15

VISUALIZATION METHODS

Chapter Objectives

- Recognize the importance of a visual-perception analysis in humans to discover appropriate data-visualization techniques.
- Distinguish between scientific-visualization and information-visualization techniques (IVT).
- Understand the basic characteristics of geometric, icon-based, pixel-oriented, and hierarchical techniques in visualization of large data sets
- Explain the methods of parallel coordinates and radial visualization for n-dimensional data sets.
- Analyze the requirements for advanced visualization systems in data mining.

How are humans capable of recognizing hundreds of faces? What is our "channel capacity" when dealing with the visual or any other of our senses? How many distinct visual icons and orientations can humans accurately perceive? It is important to factor all these cognitive limitations when designing a visualization technique that avoids delivering ambiguous or misleading information. Categorization lays the foundation

Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition. Mehmed Kantardzic.

for a well-known cognitive technique: the "chunking" phenomena. How many chunks can you hang onto? That varies among people, but the typical range forms "the magical number seven, plus or minus two." The process of reorganizing large amounts of data into fewer chunks with more bits of information per chunk is known in cognitive science as "recoding." We expand our comprehension abilities by reformatting problems into multiple dimensions or sequences of chunks, or by redefining the problem in a way that invokes relative judgment, followed by a second focus of attention.

15.1 PERCEPTION AND VISUALIZATION

Perception is our chief means of knowing and understanding the world; images are the mental pictures produced by this understanding. In perception as well as art, a meaningful whole is created by the relationship of the parts to each other. Our ability to see patterns in things and pull together parts into a meaningful whole is the key to perception and thought. As we view our environment, we are actually performing the enormously complex task of deriving meaning out of essentially separate and disparate sensory elements. The eye, unlike the camera, is not a mechanism for capturing images so much as it is a complex processing unit that detects changes, forms, and features, and selectively prepares data for the brain to interpret. The image we perceive is a mental one, the result of gleaning what remains constant while the eye scans. As we survey our three-dimensional (3-D) ambient environment, properties such as contour, texture, and regularity allow us to discriminate objects and see them as constants.

Human beings do not normally think in terms of data; they are inspired by and think in terms of images—mental pictures of a given situation—and they assimilate information more quickly and effectively as visual images than as textual or tabular forms. Human vision is still the most powerful means of sifting out irrelevant information and detecting significant patterns. The effectiveness of this process is based on a picture's submodalities (shape, color, luminance, motion, vectors, texture). They depict abstract information as a visual grammar that integrates different aspects of represented information. Visually presenting abstract information, using graphical metaphors in an immersive 2-D or 3-D environment, increases one's ability to assimilate many dimensions of the data in a broad and immediately comprehensible form. It converts aspects of information into experiences our senses and mind can comprehend, analyze, and act upon.

We have heard the phrase "Seeing is believing" many times, although merely seeing is not enough. When you understand what you see, seeing becomes believing. Recently, scientists discovered that seeing and understanding together enable humans to discover new knowledge with deeper insight from large amounts of data. The approach integrates the human mind's exploratory abilities with the enormous processing power of computers to form a powerful visualization environment that capitalizes on the best of both worlds. A computer-based visualization technique has to incorporate the computer less as a tool and more as a communication medium. The power of visualization to exploit human perception offers both a challenge and an opportunity. The challenge is to avoid visualizing incorrect patterns leading to incorrect decisions and actions. The opportunity is to use knowledge about human perception when designing

visualizations. Visualization creates a feedback loop between perceptual stimuli and the user's cognition.

Visual data-mining technology builds on visual and analytical processes developed in various disciplines including scientific visualization, computer graphics, data mining, statistics, and machine learning with custom extensions that handle very large multidimensional data sets interactively. The methodologies are based on both functionality that characterizes structures and displays data and human capabilities that perceive patterns, exceptions, trends, and relationships.

15.2 SCIENTIFIC VISUALIZATION AND INFORMATION VISUALIZATION

Visualization is defined in the dictionary as "a mental image." In the field of computer graphics, the term has a much more specific meaning. Technically, visualization concerns itself with the display of behavior and, particularly, with making complex states of behavior comprehensible to the human eye. Computer visualization, in particular, is about using computer graphics and other techniques to think about more cases, more variables, and more relations. The goal is to think clearly, appropriately, with insight, and to act with conviction. Unlike presentations, visualizations are typically interactive and very often animated.

Because of the high rate of technological progress, the amount of data stored in databases increases rapidly. This proves true for traditional relational databases and complex 2-D and 3-D multimedia databases that store images, computer-aided design (CAD) drawings, geographic information, and molecular biology structure. Many of the applications mentioned rely on very large databases consisting of millions of data objects with several tens to a few hundred dimensions. When confronted with the complexity of data, users face tough problems: Where do I start? What looks interesting here? Have I missed anything? What are the other ways to derive the answer? Are there other data available? People think iteratively and ask ad hoc questions of complex data while looking for insights.

Computation, based on these large data sets and databases, creates content. Visualization makes computation and its content accessible to humans. Therefore, visual data mining uses visualization to augment the data-mining process. Some data-mining techniques and algorithms are difficult for decision makers to understand and use. Visualization can make the data and the mining results more accessible, allowing comparison and verification of results. Visualization can also be used to steer the data-mining algorithm.

It is useful to develop a taxonomy for data visualization, not only because it brings order to disjointed techniques, but also because it clarifies and interprets ideas and purposes behind these techniques. Taxonomy may trigger the imagination to combine existing techniques or discover a totally new technique.

Visualization techniques can be classified in a number of ways. They can be classified as to whether their focus is geometric or symbolic, whether the stimulus is 2-D, 3-D, or n-dimensional, or whether the display is static or dynamic. Many visualization tasks involve detection of differences in data rather than a measurement of absolute

values. It is the well-known Weber's Law that states that the likelihood of detection is proportional to the relative change, not the absolute change, of a graphical attribute. In general, visualizations can be used to explore data, to confirm a hypothesis, or to manipulate a view.

In *exploratory visualizations*, the user does not necessarily know what he/she is looking for. This creates a dynamic scenario in which interaction is critical. The user is searching for structures or trends and is attempting to arrive at some hypothesis. In *confirmatory visualizations*, the user has a hypothesis that needs only to be tested. This scenario is more stable and predictable. System parameters are often predetermined and visualization tools are necessary for the user to confirm or refute the hypothesis. In *manipulative (production) visualizations*, the user has a validated hypothesis and so knows exactly what is to be presented. Therefore, he/she focuses on refining the visualization to optimize the presentation. This type is the most stable and predictable of all visualizations.

The accepted taxonomy in this book is primarily based on different approaches in visualization caused by different types of source data. Visualization techniques are divided roughly into two classes, depending on whether physical data are involved. These two classes are *scientific visualization* and *information visualization*.

