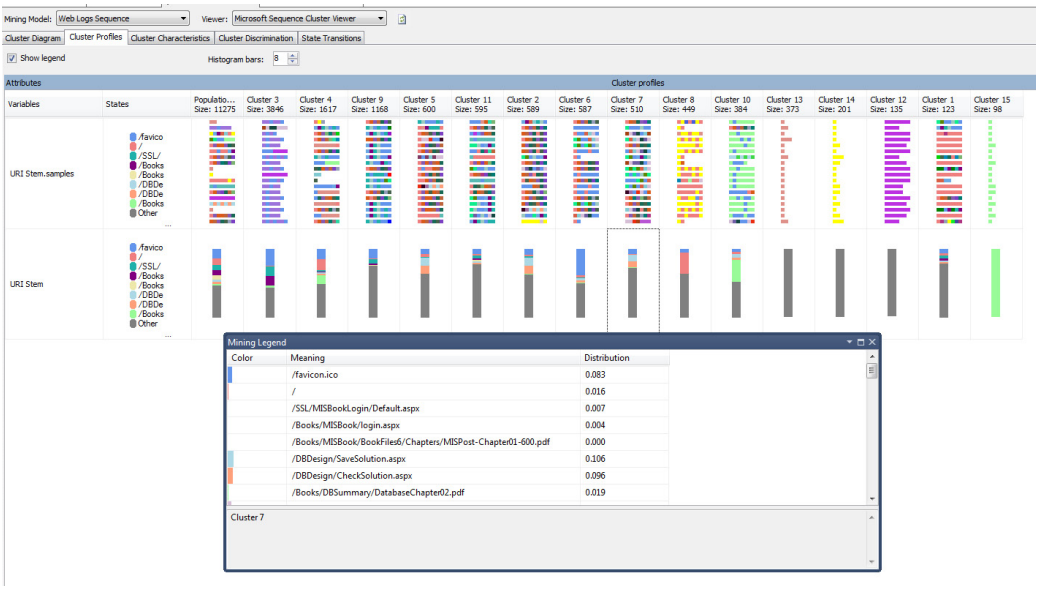


Figure 9.16

Cluster profile pages The Mining Legend for the pages contains a percentage value to indicate which pages were most common. For Cluster 7 the main pages are the Save and Grade pages for the DBDesign tool.

As shown in Figure 9.16, the profile display for the pages contains some additional information—the percentage of hits for each page. For Cluster 7, the main pages are the Save and Grade pages from the DBDesign tool on the Web site. So this particular sequence represents usage of that tool. As these results are discovered, the clusters can be renamed to indicate the primary sequence usage.

Figure 9.17 shows more detail for the pages within a cluster. Notice the emphasis on the database book login, database design, and chapters 1 – 3 of the database textbook. These pages again indicate that Cluster 7 represents a group of students studying database design. Identifying the users of the clusters is a useful first step in understanding the results. However, it requires knowledge of the Web site and the various types of users.

Sometimes it is difficult to understand a cluster—particularly when clusters include seemingly unrelated pages. In these situations, it is possible that the model contains too many clusters; or that user sequences simply do not group together. The Cluster Discrimination view helps understand a cluster by displaying pages that are important to the cluster versus pages that are important to other clusters. As shown in Figure 9.18, the researcher can choose to compare two clusters side-by-side. However, the default is to compare the selected cluster (7 in this case) to its complement—things that are not in the chosen cluster. The drop-down box makes it easy to select a second cluster for situations where two clusters seem to contain similar data sequences.

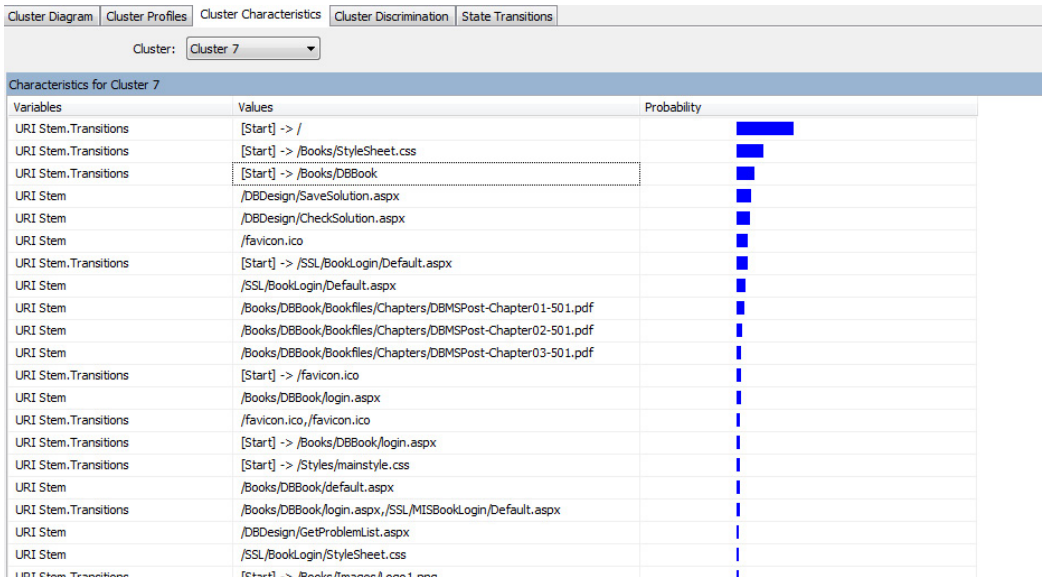


Figure 9.17

Cluster characteristics. The main pages for Cluster 7 are the database book login, DBDesign, and chapters 1-3 for the database book. So Cluster 7 represents students working on database design.

Figure 9.18

Cluster discrimination. Discrimination compares clusters side-by-side. The default is to compare a cluster to its complement—things not in that cluster, but the drop-down box makes it easy to pick a second cluster. Here, Cluster 7 emphasizes database concepts instead of the general MIS book.

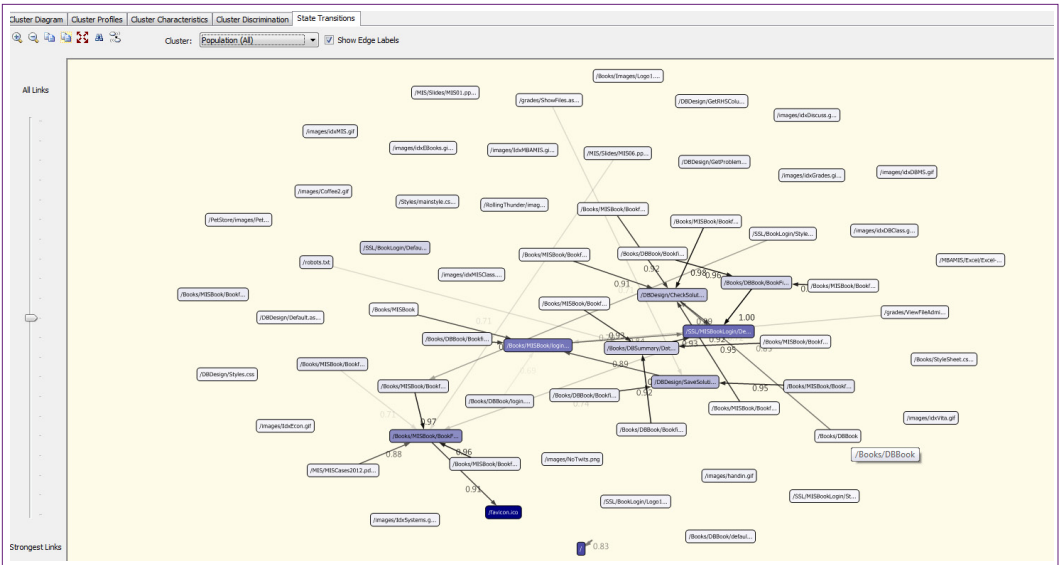


Figure 9.19

State transitions. Transitions are the probabilities of moving from one page to another. The result here is based on the population cluster or everyone. The drop-down box makes it easy to see how these transitions differ for each cluster or group of users.

