43 Visual Data-Mining Techniques*

DANIEL A. KEIM, MIKE SIPS
University of Konstanz, Germany
MIHAEL ANKERST
The Boeing Company, USA

43.1 Introduction

Never before in history have data been generated at such high volumes as it is today. Exploring and analyzing the vast volumes of data has become increasingly difficult. Information visualization and visual data mining can help to deal with the flood of information. The advantage of visual data exploration is that the user is directly involved in the data-mining process. There are a large number of information visualization techniques that have been developed over the last few years to support the exploration of large datasets. In this chapter, we provide an overview of information visualization and visual data-mining techniques and illustrate them using a few examples.

The progress made in hardware technology allows today's computer systems to store very large amounts of data. Researchers from the University of Berkeley estimate that every year about 1 exabyte (1 million terabytes) of data is generated, of which a large portion is available in digital form. This means that in the next three years more data will be generated than in all of human history to date. The data is often automatically recorded via sensors and monitoring systems. Even simple transactions of everyday life, such as paying by credit card or using the telephone, are typically recorded by computers. Usually many parameters are recorded, resulting in data with high dimensionality. The data is collected because people believe that it is a potential source of valuable information, providing a competitive advantage (at some point). Finding the valuable information hidden in the data, however, is a difficult task. With today's data-management systems, it is possible to view only small portions of the data. If the data is presented textually, the amount of data that can be displayed is in the range of some one hundred data items, but this is like a drop in the ocean when you are dealing with datasets containing millions of data items. Having no possibility to adequately explore the large amounts of data that have been collected because of their potential usefulness, the data becomes useless and the databases become data 'dumps.' Information visualization focuses on datasets lacking inherent 2D or 3D semantics and therefore also lacking a standard mapping of the abstract data onto the physical screen space. There are a number of well known techniques for visualizing such datasets, such as x-y plots, line plots, and histograms. These techniques are useful for data exploration but are limited to relatively small and low-dimensional datasets. In the last few years, a large number of novel information visualization techniques have been developed, allowing visualizations of multidimensional datasets without inherent 2D or 3D semantics. Nice overviews of the approaches can be found in a number of recent books [8,38,38,28]. The techniques can be classified based on three

*A earlier version of this paper with focus on visualization techniques and their classification (see section I) has been published in [21]

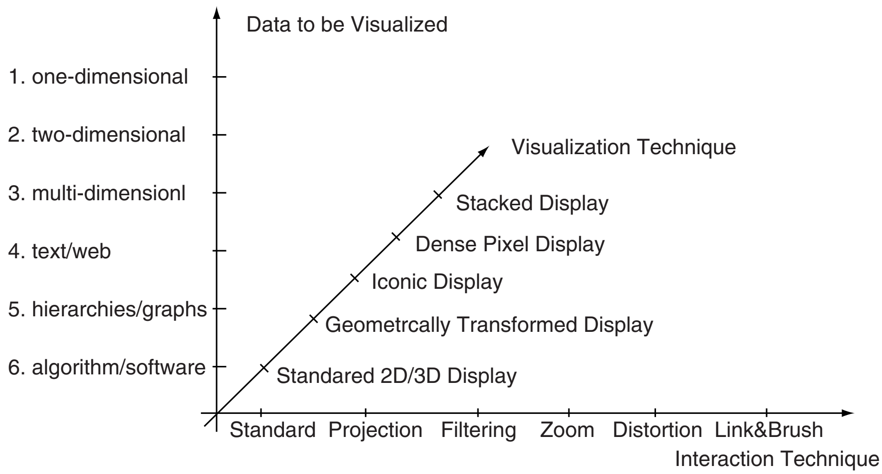


Figure 43.1 Classification of Information Visualization Techniques.

criteria [20] (Fig. 43.1): The data to be visualized, the visualization technique, and the interaction technique used.

The *data type to be visualized* [32] may be 1D data, such as temporal (time-series) data; 2D data, such as geographical maps; multidimensional data, such as relational tables text, hypertext news articles, and web documents; or hierarchies and graphs, such as telephone calls and Web documents, algorithms, and software.

The *visualization technique* used may be classified as standard 2D/3D displays, such as bar charts and x-y plots, geometrically transformed displays, such as hyperbolic plane [36] (Fig. 43.2a) and parallel coordinates [18], icon-based displays, such as chernoff faces [9] and stick figures [24,23] (Fig. 43.2c), dense pixel displays, such as the recursive pattern [4] (Fig. 43.2b) and circle segments [5], stacked displays, such as treemaps [31,19] (Fig. 43.2d) and dimensional stacking [37]. The third dimension of the classification is the *interaction technique* used. Interaction techniques allow users to directly navigate and modify the visualizations, as well as select subsets of the data for further operations. Examples include dynamic projection, interactive filtering, interactive zooming, interactive distortion, interactive linking, and brushing. Note that the three dimensions of our classification—data type to be visualized, visualization technique, and interaction technique—can be assumed to be orthogonal. Orthogonality means that any of the visualization techniques may be used in conjunction with any of the interaction techniques for any data type. Note also that a specific system may be designed to support different data types and that it may use a combination of visualization and interaction techniques. More details can be found in Keim and Ward [21].