Scientific visualization focuses primarily on physical data such as the human body, the earth, and molecules. Scientific visualization also deals with multidimensional data, but most of the data sets used in this field use the spatial attributes of the data for visualization purposes, for example, computer-aided tomography(CAT) and CAD. Also, many of the Geographical Information Systems (GIS) use either the Cartesian coordinate system or some modified geographical coordinates to achieve a reasonable visualization of the data.

Information visualization focuses on abstract, nonphysical data such as text, hierarchies, and statistical data. Data-mining techniques are primarily oriented toward information visualization. The challenge for nonphysical data is in designing a visual representation of multidimensional samples (where the number of dimensions is greater than three). Multidimensional-information visualizations present data that are not primarily plenary or spatial. One-, two-, and three-dimensional, but also temporal information–visualization schemes can be viewed as a subset of multidimensional information visualization. One approach is to map the nonphysical data to a virtual object such as a cone tree, which can be manipulated as if it were a physical object. Another approach is to map the nonphysical data to the graphical properties of points, lines, and areas.

Using historical developments as criteria, we can divide IVT into two broad categories: traditional IVT and novel IVT. Traditional methods of 2-D and 3-D graphics offer an opportunity for information visualization, even though these techniques are more often used for presentation of physical data in scientific visualization. Traditional visual metaphors are used for a single or a small number of dimensions, and they include:

1. *bar charts* that show aggregations and frequencies;
2. *histograms* that show the distribution of variable values;
3. *line charts* for understanding trends in order;

4. *pie charts* for visualizing fractions of a total;
5. *scatter plots* for bivariate analysis.

Color-coding is one of the most common traditional IVT methods for displaying a 1-D set of values where each value is represented by a different color. This representation becomes a continuous tonal variation of color when real numbers are the values of a dimension. Normally, a color spectrum from blue to red is chosen, representing a natural variation from "cool" to "hot," in other words, from the smallest to the highest values.

With the development of large data warehouses, data cubes became very popular IVT. A *data cube*, the raw-data structure in a multidimensional database, organizes information along a sequence of categories. The categorizing variables are called dimensions. The data, called measures, are stored in cells along given dimensions. The cube dimensions are organized into hierarchies and usually include a dimension representing time. The hierarchical levels for the dimension time may be year, quarter, month, day, and hour. Similar hierarchies could be defined for other dimensions given in a data warehouse. Multidimensional databases in modern data warehouses automatically aggregate measures across hierarchical dimensions; they support hierarchical navigation, expand and collapse dimensions, enable drill down, drill up, or drill across, and facilitate comparisons through time. In a transaction information in the database, the cube dimensions might be product, store, department, customer number, region, month, year. The dimensions are predefined indices in a cube cell and the measures in a cell are roll-ups or aggregations over the transactions. They are usually sums but may include functions such as average, standard deviation, and percentage.

For example, the values for the dimensions in a database may be

1. region: north, south, east, west;
2. product: shoes, shirts;
3. month: anuary, February, March, . . . , December.

Then, the cell corresponding to (north, shirt, February) is the total sales of shirts for the northern region for the month of February.

Novel IVT can simultaneously represent large data sets with many dimensions on one screen. The widely accepted classifications of these new techniques are

1. geometric-projection techniques,
2. icon-based techniques,
3. pixel-oriented techniques, and
4. hierarchical techniques.

Geometric-projection techniques aim to find interesting projections of multidimensional data sets. We will present some illustrative examples of these techniques.

The Scatter-Plot Matrix Technique is an approach that is very often available in new data-mining software tools. A grid of 2-D scatter plots is the standard means of

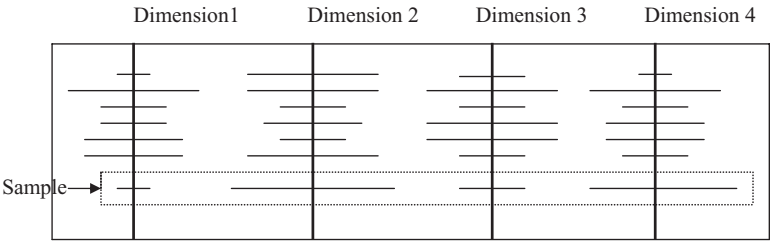

Figure 15.1. A 4-dimensional survey plot.

extending a standard 2-D scatter plot to higher dimensions. If you have 10-D data, a 10×10 array of scatter plots is used to provide a visualization of each dimension versus every other dimension. This is useful for looking at all possible two-way interactions or correlations between dimensions. Positive and negative correlations, but only between two dimensions, can be seen easily. The standard display quickly becomes inadequate for extremely large numbers of dimensions, and user interactions of zooming and panning are needed to interpret the scatter plots effectively.

The Survey Plot is a simple technique of extending an n-dimensional point (sample) in a line graph. Each dimension of the sample is represented on a separate axis in which the dimension's value is a proportional line from the center of the axis. The principles of representation are given in Figure 15.1.

This visualization of n-dimensional data allows you to see correlations between any two variables, especially when the data are sorted according to a particular dimension. When color is used for different classes of samples, you can sometimes use a sort to see which dimensions are best at classifying data samples. This technique was evaluated with different machine-learning data sets and it showed the ability to present exact IF-THEN rules in a set of samples.

The *Andrews's curves* technique plots each n-dimensional sample as a curved line. This is an approach similar to a Fourier transformation of a data point. This technique uses the function f(t) in the time domain t to transform the n-dimensional point X = (x1, x2, x3, . . . , xn) into a continuous plot. The function is usually plotted in the interval $-\pi \le t \le \pi$. An example of the transforming function f(t) is

$$f(t) = x1/1.41 + x2\sin(t) + x3\cos(t) + x4\sin(2t) + x5\cos(2t) + \ldots$$

One advantage of this visualization is that it can represent many dimensions; the disadvantage, however, is the computational time required to display each n-dimensional point for large data sets.

The class of geometric-projection techniques also includes techniques of exploratory statistics such as principal component analysis (PCA), factor analysis, and multidimensional scaling. Parallel coordinate–visualization technique and radial-visualization technique belong in this category of visualizations, and they are explained in the next sections.