State Transitions

Microsoft Sequence Clustering uses a Markov chain to estimate parameters within sequences. A particularly useful parameter is the state transition or a probability associated with the next page to be selected. The result-viewer provides a graphical method to view these probabilities. Figure 9.19 shows the base diagram. This diagram is displayed based on the entire population cluster so it shows the page probabilities for a given page based on everyone. The drop-down box makes it easy to display the transition probabilities for any of the clusters. More than any of the other views, the transition probabilities provide an insight into the sequence or flow of the action. The diagram and the probabilities show how the pages are related to each other in terms of sequential flow. By switching cluster views, it becomes possible to see how each group navigates the Web site differently.

The detailed transition states are also stored within the Analysis database. The *Microsoft Generic Content Tree Viewer* retrieves these values and displays them in a table or list. To see the probabilities, use the Viewer drop-down-list to change to the generic tree viewer instead of the Sequence Cluster viewer. As shown in Figure 9.20, under each cluster is a list of the pages labeled sequence state 0 through 63. This state is the starting item or page. Selecting one of these entries creates a list of the target or next pages along with two main values: support and probability. The probability is the transition probability computed from the Markov chain for moving from the existing state (page) to the specified next page for individuals within the selected cluster. The support value is based on the number of cases in the training data that moved from the starting item to the targeted next item. However, the count value is modified by the probability of the cluster being selected so the support values are often fractions. Remember that a

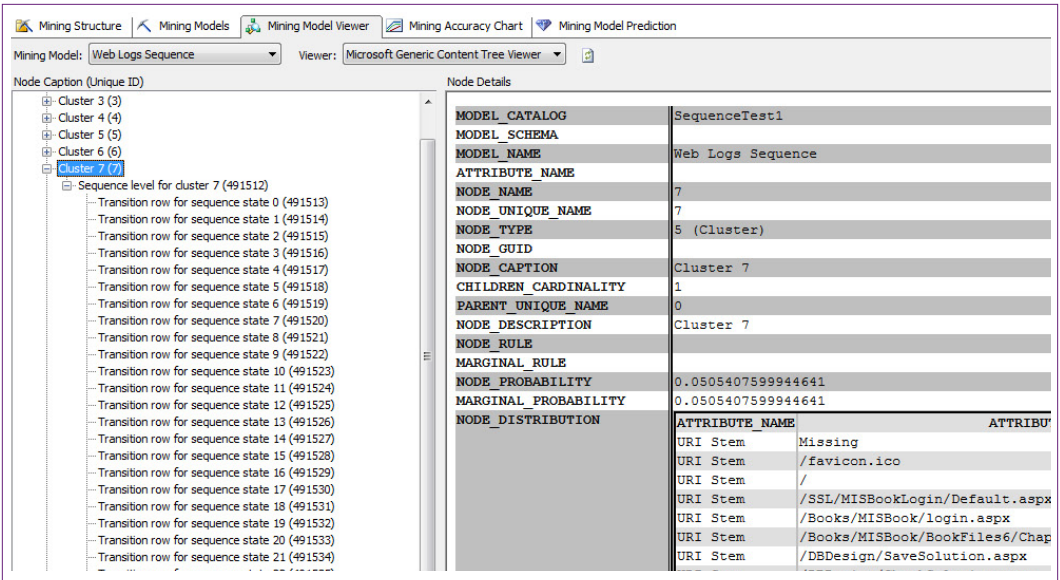


Figure 9.20

Generic tree viewer. The tree viewer shows the transition probabilities overall and for each cluster. Choose a starting state in the left window (64 entries). The middle scrolling window lists the target states and associated probabilities for the selected cluster.

given item can be associated with multiple clusters—up to a level of probability. Even for a specified cluster, the transition probabilities form a large matrix—from each possible state (page) value to each possible target page. The tree viewer displays one row of the matrix at a time: Select a starting value and it displays each of the target values and the associated probabilities.

Prediction

Prediction is not particularly useful, or supported, in sequence clustering. The transition state probabilities are the most useful for predicting behavior within a sequence. By themselves, the values are the conditional probabilities for each state. They are easily interpreted for a given a cluster or group, and a current page; they represent the probability of any moving to a specific page. Most Web sites are already designed to support certain types of flows. For instance, a product page supports search, then adding items to a shopping cart, followed by checkout pages. Users have the ability to choose options (such as return to add more products), so the transition probabilities show the chance that people within a given cluster will follow those paths. If some groups turn out to be more likely to follow detours, this information might be useful for altering the path options or finding ways to encourage others to follow them.

The clusters themselves are also defined in terms of probabilities. In the EM model, a group can be associated with multiple clusters—where the association is defined in terms of a probability. In traditional clustering, it is possible to enter values for dimensions and identify the clusters and probabilities associated with those specific values. This approach is difficult with sequences because it would be necessary to enter an entire sequence—not just a few dimension values.

Other Tools

What features are provided by other sequence mining tools?

The topic of data mining sequences is somewhat specialized. A few general tools exist, but almost all of them are commercial implementations somewhat similar to the Microsoft tools. Several add-ins also exist for the "R" statistical system to provide various tools for sequence analysis. However, most of the advanced sequence mining work has occurred in biology. In particular, the need to find patterns, align sequences, handle gaps, and other "messy" problems frequently occur in the analysis of DNA. Several specialized tools have been developed for these problems. Most are geared towards the DNA issues—notably a small number of items in the alphabet (CGAT), and potentially huge sequences. If a problem is encountered that has similar characteristics, it might be possible to define the problem in a form that can be handled by these specialized tools. For example, the U.S. National Institute for Health, specifically the National Center for Biotechnology Information) supports the BLAST (Basic Local Alignment Search Tool) tool (see the reference list for Web sites).

It is a little surprising that some of the DNA-specific features have seen little adaptation to business problems. However, the algorithms for biological research are highly optimized for their specific tasks in terms of performance and usability. The main task that they perform is comparing two or more sequences—in particular, the ability to search for a pattern from one sequence through a large collection of similar sequences. In DNA terms, a researcher might be interested in a specific pattern found in one or two sequences and then want to see if that pattern exists in other people.

In business terms, a close analogy might be found in terms of loan repayments. It is possible that a pattern of late or missing loan payments is an indicator that a person will file bankruptcy or fail to pay off a loan. If the sequence data existed for a large number of borrowers, it might be possible to search for similar patterns and then use those to identify future defaults. In a broader context, researchers have used similar techniques to examine young people moving through educational programs and into the workforce. So the tools have some value, but they can be difficult to apply to common business situations.

Sequence Summary

The analysis of sequences is relatively new compared to most of the other tools. Much of the advanced research is taking place in biology in the form of DNA and protein-sequence analysis. These tasks are relatively specialized and the high-end tools developed for those tasks may be challenging to use in business research. Nonetheless, some business applications can benefit from analyzing sequences. A sequence is an ordered list of items or events. The ordering is an important characteristic of the problem. Additionally, the items must be discrete and relatively small in number of values. To be useful, the researcher needs a large collection of sequences that have random orderings.

The most common analytical approach to business sequences is to search for clusters or groups of people who follow similar sequences. A typical example is the paths taken by visitors to Web sites. A related typical example is the order in which products are added to a shopping cart; which generally also requires Web site data. Other examples can be found, but most of them involve customers to provide the random actions. Think in terms of states or events where there is a need to predict which event will occur next.