Another class of techniques for visual data mining is the *icon-based techniques* or iconic-display techniques. The idea is to map each multidimensional data item to an

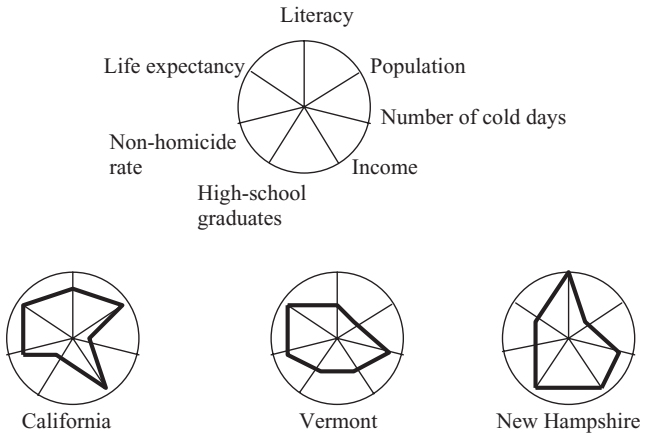

Figure 15.2. A star display for data on seven quality-of-life measures for three states.

icon. An example is the stick-figure technique. It maps two dimensions to the display dimensions and the remaining dimensions are mapped to the angles and/or limb lengths of the stick-figure icon. This technique limits the number of dimensions that can be visualized. A variety of special symbols have been invented to convey simultaneously the variations on several dimensions for the same sample. In 2-D displays, these include Chernoff's faces, glyphs, stars, and color mapping. Glyphs represent samples as complex symbols whose features are functions of data. We think of glyphs as location-independent representations of samples. For a successful use of glyphs, however, some sort of suggestive layout is often essential, because comparison of glyph shapes is what this type of rendering primarily does. If glyphs are used to enhance a scatter plot, the scatter plot takes over the layout functions. Figure 15.2 shows how the other icon-based technique, called a *star display,* is applied to quality of life measures for various states. Seven dimensions represent seven equidistant radiuses for a circle: one circle for each sample. Every dimension is normalized on interval [0, 1], where the value 0 is in the center of the circle and the value 1 is at the end of the corresponding radius. This representation is convenient for a relatively large number of dimensions but for a very small number of samples. It is usually used for comparative analyses of samples, and it may be included as a part of more complex visualizations.

The other approach is an icon-based, shape-coding technique that visualizes an arbitrary number of dimensions. The icon used in this approach maps each dimension to a small array of pixels and arranges the pixel arrays of each data item into a square or a rectangle. The pixels corresponding to each of the dimensions are mapped to a gray scale or color according to the dimension's data value. The small squares or rectangles corresponding to the data items or samples are then arranged successively in a line-by-line fashion.

The third class of visualization techniques for multidimensional data aims to map each data value to a colored pixel and present the data values belonging to each attribute in separate windows. Since the *pixel-oriented techniques* use only one pixel per data

value, the techniques allow a visualization of the largest amount of data that are possible on current displays (up to about 1,000,000 data values). If one pixel represents one data value, the main question is how to arrange the pixels on the screen. These techniques use different arrangements for different purposes. Finally, the *hierarchical techniques* of visualization subdivide the k-dimensional space and present the subspaces in a hierarchical fashion. For example, the lowest levels are 2-D subspaces. A common example of hierarchical techniques is dimensional-stacking representation.

Dimensional stacking is a recursive-visualization technique for displaying high-dimensional data. Each dimension is discretized into a small number of bins, and the display area is broken into a grid of subimages. The number of subimages is based on the number of bins associated with the two "outer" dimensions that are user-specified. The subimages are decomposed further based on the number of bins for two more dimensions. This decomposition process continues recursively until all dimensions have been assigned.

Some of the novel visual metaphors that combine data-visualization techniques are already built into advanced visualization tools, and they include:

1. *Parabox.* It combines boxes, parallel coordinates, and bubble plots for visualizing n-dimensional data. It handles both continuous and categorical data. The reason for combining box and parallel-coordinate plots involves their relative strengths. Box plots work well for showing distribution summaries. The strength of parallel coordinates is their ability to display high-dimensional outliers, individual cases with exceptional values. Details about this class of visualization techniques are given in Section 15.3.
2. *Data Constellations.* A component for visualizing large graphs with thousands of nodes and links. Two tables parametrize Data Constellations, one corresponding to nodes and another to links. Different layout algorithms dynamically position the nodes so that patterns emerge (a visual interpretation of outliers, clusters, etc.).
3. *Data Sheet.* A dynamic scrollable text visualization that bridges the gap between text and graphics. The user can adjust the zoom factor, progressively displaying smaller and smaller fonts, eventually switching to a one-pixel representation. This process is called smashing.
4. *Time Table.* a technique for showing thousands of time-stamped events.
5. *Multiscape.* A landscape visualization that encodes information using 3-D "skyscrapers" on a 2-D landscape.

An example of one of these novel visual representations is given in Figure 15.3, where a large graph is visualized using the Data Constellations technique with one possible graph-layout algorithm.

For most basic visualization techniques that endeavor to show each item in a data set, such as scatter plots or parallel coordinates, a massive number of items will overload the visualization, resulting in a clutter that both causes scalability problems and hinders the user's understanding of its structure and contents. New visualization

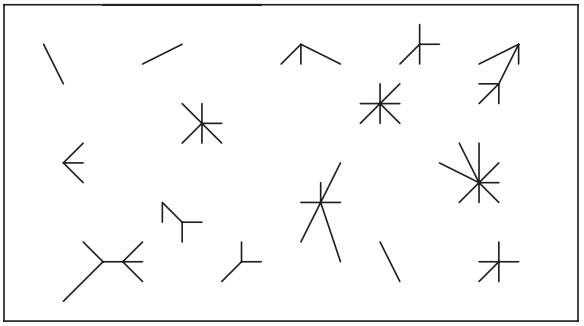

Figure 15.3. Data Constellations as a novel visual metaphor.

techniques have been proposed to overcome data overload problem, and to introduce abstractions that reduce the amount of items to display either in data space or in visual space. The approach is based on coupling aggregation in data space with a corresponding visual representation of the aggregation as a visual entity in the graphical space. This visual aggregate can convey additional information about the underlying contents, such as an average value, minima and maxima, or even its data distribution.

Drawing visual representations of abstractions performed in data space allows for creating simplified versions of visualization while still retaining the general overview. By dynamically changing the abstraction parameters, the user can also retrieve details-on-demand. There are several algorithms to perform data aggregations in a visualization process. For example, given a set of data items, *hierarchical aggregation* is based on iteratively building a tree of aggregates either bottom-up or top-down. Each aggregate item consists of one or more children that are either the original data items (leaves) or aggregate items (nodes). The root of the tree is an aggregate item that represents the entire data set. One of the main visual aggregations for scatter plots involves hierarchical aggregations of data into hulls, as it is represented in Figure 15.4. Hulls are variations and extensions of rectangular boxes as aggregates. They show enhanced displayed dimensions by using 2-D or 3-D convex hulls instead of axis-aligned boxes as a constrained visual metric. Clearly, the benefit of a data aggregate hierarchy and corresponding visual aggregates is that the resulting visualization can be adapted to the requirements of the human user as well as the technical limitations of the visualization platform.

15.3 PARALLEL COORDINATES

Geometric-projection techniques include the parallel coordinate—visualization technique, one of the most frequently used modern visualization tools. The basic idea is to map the k-dimensional space onto the two-display dimensions by using k equidistant axes parallel to one of the display axes. The axes correspond to the dimensions and are linearly scaled from the minimum to the maximum value of the corresponding dimension. Each data item is presented as a polygonal line, intersecting each of the axes at the point that corresponds to the value of the considered dimension.

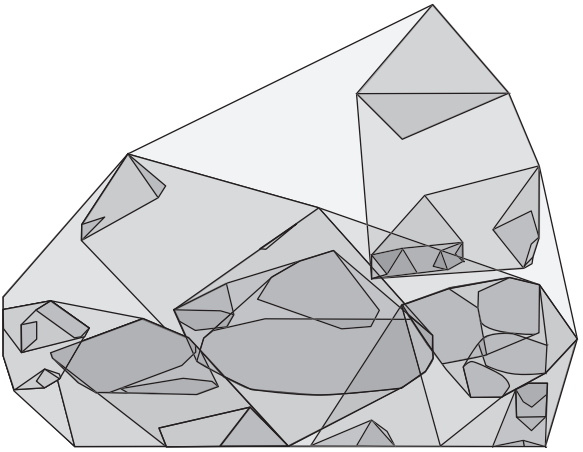

Figure 15.4. Convex hull aggregation [Elmquist 2010].

TABLE 15.1. Database with Six Numeric Attributes

Sample number	Dimensions					
	A	B	C	D	E	F
1	1	5	10	3	3	5
2	3	1	3	1	2	2
3	2	2	1	2	4	2
4	4	2	1	3	1	2

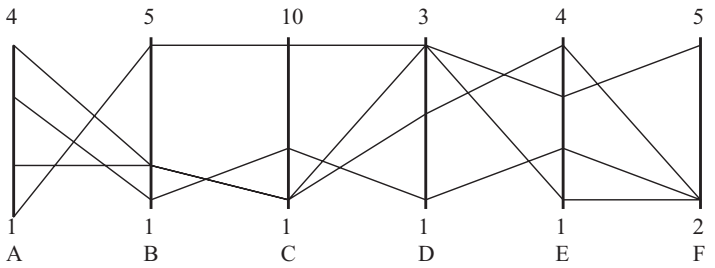

Figure 15.5. Graphical representation of 6-dimesional samples from the database given in Table 15.1 using a parallel coordinate visualization technique.

Suppose that a set of 6-D samples, given in Table 15.1, is a small relational database. To visualize these data, it is necessary to determine the maximum and minimum values for each dimension. If we accept that these values are determined automatically based on a stored database, then graphical representation of data is given on Figure 15.5.

The *anchored-visualization perspective* focuses on displaying data with an arbitrary number of dimensions, for example, between four and 20, using and combining multidimensional-visualization techniques such as weighted Parabox, bubble plots, and parallel coordinates. These methods handle both continuous and categorical data. The reason for combining them involves their relative strengths. Box plots works well for showing distribution summaries. Parallel coordinates' strength is their ability to display high-dimensional outliers, individual cases with exceptional values. Bubble plots are used for categorical data and the size of the circles inside the bubbles shows the number of samples and their respective value. The dimensions are organized along a series of parallel axes, as with parallel-coordinate plots. Lines are drawn between the bubble and the box plots connecting the dimensions of each available sample. Combining these techniques results in a visual component that excels the visual representations created using separate methodologies.

An example of multidimensional anchored visualization, based on a simple and small data set, is given in Table 15.2. The total number of dimensions is five, two of them are categorical and three are numeric. Categorical dimensions are represented by bubble plots (one bubble for every value) and numeric dimensions by boxes. The circle inside the bubbles visually shows the percentage that the given value represents in a database. Lines inside the boxes represent mean value and standard deviation for a given numeric dimension. The resulting representation in Figure 15.6 shows all six 5-D

TABLE 15.2. The Database for Visualization

Sample number	Dimensions				
	A	B	C	D	E
1	Low	Low	2	4	3
2	Medium	Medium	4	2	1
3	High	Medium	7	5	9
4	Medium	Low	1	3	5
5	Low	Low	3	1	2
6	Low	Medium	4	3	2

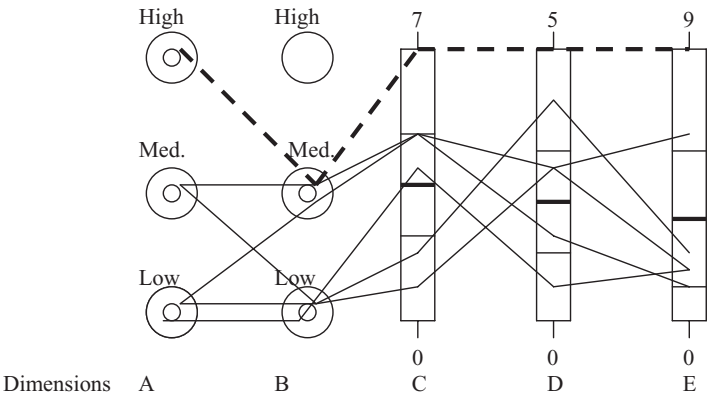

Figure 15.6. Parabox visualization of a database given in Table 15.2.

samples as connecting lines. Although the database given in Table 15.2 is small, still, by using anchored representation, we can see that one sample is an outlier for both numeric and categorical dimensions.

The circular-coordinates method is a simple variation of parallel coordinates, in which the axes radiate from the center of a circle and extend to the perimeter. The line segments are longer on the outer part of the circle where higher data values are typically mapped, whereas inner-dimensional values toward the center of the circle are more cluttered. This visualization is actually a star and glyphs visualization of the data superimposed on one another. Because of the asymmetry of lower (inner) data values from higher ones, certain patterns may be easier to detect with this visualization.

15.4 RADIAL VISUALIZATION

Radial visualization is a technique for representation of multidimensional data where the number of dimensions are significantly greater then three. Data dimensions are laid out as points equally spaced around the perimeter of a circle. For example, in the case of an 8-D space, the distribution of dimensions will be given as in Figure 15.7.

A model of springs is used for point representation. One end of n springs (one spring for each of n dimensions) is attached to n perimeter points. The other end of the springs is attached to a data point. Spring constants can be used to represent values of dimensions for a given point. The spring constant K_i equals the value of the *i*th coordinate of the given n-dimensional point where i = 1, . . . , n. Values for all dimensions are normalized to the interval between 0 and 1. Each data point is then displayed in 2-D under condition that the sum of the spring forces is equal to 0. The radial visualization of a 4-D point P(K_1, K_2, K_3, K_4) with the corresponding spring force is given in Figure 15.8.