Sequence Clustering is similar to traditional clustering, where the goal is to find groups of actions that are similar to each other but different from other clusters. Using Microsoft Sequence Clustering, the key is that a Case table is needed that contains a single entry for each person or primary sequence. A second, Nested table contains the values of the sequence for each person where the value for each person at each point in the sequence is stored in a separate row. This nested table must have a primary key that provides the correct order of the sequence values. The column containing the sequence values must be specified as an input column, but never a key column. The tool limits the number of sequence values—by default to 64. This limit can be changed but Microsoft recommends that it be kept under 100. Problems with a large number of items, such as complex Web sites, should consider reducing the extraneous pages (such as GIF files) or perform multiple analyses on sections of the site.

The results of sequence clustering consist of two main components: (1) The cluster estimates, and (2) transition state probabilities. The cluster results are similar to other clusters but they need to be evaluated in terms of the items (pages) and the sequence of items. The transition probabilities are better at highlighting the dynamic nature of the sequence. They are derived from a Markov chain model and estimate the chance that a new item will be chosen next in the sequence given the current position. In Web site terms, if a person is viewing one page, the transition probabilities show the probability for each page that the person might move to next. These probabilities can be a general statement for everyone, or they can be found for each cluster or group.

Geographic Analysis

What role does location have in decisions? Many people are familiar with the common mapping tools available on the Web, cell phones, and **global positioning system (GPS)** devices. These tools play a role in making some types of decisions, but business decisions are better supported by a more powerful **geographic information system (GIS)**. GIS tools include the basic mapping functions but focus on the ability to display layers of business, demographic, and technical data on top of the maps. In essence, the analytic decisions examine the question of how the business data elements are correlated due to location.

Increasing amounts of business data are available that are related to location. In the U.S., the federal government collects and distributes large amounts of demographic and economic data. For example, the demographic data available from the Census Bureau is tied to ZIP Code and often the more-detailed Census district codes. Consequently, basic information on consumers is readily available at many geographic levels. Also, several private companies collect detailed location data on businesses (particularly Google). Some sell the data and others sell geographic analyses using customized data sets. Local and state governments routinely use GIS tools to track services—such as utility (water) lines and crime incidents.

Many business decisions involve the use of location, such as siting for retail stores—which requires knowing the customer demographics of many areas. Building new factories raises similar issues of access to transportation (highways, ports, and railroads), as well as number, age, and educational level of potential workers. As more data becomes available, more complex questions can be addressed, including locations of specialty provider firms (such as machine shops), and environmental issues such as access to water, waste disposal, pollution levels, and so on.

Figure 9.21

Sample GIS color-coded layer. A GIS such as Microsoft MapPoint can display data within regions by color-coding the display. Here, Per Capita Income for 2007 is displayed at the County level. However, MapPoint can display only one color-coded layer at a time, so business data has to be displayed in a different form such as small charts or variable-sized circles.

Several governmental policy questions focus on geographic questions. For instance, health care potentially has many different regional differences. The Dartmouth Atlas of Health Care has some interesting charts online for comparing various national issues in health care. The term **geographic correlation** appears most commonly in health care research—where researchers test for differences in health care statistics across regions. The basic practice is to define a set of regions (such as hospital referral regions), collect data for various events within those regions and then compute and evaluate correlations for the various dimensions across the regions.

A GIS is different from a basic mapping system in that a GIS has the ability to display various data types on maps as layers. A **layer** is just one set of data tied to location. It might be displayed as a segment—such as a pipeline or travel path, a marker or pin, a small chart, or a color-coded region. Figure 9.21 shows a simple GIS presentation of per capita income at the county level, shown as a color-coded map. Microsoft MapPoint is used to display per capita income for 2007 at the county level. The data is from the Census Bureau and is built into MapPoint, making it relatively easy to display. However, MapPoint supports only one color-coded layer at a time. Any additional business data has to be displayed using small charts, pins, or variable-sized circles.

Business Situation

What types of business problems are related to location? The location question is challenging to answer—at some levels almost any business decision has elements of geography or location embedded in the question. Selling almost anything to customers could be impacted by the customer's physical location. Even for online and digital sales, companies will want to track a customer's location—partly for taxation and currency issues; but also for longer-term questions of how to appeal to more customers in the same location.

The heart of a geographic analysis is that the data can somehow be tagged to a specific physical location. The location is not always specified down to an exact spot. As the availability of GPS data increases, more items can be identified to fairly tight locations. However, in many situations, location is identified at a wider level, such as by ZIP Code, county, or state. In fact, in most cases location at a specific point is too narrow. Rarely does a business need to know an exact spot—which might represent only a single person—instead, the area is more important because it represents a collection of potential customers, or employees, or suppliers. Along the same lines, location is typically viewed as a hierarchical concept—such as: Nation – State – County – ZIP Code. And managers often want to examine values at each level of the hierarchy—zooming in to see details when necessary.

In other cases, detailed local geography can be critical. For instance, consumers might be reluctant to cross a specific railroad track, highway, or river to shop at a store on the other side. A challenging aspect of geographical analysis is that local knowledge is often important for detailed analysis. At a higher level, sales by state or county are straightforward, but as the problem becomes more local, then additional information becomes useful to understanding the results. It is these areas where specialist firms hire people to collect detailed local data.

The bottom line is that many business problems are related to location. The key is to be able to collect the data with the associated location information. Sales are often easy to tag because individual stores or shipments automatically contain addresses. Even online sales generally collect address data. Likewise, purchases from suppliers and shipping data typically contain location information. Large companies that have production operations in multiple locations can also track geographical data on employees, production times, quality, and other measures that might be affected by geography or local demographic issues. Similarly issues in human resources that might seem to be unrelated to location can be affected by local cultures, such as overtime, sick days, and overall health costs. In some situations, data that might not seem to be related to location could be affected by other variables that are local—including culture, weather, income, and local demographics. In many cases, the only way to identify these effects is to look at the data and then search for explanations when unexpected differences do arise.

Data

Probably the most important data source for geographical information is the U.S. Census Bureau. Census collects a vast amount of demographic and some economic data on a regular basis and all of it is geographically tagged. Almost all of the publicly-available data is available for download. Census also created some of the early mapping datasets through its TIGER mapping system—and much of this data is also available for download. Although, today, it is simpler to rely on

commercial software packages to draw the detailed maps—and they might use Census mapping data automatically. Census demographic data includes hundreds of series including "traditional" items such as age, gender, marital status, ethnicity, crime, and household size. It also includes economic and other data including income, employment, occupation, housing, industry data, commuting time, and weather. Census data is available at several geographic levels, including the Census District. The Bureau collects data at the household level but it is forbidden from releasing it until 70 years after it was collected. Interestingly, private companies have filled in some of the gaps. For example, Google routinely makes available the location of individual houses—including satellite photos; but Census (and the Post Office) is not allowed to provide this data. Another drawback to government-provided data is that release can be delayed—often for several years. For instance, the economic values from the Census might be two or more years out of date; which means that data collected every five years could be six or seven years out of date.

Companies also collect internal data—largely in terms of sales revenue and production costs. To handle this data geographically, it needs to be tagged or identified with location information. At the simple level, most data will have an address which includes the state or ZIP Code. ZIP Code level is probably the most detailed level available to most organizations. ZIP Codes are defined by the Post Office and loosely correspond to individual offices. Larger cities have multiple ZIP Codes that are organized around the physical Postal buildings. Smaller cities might include several cities within the same ZIP Code. The ZIP Code level is typically considered to be narrower than an individual city yet refers to a defined geographic area. The specific area is defined by the Post Office which defines the specific geographic region. GIS tools generally contain the mapping points for those regions.