Using basic laws from physics, we can establish a relation between coordinates in an n-dimensional space and in 2-D presentation. For our example of 4-D representation given in Figure 15.8, point P is under the influence of four forces, F_1, F_2, F_3, and F_4. Knowing that every one of these forces can be expressed as a product of a spring constant and a distance, or in a vector form

$$F = K \cdot d$$

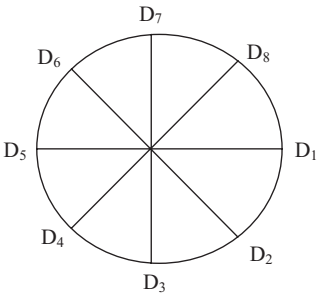

Figure 15.7. Radial visualization for an 8-dimensional space.

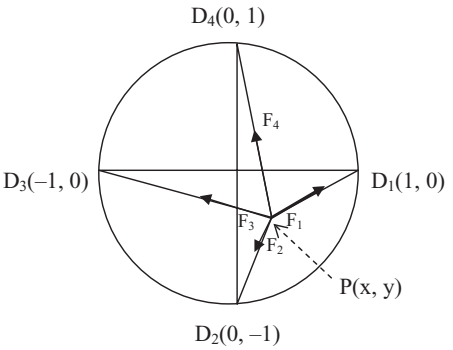

Figure 15.8. Sum of the spring forces for the given point P is equal to 0.

it is possible to calculate this force for a given point. For example, force F_1 in Figure 15.8 is a product of a spring constant K_1 and a distance vector between points P(x, y) and $D_1(1,0)$:

$$F_1 = K_1([x-1]i + yj)$$

The same analysis will give expressions for F_2, F_3, and F_4. Using the basic relation between forces

$$F_1 + F_2 + F_3 + F_4 = 0$$

we will obtain

$$K_1([x-1]i + yj) + K_2(xi + [y+1]j) + K_3([x+1]i + yj) + K_4(xi + [y-1]j) = 0$$

Both the i and j components of the previous vector have to be equal to 0, and therefore:

$$K_1(x-1) + K_2x + K_3(x+1) + K_4x = 0$$
$$K_1y + K_2(y+1) + K_3y + K_4(y-1) = 0$$

or

$$x = (K_1 - K_3)/(K_1 + K_2 + K_3 + K_4)$$
$$y = (K_4 - K_2)/(K_1 + K_2 + K_3 + K_4)$$

These are the basic relations for representing a 4-D point $P^*(K_1,K_2,K_3,K_4)$ in a 2-D space P(x, y) using the radial-visualization technique. Similar procedures may be performed to get transformations for other n-dimensional spaces.

We can analyze the behavior of n-dimensional points after transformation and representation with two dimensions. For example, if all n coordinates have the same value, the data point will lie exactly in the center of the circle. In our 4-D space, if the

initial point is P_1*(0.6, 0.6, 0.6, 0.6), then using relations for x and y its presentation will be P_1(0, 0). If the n-dimensional point is a unit vector for one dimension, then the projected point will lie exactly at the fixed point on the edge of the circle (where the spring for that dimension is fixed). Point P_2*(0, 0, 1, 0) will be represented as P_2(−1, 0). Radial visualization represents a nonlinear transformation of the data, which preserves certain symmetries. This technique emphasizes the relations between dimensional values, not between separate, absolute values. Some additional features of radial visualization include:

1. Points with approximately equal coordinate values will lie close to the center of the representational circle. For example, P_3*(0.5, 0.6, 0.4, 0.5) will have 2-D coordinates P_3(0.05, −0.05).
2. Points that have one or two coordinate values greater than the others lie closer to the origins of those dimensions. For example, P_4*(0.1, 0.8, 0.6, −0.1) will have a 2-D representation P_4(−0.36, −0.64). The point is in a third quadrant closer to D2 and D3, points where the spring is fixed for the second and third dimensions.
3. An n-dimensional line will map to the line or in a special case to the point. For example, points P_5*(0.3, 0.3, 0.3, 0.3), P_6*(0.6, 0.6, 0.6, 0.6), and P_7*(0.9, 0.9, 0.9, 0.9) are on a line in a 4-D space, and all three of them will be transformed into the same 2-D point P_{567}(0, 0).
4. A sphere will map to an ellipse.
5. An n-dimensional plane maps to a bounded polygon.

The *Gradviz method* is a simple extension of a radial visualization that places the dimensional anchors on a rectangular grid instead of the perimeter of a circle. The spring forces work the same way. Dimensional labeling for *Gradviz* is difficult, but the number of dimensions that can be displayed increases significantly in comparison to the Radviz technique. For example, in a typical Radviz display 50 seems to be a reasonable limit to the points around a circle. However, in a grid layout supported by the Gradviz technique you can easily fit 50×50 grid points or dimensions into the same area.

15.5 VISUALIZATION USING SELF-ORGANIZING MAPS (SOMs)

SOM is often seen as a promising technique for exploratory analyses through visualization of high-dimensional data. It visualizes a data structure of a high-dimensional data space usually as a 2-D or 3-D geometrical picture. SOMs are, in effect, a nonlinear form of PCA, and share similar goals to multidimensional scaling. PCA is much faster to compute, but it has the disadvantage, compared with SOMs, of not retaining the topology of the higher dimensional space.

The topology of the data set in its n-dimensional space is captured by the SOM and reflected in the ordering of its output nodes. This is an important feature of the

SOM that allows the data to be projected onto a lower dimension space while roughly preserving the order of the data in its original space. Resultant SOMs are then visualized using graphical representations. SOM algorithm may use different data-visualization techniques including a cell or U-matrix visualization (a distance matrix visualization), projections (mesh visualization), visualization of component planes (in a multiple-linked view), and 2-D and 3-D surface plot of distance matrices. These representations use visual variables (size, value, texture, color, shape, orientation) added to the position property of the map elements. This allows exploration of relationships between samples. A coordinate system enables to determine distance and direction, from which other relationships (size, shape, density, arrangement, etc.) may be derived. Multiple levels of detail allow exploration at various scales, creating the potential for hierarchical grouping of items, regionalization, and other types of generalizations. Graphical representations in SOMs are used to represent uncovered structure and patterns that may be hidden in the data set and to support understanding and knowledge construction. An illustrative example is given in Figure 15.9 where linear or nonlinear relationships are detected by the SOM.