The other way to tag data is to find a way to collect it along with GPS coordinates. GPS hardware prices have declined and are now embedded in many devices. However, collecting that data is a challenge—particularly in terms of privacy issues. Certainly, in-house data can be collected and stored by GPS-provided latitude and longitude. But data on customers can be problematic. Although cell phones collect GPS data, many people block this information (except to emergency providers—which cannot be blocked). Still, if a business knows the exact location of its stores or kiosks, then it knows the location of most customer interactions. Data at this level could be useful when examining detailed traffic patterns, and when sales are made at many small locations.

The Rolling Thunder Bicycle data will be used in the following sections to demonstrate some of the techniques of geographical analysis. In particular, the sales data is available at the city level. Although the data is simulated, it was generated with few specific patterns that can be investigated geographically. The database also contains valid Census data for several thousand cities which can be used to highlight comparisons.

Model

The GIS approach is still commonly used to examine geographic data. The focus is on mapping the data to make it easier to see geographical patterns. For instance, sales by state might reveal stronger sales in southern or western states. This approach works well using color-coded shading when the focus of the analysis is a single data series. When multiple elements are being modeled, the maps can be-

come more difficult to create and read. For example, consider the Rolling Thunder Bicycle case and think about trying to display sales of bicycles. If the goal is to display just the total sales value of all bicycles, only a single series is needed and a color-coded region can display the relevant data. But what if you need to display sales by model type as well as location? Or what if it is important to display demographic data as well as sales, such as population or income levels? These additional dimensions are known as layers in geographic models.

In the old days when geographic data was largely handled manually, a layer of data was often created on a separate sheet—typically on acetate (clear plastic). Then any layer could be overlaid on the map—making it possible to stack up several layers to see multiple dimensions at the same time. Most GIS tools support a similar concept where additional dimensions can be added to any map. However, the issue of color-coded regions is a problem. Trying to display a second color-coded region on top of an existing one can make it difficult to see both sets of data. For this reason, few tools support multiple color-coded layers. It can work—it just requires careful consideration of the colors and intensities.

The issue of multiple dimensions is important in the analysis of most geographical data. In some cases, the effect of location is primary—perhaps sales near beaches or in the south are different from other locations. However, many situations require examining the correlation between multiple dimensions. For example, sales are likely to be correlated with per capita income or even population. Visually demonstrating these relationships requires creating multiple layers of data on a map.

A second type of correlation involves point data. A classic example is local data on crime statistics. For example, a police department could use push-pin markers to indicate the location of various crimes. One color could be used for drug busts, a second color for gun or knife attacks. If the two colors commonly appear together, it would indicate a correlation between the two types of crimes. It is relatively easy to plot these types of layers with most existing GIS tools. However, the point-location data is generally less useful for business applications because most businesses operate from a few fixed locations. But the tools could be helpful for companies with thousands of small locations (such as a coffee chain), or when tracking data through individual resellers such as convenience stores and other retail outlets.

The issue of data correlation is a critical element in geographical analysis. The GIS tools can make it easier to spot patterns, and they definitely make it easier to show others that a pattern exists. But it can be difficult to see correlations when the data dimensions are related to each other instead of directly to location. Examples are covered in the next section. But, the point is that sometimes traditional correlation or regression analysis is more useful. For instance, medical researchers rely on traditional statistics to analyze health data. The basic process is to examine the desired dimensions in terms of a specific region. The Dartmouth Atlas of Health Care defines hospital referrer regions (HRR) that correspond to major regional hospitals. Each hospital draws physicians and patients from a surrounding area. By collecting statistics within each region, it becomes possible to compare the data statistically.

Figure 9.22 shows a small example of correlation. Hypothetical data was created for sales and income. Both dimensions are plotted on the map with the income shown using region shading and the sales denoted by the size of a circle. Even with this small number of states, is it possible to see a correlation between

Figure 9.22

Sample correlation problem. The shading represents income and the circle size is defined as the sales. Is there a correlation between these two values? Perhaps, but it would be easier to see using traditional statistical tools. In this case, the simple correlation coefficient is 0.86 and a strongly significant regression coefficient from income to sales.

the two dimensions? Perhaps. But a simple statistical computation shows that the correlation is 0.86 and a regression analysis generates a strongly significant coefficient between income and sales. So, the statistical tests verify a correlation, and also provide a numerical measure and a test of the significance. The point is that there is a geographical relationship but it is actually due to a correlation between two dimensions, where one (income) is geographically related. The geographical aspect will become important later if the company wants to search for new sales territories—the data suggests focusing on high-income areas.

Microsoft MapPoint

Several tools provide basic GIS support—many of them are expensive. Microsoft sells MapPoint for a comparatively reasonable price, and trial versions are available. The interesting feature of MapPoint is that it includes preconfigured Census demographic data. The Census data can be used along with your own data. The Rolling Thunder Bicycle data is useful for exploring MapPoint capabilities. The ZIP Codes in RT are accurate only to the city level, so it is better to start with the state data.

The first step in any analysis is to configure the data in the format needed for the tool. MapPoint (and other tools) make it relatively easy to read data in formats ranging from text files to databases. Because the RT data is already in a DBMS, the main step is to create a query to select and format the data. Microsoft Access or SQL Server are the easiest to use. Assume that managers want to explore 2012

```
SELECT Bicycle.SaleState, Sum(Bicycle.SalePrice) AS SumOfSalePrice
FROM Bicycle
WHERE (Bicycle.OrderDate Between '01-Jan-2012' And '31-Dec-2012')
GROUP BY Bicycle.SaleState;
```

Figure 9.23

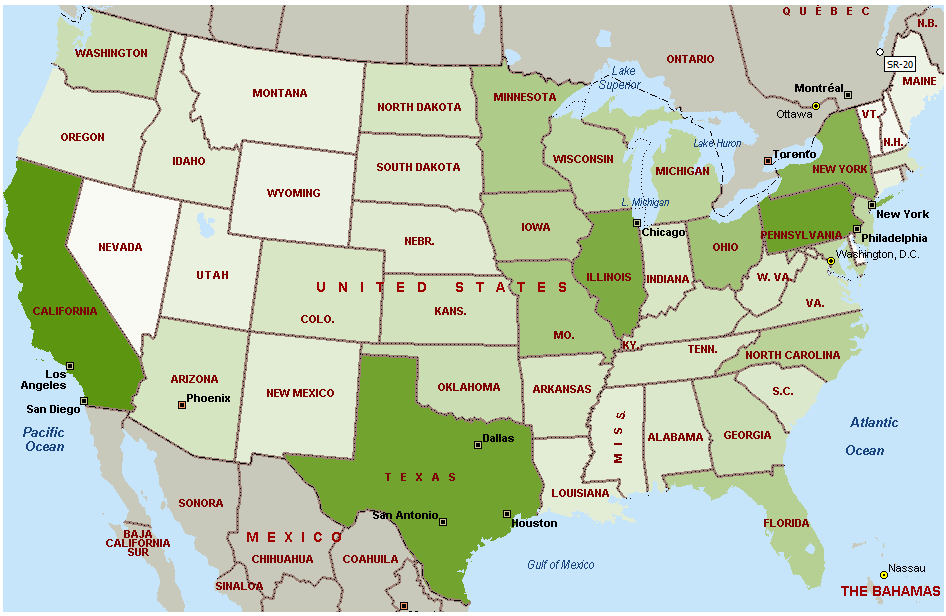
RT sales by state for 2012. The query is straightforward and should be saved as a view or query so that the GIS tool (MapPoint) can open the query directly.

sales by state. Figure 9.23 shows the basic query to examine sales in terms of value (versus quantity).

MapPoint has a Data Mapping Wizard that makes it straightforward to add data to a map. Set up the data in advance that contains a column with a location indicator. The location value should be at the most detailed level needed. MapPoint will provide any aggregation needed. For example, if data is available at the ZIP code level, use that and MapPoint will support aggregation to higher levels such as county and state. To begin, assume only one data series (sales) is going to be displayed. Start the wizard and accept the default choice of shaded area. Choose the option to link to the data (import is also acceptable but the link is dynamic). If the data is in Access, choose the Access option and find the database; otherwise pick the Link option and follow the prompts to connect to the database. Navigate and select the saved query. The wizard will automatically match the State column;

Figure 9.24

RT Sales by State for 2012 in MapPoint. The single dimension is relatively easy to see (once the road data is removed). But it is not immediately clear how the data are related to geography.



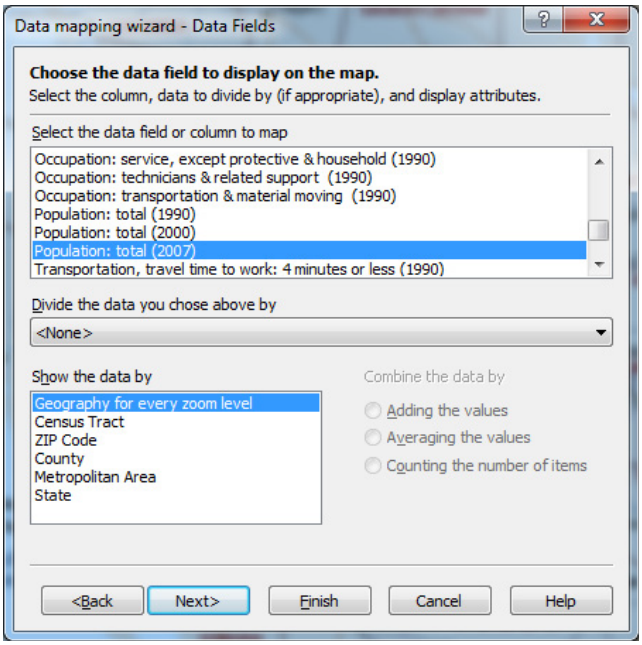


Figure 9.25

MapPoint demographic data. Many variables area available from the built-in Census database. In most cases, it is helpful to pick the option to display the data at every zoom level to obtain the best display at each level.

however, the database might include sales from Puerto Rico which has a code of PR, which is not considered a "state" by MapPoint (although it is a U.S. territory). So choose the option to skip the record.

Figure 9.24 shows the MapPoint chart. The chart is interesting, but it is not immediately clear how the sales data is related to geography. After a little thought, it appears that the states with the most sales are also the ones with the most population—which is not too surprising. So, it might be interesting to add a layer for population.

MapPoint has a large amount of Census data including population. The process is similar to adding your own data. However, MapPoint can use a shaded region for only one dimension. The second will have to be added using a different notation, such as a sized circle. So, it is necessary to think about which dimension should be shaded and which one as a circle (or other indicator). In this situation, the income can be shown as the background using the shading and then put the sales in the foreground. It is easiest to s tart over with a new map. Start the wizard but choose the option to add demographic data. Figure 9.25 shows the main step in selecting Population (Total 2007). It also shows the option to add the data for all zoom levels, which enables MapPoint to adjust the display for any point in the hierarchy.

After the demographic data has been added, restart the wizard to add the Rolling Thunder Bicycle data. Options for the display include Shaded Circle, Sized circle, Pie chart, Column chart, and push pins. The Sized circle is probably the easiest to see in this case. Figure 9.26 shows the resulting map and it is possible

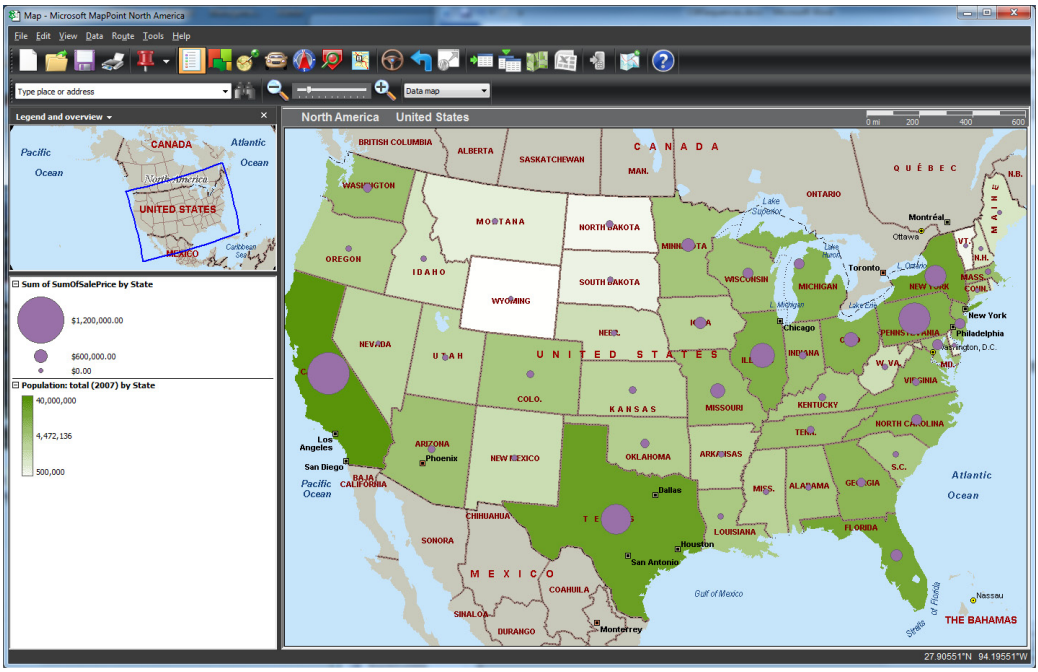


Figure 8.29

RT Sales versus population. The correlation between population and sales is relatively easy to see where larger circles appear in the states with higher populations (darker shades).

to see the correlation between sales and population. The correlation might be even easier to see if both layers could be presented as two differently-colored shaded regions, but MapPoint supports only a single layer as a shaded region. To support multiple layers might be useful but it requires colors and layers that support transparency so that the bottom layers show through to a certain extent; and it requires being cautious in its application. Note that the road layer was removed to make it easier to see the shadings. The individual layers can be edited by right-clicking the entry in the map key on the left side. The pop-up menu items support changing colors or even changing the data selection or display method.

Additional layers can be added using the same wizard, but beyond two or three layers, it becomes difficult to see the individual items. Multiple layers or dimensions work best by using push pins with a different color for each layer. This approach can highlight correlations between the dimensions by counting the number of times the colors appear together. But, it will work best at more-detailed location levels such as county, city, or ZIP Code. A comparison of only 48-50 states is usually not enough observations to highlight differences.