For years there has been visualization of primary numeric data using pie charts, colored graphs, graphs over time, multidimensional analysis, Pareto charts, and so forth. The counterpart to numeric data is unstructured, textual data. Textual data are found in many places, but nowhere more prominently than on the Web. Unstructured electronic data include emails, email attachments, PDF files, spread sheets, PowerPoint files, text files, and document files. In this new environment, the end user faces massive amounts,

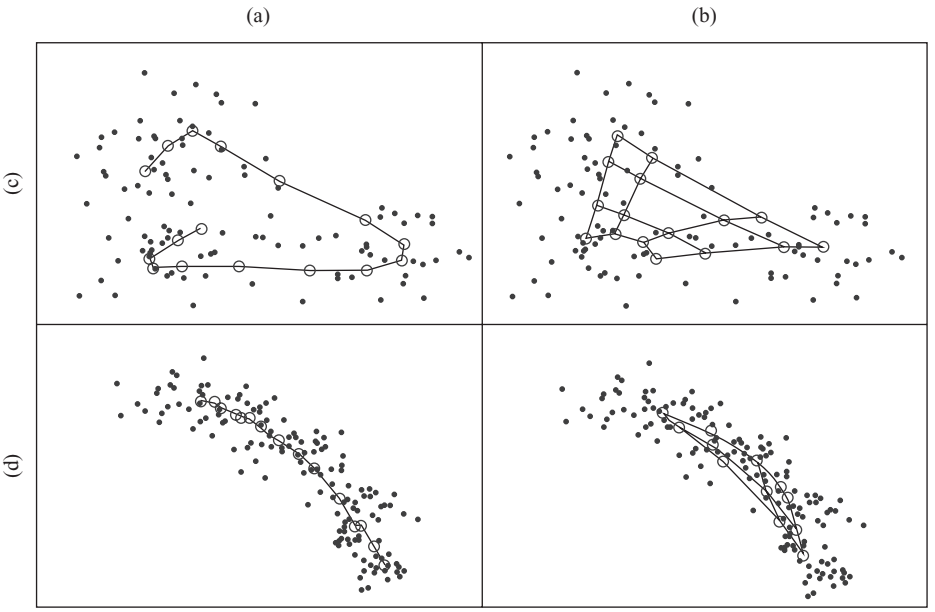

Figure 15.9. Output maps generated by the SOM detect relationships in data. (a) 1-D image map; (b) 2-D image map; (c) nonlinear relationship; (d) linear relationship.

often millions, of unstructured documents. The end user cannot read them all, and especially, there is no way he/she could manually organize or summarize them. Unstructured data run the less formal part of the organization, while structured data run the formal part of the organization. It is a good assumption, confirmed in many real-world applications, that as many business decisions are made in the unstructured environment as in the structured environment.

The SOM is one efficient solution for the problems of unstructured visualization of documents and unstructured data. With a properly constructed SOM, you can analyze literally millions of unstructured documents that can be merged into a single SOM. The SOM deals not only with individual unstructured documents but relationships between documents as well. The SOM may show text that is correlated to other text. For example, in the medical field, working with medical patient records, this ability to correlate is very attractive. The SOM also allows the analyst to see the larger picture as well as drilling down to the detailed picture. The SOM goes down to the individual stemmed-text level, and that is as accurate as textual processing can become. All these characteristics have resulted in the growing popularity of SOM visualizations in order to assist visual inspection of complex high-dimensional data. For the end user the flexibility of the SOM algorithm is defined through a number of parameters. For appropriate configuration of the network, and tuning the visualization output, user-defined parameters include grid dimensions (2-D, 3-D), grid shape (rectangle, hexagon), number of output nodes, neighborhood function, heighborhood size, learning-rate function, initial weights in the network, way of learning and number of iterations, and order of input samples.

15.6 VISUALIZATION SYSTEMS FOR DATA MINING

Many organizations, particularly within the business community, have made significant investments in collecting, storing, and converting business information into results that can be used. Unfortunately, typical implementations of business "intelligence software" have proven to be too complex for most users except for their core reporting and charting capabilities. Users' demands for multidimensional analysis, finer data granularity, and multiple-data sources, simultaneously, all at Internet speed, require too much specialist intervention for broad utilization. The result is a report explosion in which literally hundreds of predefined reports are generated and pushed throughout the organization. Every report produces another. Presentations get more complex. Data are exploding. The best opportunities and the most important decisions are often the hardest to see. This is in direct conflict with the needs of frontline decision makers and knowledge workers who are demanding to be included in the analytical process.

Presenting information visually, in an environment that encourages the exploration of linked events, leads to deeper insights and more results that can be acted upon. Over the past decade, research on information visualization has focused on developing specific visualization techniques. An essential task for the next period is to integrate these techniques into a larger system that supports work with information in an interactive way, through the three basic components: *foraging the data*, *thinking about data*, and *acting on data*.

The vision of a visual data-mining system stems from the following principles: simplicity, visibility, user autonomy, reliability, reusability, availability, and security. A visual data-mining system must be syntactically simple to be useful. Simple does not mean trivial or non-powerful. Simple to learn means use of intuitive and friendly input mechanisms as well as instinctive and easy-to-interpret output knowledge. Simple to apply means an effective discourse between humans and information. Simple to retrieve or recall means a customized data structure that facilitates fast and reliable searches. Simple to execute means a minimum number of steps needed to achieve the results. In short, simple means the smallest, functionally sufficient system possible.

A genuinely visual data-mining system must not impose knowledge on its users, but instead guide them through the mining process to draw conclusions. Users should study the visual abstractions and gain insight instead of accepting an automated decision. A key capability in visual analysis, called visibility, is the ability to focus on particular regions of interest. There are two aspects of visibility: excluding and restoring data. The exclude process eliminates the unwanted data items from the display so that only the selected set is visible. The restore process brings all data back, making them visible again.

A reliable data-mining system must provide for estimated error or accuracy of the projected information in each step of the mining process. This error information can compensate for the deficiency that an imprecise analysis of data visualization can cause. A reusable, visual, data-mining system must be adaptable to a variety of environments to reduce the customization effort, provide assured performance, and improve system portability. A practical, visual, data-mining system must be generally and widely available. The quest for new knowledge or deeper insights into existing knowledge cannot be planned. It requires that the knowledge received from one domain adapt to another domain through physical means or electronic connections. A complete, visual, data-mining system must include security measures to protect the data, the newly discovered knowledge, and the user's identity because of various social issues.

Through data visualization we want to understand or get an overview of the whole or a part of the n-dimensional data, analyzing also some specific cases. Visualization of multidimensional data helps decision makers to

1. slice information into multiple dimensions and present information at various levels of granularity,
2. view trends and develop historical tracers to show operations over time,
3. produce pointers to synergies across multiple dimensions,
4. provide exception analysis and identify isolated (needle in the haystack) opportunities,
5. monitor adversarial capabilities and developments,
6. create indicators of duplicative efforts,
7. conduct What-If Analysis and Cross-Analysis of variables in a data set.