Other Tools

What tools exist for analyzing geographic data? In particular, what online tools are available? Microsoft MapPoint was used in the earlier sections because it is available are a relatively low cost and is straightforward to use. However, several other tools offer additional features. The major commercial GIS

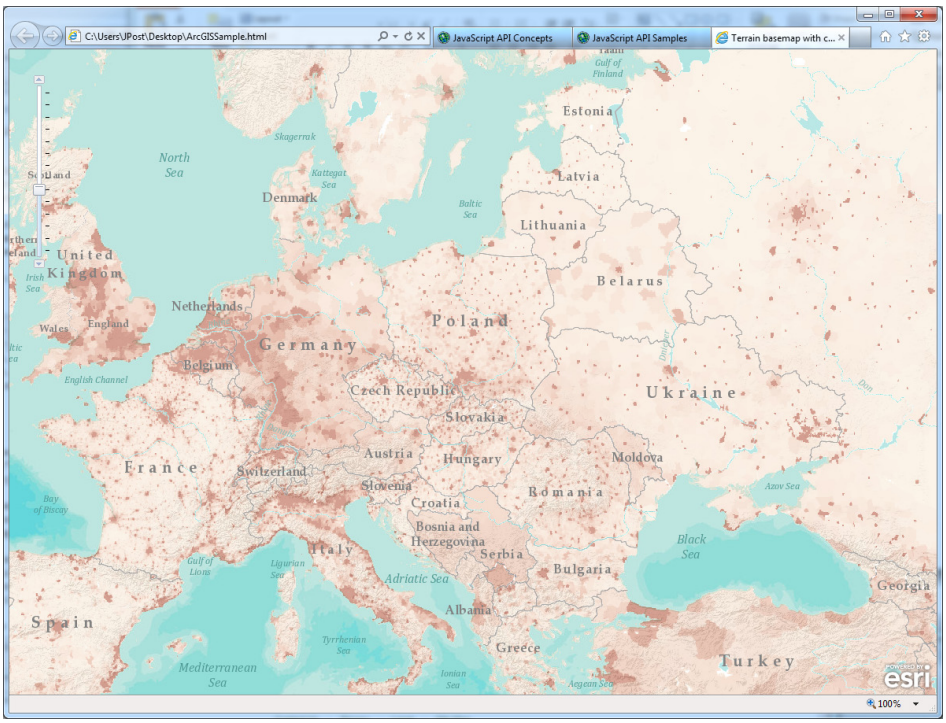


Figure 9.27

Esri ArcGIS map created with an HTML/Javascript file. The HTML file comes from the Esri documentation and uses Esri servers to load the base map and the population data layer. The country and names data are loaded as a third layer.

tool is sold by Esri, and it is heavily used by large organizations and government agencies.

Largely led by Google and its purchases and research in online tools, many people have turned to online GIS facilities. Google and Microsoft's Bing, the leading search engines, both have extensive mapping tools that can be integrated into other Web sites. Both are free to explore with documentation and sample code online. Both also charge for extended use of the services. Commercial Web sites with millions of visitors could end up paying substantial amounts of money if everyone uses the online mapping tools. However, a site with that many visitors could probably make money through sales or advertising. More recently, Esri has added online services as well—many of them can be used for free; and Esri tools provide detailed options that might not be available with the other tools.

Esri

One of the earliest and most powerful GIS tools is ArcInfo by Esri. The current product name is ArcGIS and it is available as a standalone desktop standalone software tool or on online Web-based system. ArcInfo/ArcGIS is commonly used in government organizations. It can handle complex graphing and large datasets, but is relatively expensive. On the other hand, the company has been increasing access to free offerings online, including base maps such as USA topographic, ocean, and the National Geographic World map. The online system can be inte-

```
<!DOCTYPE HTML PUBLIC “-//W3C//DTD HTML 4.01//EN”
   “http://www.w3.org/TR/html4/strict.dtd”>
<html>
<head>
<meta http-equiv=”Content-Type” content=”text/html; charset=utf-8”>
<meta http-equiv=”X-UA-Compatible” content=”IE=7,IE=9” />
<!--The viewport meta tag is used to improve the presentation and behavior of the
samples on iOS devices-->
<meta name=”viewport” content=”initial-scale=1, maximum-scale=1,user-scalable=no”/>
<title>Terrain basemap with custom data</title>
<link rel=”stylesheet” type=”text/css” href=”http://serverapi.arcgisonline.com/jsapi/
arcgis/3.0/js/dojo/dijit/themes/claro/claro.css”>
<style>
html, body { height: 100%; width: 100%; margin: 0; padding: 0; }
#map{padding:0;}
</style>
<script type=”text/javascript”>var djConfig = {parseOnLoad: true};</script>
<script type=”text/javascript” src=”http://serverapi.arcgisonline.com/jsapi/arcgis/?v=3.0”>
</script>
```

Figure 9.28

Initialization of the Esri HTML and Javascript. The first few tags simply define the location of the Esri server script and style files.

grated into Web sites to display basic location data such as a identifying a store location. More sophisticated data integration requires a subscription. Esri also has substantial amounts of geocoded data, including Census demographic data. But most of this data requires a subscription to use. Esri also has an “analytics” package to help integrate company data with the Esri databases. The free online version makes it possible to add up to 1,000 points to a standard Esri map and share that map with others or make it part of a Web site. The process is similar to that used by Google.

One of the nice features of ArcGIS is the ability to set the opacity (conversely the transparency) of a layer—allowing lower layers to show through. This feature also works with the online tools. Figure 9.27 shows a sample map created using Esri’s documentation. It contains three layers: The base map of continents and water features, population shown as shaded regions, and the political boundaries and names. The interesting feature of the map is that it was generated from a single HTML file using Javascript API commands to the Esri servers to obtain the base map and data layers. It is also possible to add custom data instead of the population values. Most applications today use **application programming interfaces (APIs)** or function calls to pass information from code to the server.

Figure 9.28 shows the first part of the HTML/Javascript file to create the Esri map. The first steps are largely html commands to identify the style sheet and location of the script files on the Esri server.

Figure 9.29 shows the main Javascript commands to handle the mapping tasks. The bottom (first) layer is simply the base map that is provided from Esri servers. The top (last) layer is also from an Esri server that provides the national boundaries and names. These two layers (or similar ones) will likely be a part of almost any map you generate. The middle layer retrieves and displays the population data using shaded regions. It comes from a separate Esri server used for demonstra-

```
<script type="text/javascript">
dojo.require("dijit.layout.BorderContainer");
dojo.require("dijit.layout.ContentPane");
dojo.require("esri.map");
var map;

function init() {
var initExtent = new esri.geometry.Extent({"xmin":-5.54,"ymin":40.81,"xmax":44.46,"ymax":
58.35,"spatialReference":{"wkid":4326}});
map = new esri.Map("map",{ extent:esri.geometry.geographicToWebMercator(initExtent)
});
//Add the terrain service to the map. View the ArcGIS Online site for services http://
arcgisonline/home/search.html?t=content&f=typekeywords:service
var basemap = new esri.layers.ArcGISTiledMapServiceLayer("http://server.arcgisonline.
com/ArcGIS/rest/services/World_Terrain_Base/MapServer");
map.addLayer(basemap);

//add custom services in the middle. This is typically data like demographic data, soils,
geology etc.
//This layer is typically partly opaque so the base layer is visible.
var operationalLayer = new esri.layers.ArcGISDynamicMapServiceLayer("http://
sampleserver1.arcgisonline.com/ArcGIS/rest/services/Demographics/ESRI_Population_
World/MapServer", {"opacity":0.5});
map.addLayer(operationalLayer);

//add the reference layer
var referenceLayer = new esri.layers.ArcGISTiledMapServiceLayer("http://server.
arcgisonline.com/ArcGIS/rest/services/Reference/World_Reference_Overlay/
MapServer");
map.addLayer(referenceLayer);

dojo.connect(map, 'onLoad', function(theMap) {
//resize the map when the browser resizes
dojo.connect(dijit.byId('map'), 'resize', map,map.resize);
});
}

dojo.addOnLoad(init);
</script>
```

Figure 9.29

Esri Javascript to load three layers for the map. The bottom (first) layer is the Esri base map. The top (last) layer is the Esri reference layer that loads the country boundaries and names. The middle layer is the data layer that shows the population, which could be replaced with a custom service.

tions. This data layer would be replaced by calls to your own data source to plot custom data.

Figure 9.30 shows the final HTML needed to display the actual map. The work is handled by the two tags by using the Esri-specific "dojotype" attribute. Almost all of the work is handled automatically by the Esri pre-written scripts. Defining your own data source takes a few more steps, but the Esri documentation shows how to load data from a variety of sources, including sample text data from online Web sites. Note that the HTML file itself can exist on almost any device. For example, the lines can be copied and pasted into a text file and the graph will be displayed simply by opening the file in a browser.

```
</head>

<body class="claro">
<div dojotype="dijit.layout.BorderContainer" design="headline" gutters="false"
style="width: 100%; height: 100%; margin: 0;">
<div id="map" dojotype="dijit.layout.ContentPane" region="center"
style="overflow:hidden;">
</div>
</div>
</body>

</html>
```

Figure 9.30

Simple HTML for the Esri map file. The remaining portion of the HTML file simply defines the divs needed to display the actual map.

Google

Google is probably the best-known online map provider, and Google continues to add features to its tools. It also provides the desktop Google Earth application which supports more interaction than the standard online maps. In terms of GIS capabilities, Google supports several online tools to embed maps in Web sites and add points of interest. To a point, Google tools are provided free of charge. But even for publicly-available data, the tools impose limitations—largely on the number of daily users. The developer Web site (https://developers.google.com/maps/licensing) lists the details.

Google also uses HTML and Javascript to pass data to the servers and display maps. Documentation is provided on the Google maps Web site. One catch is that you have to sign up for a personal account first to obtain a personal key value. A second issue is that it is relatively easy to display markers (icons) and paths or routes on Google maps. It is more challenging to created filled areas (using polygons) to display data. Consequently, most online Google maps are basic maps with markers.

Google has many other online tools and some of them are being integrated into the mapping environment. In particular, a Fusion Table is a simple row-column database table that can be stored online. A Fusion table that holds location data can be used to generate a layer of data for a Google map. For instance, a Fusion table could be created that lists the location retail stores for a company (based on its address). Google maps can plot the location of each of the stores simply by passing it the data from the Fusion table. Some data limits are imposed, such as no more than 100,000 rows of data for mapping and a limit of one million bytes of data.

The terminology, APIs, and integration are unique to Google. Documentation is available online and some samples exist to help learn the system. However, Google is also known for continually "upgrading" its capabilities, so expect changes over time. Additionally, you will have to register for a personal account to obtain a special key that must be included in all of your code. This key value is used to monitor traffic and usage so that if the limits are exceeded, you can be billed for the additional data usage.

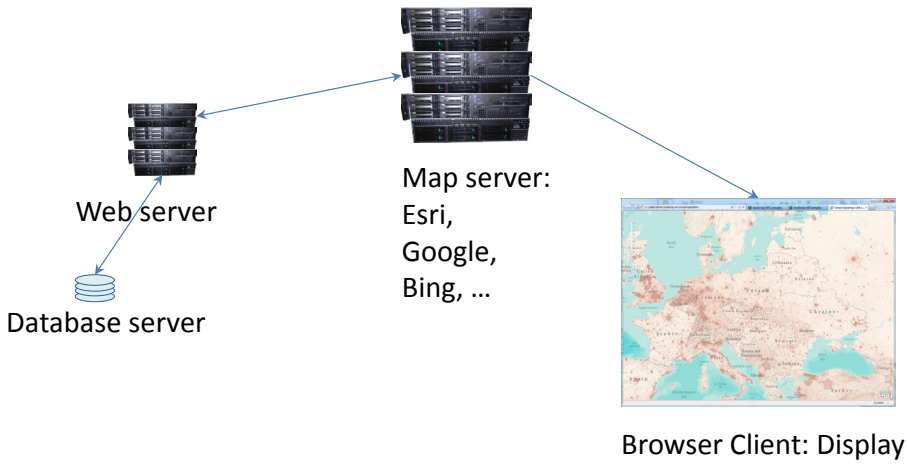


Figure 9.31

Web tiers for mapping. The map and data layers are generated by a map server in a form that is displayed by the client browser window. Any custom data must be generated by a Web application that retrieves the desired data, converts it to a form acceptable to the map server and functions as a Web service.

Bing

Microsoft's Bing maps has similar features to Google and Esri. In particular, you have to sign up for an account and use the AppID in all of the files. Accounts with high numbers of transactions will be billed. The overall development process is similar to Esri and Google—develop HTML and javascript code as a Web page that connects to the server. Documentation is available at http://www.microsoft.com/maps/developers. One useful tool is an AJAX Control 7.0 which makes it easier to find the special APIs and install them into your file. Selecting items from a list automatically install the corresponding API code into your file, which you then customize for your specific needs.

Microsoft also provides many other tools, including compilers and databases, including the online Azure SQL tools. Although SQL Server supports spatial data types, as of 2012, Azure does not automatically connect to Bing maps. However, several third-party companies provide connectors to facilitate the transfer and conversion of data into a form that Bing maps can handle. Search the Codeplex sharing site for examples such as the ajaxmapdataconnector that works with Microsoft's AJAX 7.0 API tool.

A key step to remember whether using Google, Esri, or Microsoft is that Web mapping involves three major elements: (1) The client browser, (2) The map server, and (3) A data server. Figure 9.31 applies to all of the online mapping services. To provide custom data on a map, you need to develop a Web application that runs on your own server to retrieve data, format it for the map server, and provide it on command as a Web service. In many cases, this step requires the help of a Web server programmer.

Census Bureau
http://www.census.gov
USGS: United States Geological Survey
http://www.usgs.gov
Geospatial
http://geo.data.gov
FAA
http://aeronav.faa.gov
NOAA: National Oceanic and Atmospheric Administration
http://www.nauticalcharts.noaa.gov
NGA: National Geospatial Intelligence Agency
http://msi.nga.mil
Others: Many agencies maps to display data
Local/County: Many use Esri to display property maps and local data. Search for the county assessor.
Canada
http://www.fedmaps.com
United Nations
http://www.un.org/Depts/Cartographic/english/htmain.hml
http://www.grida.no/graphicslib

Figure 9.32

Government sources of maps and data. Some online data and maps are free but watch the dates. It costs time and money to update map data so some sources take time to get current data.

Federal Government

Several government agencies have begun providing mapping data online. In many cases, this data is already integrated with standard mapping services—particularly Esri. In a few cases, the maps are created automatically, but with limited options. For example, the Census Bureau produces several maps to highlight interesting data. The government access portals have changed over time so custom searches might be needed to find specific data and maps. However, the geospatial site http://geo.data.gov is designed to handle some standard mapping tasks.

For many years (centuries), the U.S. Geological Survey (USGS) has been responsible for producing accurate maps of various features. In particular, the USGS is well-known for its detailed topographical maps. Most of these maps were printed and sold at various locations. Most of the printed maps are still available, and the large format makes them convenient for many purposes. However, the USGS is in the process of converting the maps to digital format. A few private companies resell this data for GPS units. The USGS is working to provide the topo maps for free download. The maps show detailed points of interest, but are often out of date. Still, they can be critical when examining and displaying data at a very local level of detail. Figure 9.32 lists some of the common sources of government maps. Some of the online maps and data are free, but it takes time and money to update and create maps so current, high-value data often carries a charge. One interesting source of map data is not a government agency but an organization that uses crowd sourcing to collect and share GPS data: www.openstreetmap.org.

By itself a basic map is not a GIS and it is not a geographical analysis. However, base maps provide the foundation for many of the displays. Placing your data on a map with the proper content and level of detail will make it easier to see and

understand the analysis. Hence, it helps to have multiple choices when selecting a base map.