Visualization tools transform raw experimental or simulated data into a form suitable for human understanding. Representations can take on many different forms,

depending on the nature of the original data and the information that is to be extracted. However, the visualization process that should be supported by modern, visualization-software tools can generally be subdivided into three main stages: data preprocessing, visualization mapping, and rendering. Through these three steps the tool has to answer the questions: What should be shown in a plot? How should one work with individual plots? How should multiple plots be organized?

Data preprocessing involves such diverse operations as interpolating irregular data, filtering and smoothing raw data, and deriving functions for measured or simulated quantities. Visualization mapping is the most crucial stage of the process, involving design and adequate representation of the filtered data, which efficiently conveys the relevant and meaningful information. Finally, the representation is often rendered to communicate information to the human user.

Data visualization is essential for understanding the concept of multidimensional spaces. It allows the user to explore the data in different ways and at different levels of abstraction to find the right level of details. Therefore, techniques are most useful if they are highly interactive, permit direct manipulation, and include a rapid response time. The analyst must be able to navigate the data, change its grain (resolution), and alter its representation (symbols, colors, etc.).

Broadly speaking, the problems addressed by current information-visualization tools and requirements for a new generation fall into the following classes:

1. *Presentation Graphics.* These generally consist of bars, pies, and line charts that are easily populated with static data and drop into printed reports or presentations. The next generation of presentation graphics enriches the static displays with a 3-D or projected n-dimensional information landscape. The user can then navigate through the landscape and animate it to display time-oriented information.
2. *Visual Interfaces for Information Access.* They are focused on enabling users to navigate through complex information spaces to locate and retrieve information. Supported user tasks involve searching, backtracking, and history logging. User-interface techniques attempt to preserve user-context and support smooth transitions between locations.
3. *Full Visual Discovery and Analysis.* These systems combine the insights communicated by presentation graphics with an ability to probe, drill down, filter, and manipulate the display to answer the "why" question as well as the "what" question. The difference between answering a "what" and a "why" question involves an interactive operation. Therefore, in addition to the visualization technique, effective data exploration requires using some *interaction* and *distortion* techniques. The *interaction techniques* let the user directly interact with the visualization. Examples of interaction techniques include interactive mapping, projection, filtering, zooming, and interactive linking and brushing. These techniques allow dynamic changes in the visualizations according to the exploration objectives, but they also make it possible to relate and combine multiple, independent visualizations. Note that connecting multiple visualizations by linking and brushing, for example, provides more information than

considering the component visualizations independently. The *distortion techniques* help in the interactive exploration process by providing a means for focusing while preserving an overview of the data. Distortion techniques show portions of the data with a high level of detail while other parts are shown with a much lower level of detail.

Three tasks are fundamental to data exploration with these new visualization tools:

1. *Finding Gestalt.* Local and global linearities and nonlinearities, discontinuities, clusters, outliers, unusual groups, and so on are examples of gestalt features that can be of interest. Focusing through individual views is the basic requirement to obtain a qualitative exploration of data using visualization. Focusing determines what gestalt of the data is seen. The meaning of focusing depends very much on the type of visualization technique chosen.
2. *Posing Queries.* This is a natural task after the initial gestalt features have been found, and the user requires query identification and characterization technique. Queries can concern individual cases as well as subsets of cases. The goal is essentially to find intelligible parts of the data. In graphical data analysis it is natural to pose queries graphically. For example, familiar brushing techniques such as coloring or otherwise highlighting a subset of data means issuing a query about this subset. It is desirable that the view where the query is posed and the view that present the response are linked. Ideally, responses to queries should be instantaneous.
3. *Making Comparisons.* Two types of comparisons are frequently made in practice. The first one is a comparison of variables or projections and the second one is a comparison of subsets of data. In the first case, one compares views "from different angles"; in the second, comparison is based on views "of different slices" of the data. In either case, it is likely that a large number of plots are generated, and therefore it is a challenge to organize the plots in such a way that meaningful comparisons are possible.

Visualization has been used routinely in data mining as a presentation tool to generate initial views, navigate data with complicated structures, and convey the results of an analysis. Generally, the analytical methods themselves do not involve visualization. The loosely coupled relationships between visualization and analytical data-mining techniques represent the majority of today's state-of-the-art in visual data mining. The process-sandwich strategy, which interlaces analytical processes with graphical visualization, penalizes both procedures with the other's deficiencies and limitations. For example, because an analytical process cannot analyze multimedia data, we have to give up the strength of visualization to study movies and music in a visual data-mining environment. A stronger strategy lies in tightly coupling the visualization and analytical processes into one data-mining tool. Letting human visualization participate in the decision making in analytical processes remains a major challenge. Certain mathematical steps within an analytical procedure may be substituted by human decisions based on visualization to allow the same procedure to analyze a broader scope of information.

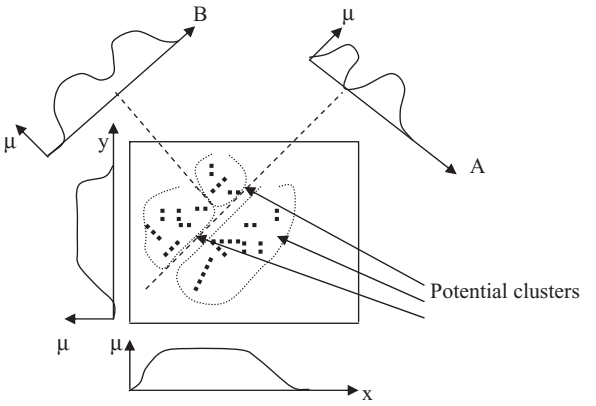

Figure 15.10. An example of the need for general projections, which are not parallel to axes, to improve clustering process.

Visualization supports humans in dealing with decisions that can no longer be automated.

For example, visualization techniques can be used for efficient process of "visual clustering." The algorithm is based on finding a set of projections $P = [P_1, P_2, \ldots, P_k]$ useful for separating the initial data into clusters. Each projection represents the histogram information of the point density in the projected space. The most important information about a projection is whether it contains well-separated clusters. Note that well-separated clusters in one projection could result from more than one cluster in the original space. Figure 15.10 shows an illustration of these projections. You can see that the axes' parallel projections do not preserve well the information necessary for clustering. Additional projections A and B, in Figure 15.10, define three clusters in the initial data set.