Geographic Summary

Many business decisions are related to location. In particular, almost any data related to customers or suppliers can be tied to location—even if only at a national or state level. In one sense, every organization constantly searches for new customers, so it makes sense to search for new customers in the same location as existing customers. More broadly, the phrase "same location" actually means locations with similar demographics. So a key aspect of geographical analysis is to identify common demographic traits among existing customers and then find other areas that have similar demographics. In the U.S., the Census Bureau demographic data is useful for estimating customer characteristics and for identifying similar locations across the United States.

Geographic analysis goes beyond basic mapping. The analytical component often consists of finding correlations across dimensions and demographics. Data variables can be directly correlated with geography—such as latitude, access to water, days of sunshine, and so on. Or, multiple data sets can be correlated with each other through their location. For example, sales might be tied to a certain level of per capita income and people with that level of income might tend to live in specific areas—such as suburbs. The main key to beginning a geographic analysis is to collect and tag data by location. The location can be highly detailed—such as GPS coordinates; or simpler items including address, ZIP Code, city, or state. GIS tools display the data on a base map, using shaded regions, markers, or charts. The displays are useful for visualizing relationships, but statistical analyses on the data by region can provide more detailed estimates and measures of significance.

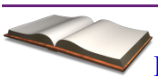
Key Words

application programming interfaces (APIs)
bins
classification
clustering
comma-separated-values (CSV)
discretize
edit distance
gap
geographic correlation
geographic information system (GIS)
global positioning system (GPS)
layer
Levenshtein distance
Markov chain
sequence
transition probability

Review Questions

1. What are the two primary types of sequence analysis?
2. What are the key data elements needed to perform sequence clustering analysis?
3. Why do sequence values need to be discrete instead of continuous?
4. How is missing data handled in sequence analysis?
5. What are the sequence and random elements available from Web server logs?
6. With Microsoft Sequence Clustering what are the roles of the Case and Nested tables?
7. With Microsoft Sequence Clustering is a sequence uniquely identified with a single cluster?
8. What are the differences between cluster views in Microsoft Sequence Clustering?
9. How do state transition probabilities reflect the dynamic sequence?
10. How is geographic analysis different from a mapping system?
11. How is indirect geographic correlation different from direct correlation?
12. What is the standard hierarchy for location data?
13. Why is it important for GIS tools like MapPoint and ArcGIS to include demographic data?
14. Which GIS tools support layer opacity to make it easier to display multiple data series?
15. What basic tools are needed to provide custom data to an online mapping system for use by your customers?

Exercises



Book

1. Set up and run the sequence clustering example using the Web site data. Write a brief report to the site manager about the results.

2. Research how the BLAST tool is used in DNA research and summarize the main characteristics of the types of problems it solves. Provide a business or social (non-biological) example where the tool might be used if it can be modified.

3. Explain how sequence mining might be used in a university to help students.

4. Select a business category and briefly describe how sequence analysis could be used to help a company in that business. Identify the data that would need to be collected.

5. Use an online tool (Esri, Google, or Bing) to create a map with markers for at least three locations that are of interest to you.

6. Find two government data series and identify the geographic levels available.

7. Research Esri's online ArcGIS and briefly explain the command needed to retrieve text data from an existing Web site and display it on a map.

Rolling Thunder Database

8. Compute total sales by month for each year. Let each year become a sequence, discretize the sales data (try using standard deviations). Use sequence analysis to see if clusters exist.

9. Similar to the previous question, compute total sales by month by model type for each year. Each model type year is a separate sequence. Count the number of bikes and discretize the data. Use sequence clustering and include the ModelType as an additional input dimension to see if any patterns exist.

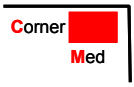
10. Use MapPoint or similar tool to explore correlations between sales value, population, and income.

11. Use MapPoint or similar GIS tool to see if there are differences in sales of model types based on elevation (altitude). That is, do people in mountainous regions buy more mountain bikes than road bikes?

Diner

12. The diner data contains a DOW (day of week) column. Use that column as the sequence order and compute total sales by DOW for each week in the year. Each week represents one sequence. Discretize the sales data—such as by using standard deviations. Run sequence clustering to find any patterns in weekly sales data.

13. Find state or regional data on the amount of money consumers spend on restaurant meals to help find a location for a new diner.



Corner Med

14. Count the number of procedures by each employee on each day and create a day-of-week (DOW) column for the date. Essentially, build a sequence for each employee for each week by DOW to count the number of procedures performed. Analyze those sequences by Week (case) to see if any clusters or interesting patterns exist.

15. Try to find health care statistics and use them along with income data to help identify potential locations for new Corner Med outlets. You can search nationwide or within a specific city.

Basketball

16. Compute the total points scored by a team by game for a season and run sequence clustering on the teams to see if any groupings exist. Add dimensions for whether the team made the playoffs and the division or region to see if they affect the clustering.

17. Compute total points scored by person by game for a season. Be sure to include missing values for games not played by a specific person. Run sequence clustering to see if groupings exist and comment on the results.

18. Compute the average points scored per game for the main NBA Divisions and plot them using a GIS tool to see if there are regional differences. Hint: Assign states to the regions.

19. Find the teams that won the final championship game for the past 20 years or so. Build a table with the list of the cities and the number of times each has won. Plot the results with a GIS and comment on them.

Bakery

20. The bakery SaleItem table includes the column Seq to indicate the order in which items were purchased. Assume this value is accurate and use sequence clustering to see if patterns exist in the order in which items are purchased. Use the product category instead of item name to reduce the number of item values.

21. Using online directories choose two similar-sized cities in two different states and count the number of bakeries in each city. Combine counts from other students and see if there is a significant regional difference. Based on the data where might you open a new bakery?

Cars

22. Find data on car sales by location and plot the data using a GIS. Hint: NADA has sales by state.

Teamwork

23. Have each person in the group find one additional tool that can be used to analyze sequence data. Briefly explain the purpose, the type of data, and types of analyses. Combine the reports and as a team identify the tool that offers the most usefulness to business. Include SQL Server in the final comparison.

24. Find at least three studies that use sequence analysis. Hint: Try research studies if you cannot find business results. Briefly summarize the data, the tool used, and the results found.

25. Find at least three studies or examples that use geographic correlations. Briefly describe the data, how the data was collected, and the results. Which study provides the best example of geographic analysis?

26. Select a local region and using a GPS unit (possibly on a phone), obtain the location of all diners and restaurants within that region. Plot the results on a mapping system. (Using Google Local or similar tools is cheating.)

Additional Reading

Dong, Guozhu and Jian Pei, 2007, *Sequence Data Mining*, Springer: New York. [A relatively short but technical (mathematical) reference for sequence mining methods.]

Esri documentation center: http://edn.esri.com/index.cfm?fa=doclibrary.gateway [Links to the online documentation for Esri products.]

Google documentation for maps. https://developers.google.com/maps/documentation/javascript/ API documentation: https://developers.google.com/maps/documentation/ [Links to using V3 of Google maps.]

Microsoft documentation for Bing maps. http://msdn.microsoft.com/en-us/library/dd877180.aspx [Includes links to the latest AJAX control.]

National Institute of Health: http://blast.ncbi.nlm.nih.gov/. Downloads at: ftp://ftp.ncbi.nih.gov/blast/executables/. For a description see http://www.ncbi.nlm.nih.gov/books/NBK21097/. [BLAST tools for aligning sequences—notably genomic/DNA sequences.]

Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 2001, *The Elements of Statistical Learning*, Springer: New York. [An outstanding book on data mining, with an emphasis on theory. A graduated-level book that requires a strong mathematics background.]