Visual techniques that preserve some characteristics of the data set can be invaluable for obtaining good separators in a clustering process. In contrast to dimension-reduction approaches such as PCAs, this visual approach does not require that a single projection preserve all clusters. In the projections, some clusters may overlap and therefore not be distinguishable, such as projection A in Figure 15.10. The algorithm only needs projections that separate the data set into at least two subsets without dividing any clusters. The subsets may then be refined using other projections and possibly partitioned further based on separators in other projections. Based on the visual representation of the projections, it is possible to find clusters with unexpected characteristics (shapes, dependencies) that would be very difficult or impossible to find by tuning the parameter settings of automatic-clustering algorithms.

In general, model visualization and exploratory data analysis (EDA) are data-mining tasks in which visualization techniques have played a major role. Model visualization is the process of using visual techniques to make the discovered knowledge understandable and interpretable by humans. Techniques range from simple scatter plots and histograms to sophisticated multidimensional visualizations and animations.

These visualization techniques are being used not only to convey mining results more understandable to end users, but also to help them understand how the algorithm works. EDA, on the other hand, is the interactive exploration of usually graphical representations of a data set without heavy dependence on preconceived assumptions and models, thus attempting to identify interesting and previously unknown patterns. Visual data-exploration techniques are designed to take advantage of the powerful visual capabilities of human beings. They can support users in formulating hypotheses about the data that may be useful in further stages of the mining process.

15.7 REVIEW QUESTIONS AND PROBLEMS

1. Explain the power of n-dimensional visualization as a data-mining technique. What are the phases of data mining supported by data visualization?
2. What are fundamental experiences in human perception we would build into effective visualization tools?
3. Discuss the differences between scientific visualization and information visualization.
4. The following is the data set X:

X:	Year	A	B
	1996	7	100
	1997	5	150
	1998	7	120
	1999	9	150
	2000	5	130
	2001	7	150

   Although the following visualization techniques are not explained with enough details in this book, use your knowledge from earlier studies of statistics and other courses to create 2-D presentations.

   (a) Show a bar chart for the variable A.
   (b) Show a histogram for the variable B.
   (c) Show a line chart for the variable B.
   (d) Show a pie chart for the variable A.
   (e) Show a scatter plot for A and B variables.

5. Explain the concept of a data cube and where it is used for visualization of large data sets.
6. Use examples to discuss the differences between icon-based and pixel-oriented visualization techniques.
7. Given 7-D samples

x_1	x_2	x_3	x_4	x_5	x_6	x_7
A	1	25	7	T	1	5
B	3	27	3	T	2	9
A	5	29	5	T	1	7
A	2	21	9	F	3	2
B	5	30	7	F	1	7

(a) make a graphical representation of samples using the parallel-coordinates technique;
(b) are there any outliers in the given data set?

8. Derive formulas for radial visualization of

(a) 3-D samples
(b) 8-D samples
(c) using the formulas derived in (a) represent samples (2, 8, 3) and (8, 0, 0).
(d) using the formulas derived in (b) represent samples (2, 8, 3, 0, 7, 0, 0, 0) and (8, 8, 0, 0, 0, 0, 0, 0).

9. Implement a software tool supporting a radial-visualization technique.

10. Explain the requirements for full visual discovery in advanced visualization tools.

11. Search the Web to find the basic characteristics of publicly available or commercial software tools for visualization of n-dimensional samples. Document the results of your search.

15.8 REFERENCES FOR FURTHER STUDY

Draper, G. M., L. Y. Livnat, R. F. Riesenfeld, A Survey of Radial Methods for Information Visualization, *IEEE Transaction on Visualization and Computer Graphics*, Vol. 15, No. 5, 2009, pp. 759–776.

Radial visualization, or the practice of displaying data in a circular or elliptical pattern, is an increasingly common technique in information visualization research. In spite of its prevalence, little work has been done to study this visualization paradigm as a methodology in its own right. We provide a historical review of radial visualization, tracing it to its roots in centuries-old statistical graphics. We then identify the types of problem domains to which modern radial visualization techniques have been applied. A taxonomy for radial visualization is proposed in the form of seven design patterns encompassing nearly all recent works in this area. From an analysis of these patterns, we distill a series of design considerations that system builders can use to create new visualizations that address aspects of the design space that have not yet been explored. It is hoped that our taxonomy will provide a framework for facilitating discourse among researchers and stimulate the development of additional theories and systems involving radial visualization as a distinct design metaphor.

Fayyad, V., G. G. Grinstein, A. Wierse, *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, San Francisco, CA, 2002.

Leading researchers from the fields of data mining, data visualization, and statistics present findings organized around topics introduced in two recent international knowledge-discovery

and data-mining workshops. The book introduces the concepts and components of visualization, details current efforts to include visualization and user interaction in data mining, and explores the potential for further synthesis of data-mining algorithms and data-visualization techniques.

Ferreira de Oliveira, M. C., H. Levkowitz, From Visual Data Exploration to Visual Data Mining: A Survey, *IEEE Transactions On Visualization And Computer Graphics*, Vol. 9, No. 3, 2003, pp. 378–394.

The authors survey work on the different uses of graphical mapping and interaction techniques for visual data mining of large data sets represented as table data. Basic terminology related to data mining, data sets, and visualization is introduced. Previous work on information visualization is reviewed in light of different categorizations of techniques and systems. The role of interaction techniques is discussed, in addition to work addressing the question of selecting and evaluating visualization techniques. We review some representative work on the use of IVT in the context of mining data. This includes both visual-data exploration and visually expressing the outcome of specific mining algorithms. We also review recent innovative approaches that attempt to integrate visualization into the DM/KDD process, using it to enhance user interaction and comprehension.

Gallaghar, R. S., *Computer Visualization: Graphics Techniques for Scientific and Engineering Analysis*, CRC Press, Boca Raton, 1995.

The book is a complete reference book on computer-graphic techniques for scientific and engineering visualization. It explains the basic methods applied in different fields to support an understanding of complex, volumetric, multidimensional, and time-dependent data. The practical computational aspects of visualization such as user interface, database architecture, and interaction with a model are also analyzed.

Spence, R., *Information Visualization*, Addison Wesley, Harlow, England, 2001.

This is the first fully integrated book on the emerging discipline of information visualization. Its emphasis is on real-world examples and applications of computer-generated interactive information visualization. The author also explains how these methods for visualizing information support rapid learning and accurate decision making.

Tufte, E. R., *Beautiful Evidence*, 2nd edition, Graphic Press, LLC, Cheshire, CT, 2007.

Beautiful Evidence is a masterpiece from a pioneer in the field of data visualization. It is not often an iconoclast comes along, trashes the old ways, and replaces them with an irresistible new interpretation. By teasing out the sublime from the seemingly mundane world of charts, graphs, and tables, Tufte has proven to a generation of graphic designers that great thinking begets great presentation. In *Beautiful Evidence*, his fourth work on analytical design, Tufte digs more deeply into art and science to reveal very old connections between truth and beauty—all the way from Galileo to Google.