43.2 Methodology of Visual Data Mining

The data analyst typically specifies first some parameters to restrict the search space; data mining is then performed automatically by an algorithm, and finally the patterns found by the automatic data-mining algorithm are presented to the data analyst on the screen. For data mining to be effective, it is important to include the human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today's computers. Since there is a huge

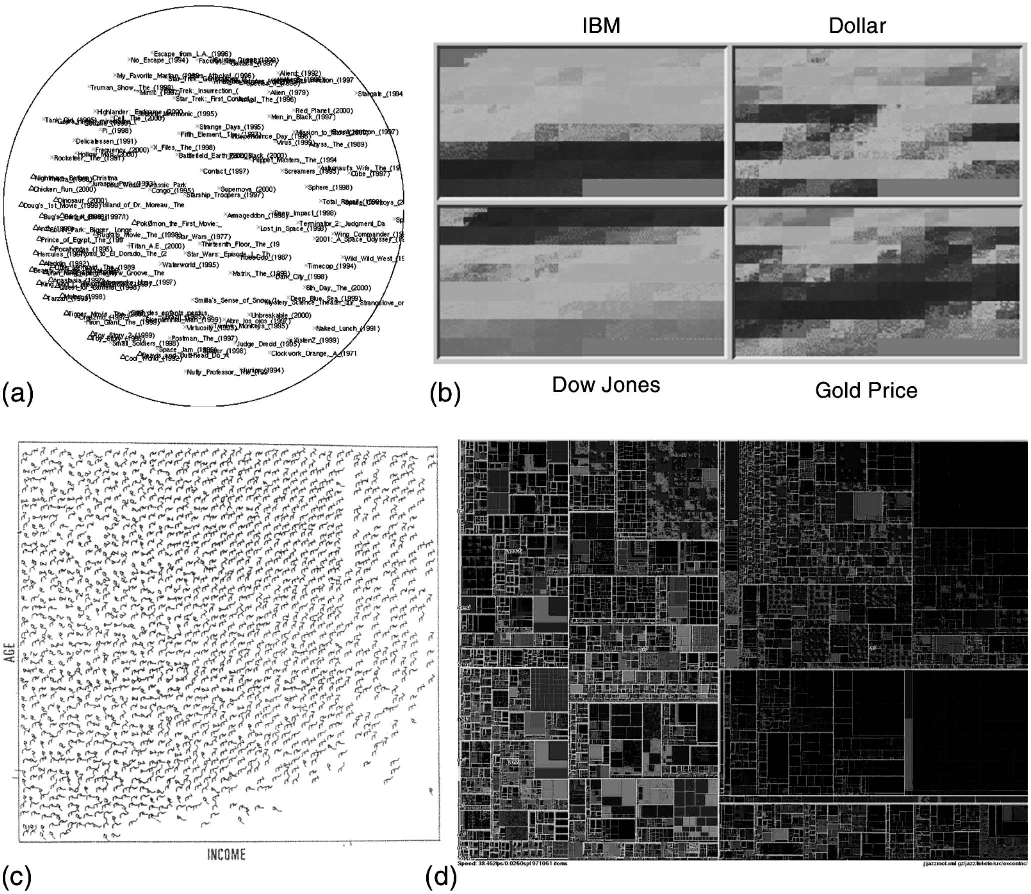


Figure 43.2 Some popular information visualization techniques. (a) Geometrically transformed displays: Interactive visualization of high-dimensional data using the hyperbolic plane [36]—Genre separation in movie space (red "x" marks science fiction, black "Δ" marks animation and green "+" movies belonging to both genres) © ACM (b) Dense pixel displays: Recursive Pattern [4]—based on a generic back-and-forth recursive arrangement schema to represent each data value as a colored pixel and each attribute in separate sub-windows (example visualization shows the stock prices for Dow Jones, Gold, IBM and US-Dollar are depicted for almost seven consecutive years, seven vertical bars correspond to the seven years (level (3)-patterns) and the subdivision of the bars to the 12 month within each year (level (2)-patterns), the coloring maps high attribute values (stock prices) to light colors and low attributes values (stock prices) to dark colors) (c) Iconic displays: Stick Figures [24,23]—visualization of multidimensional data using properties of angle and/or length of the limbs (US Census Data Median Household Income and Age of Householder) (d) Stacked displays: TreeMaps [31,9]—splitting the screen into rectangles in alternating horizontal and vertical directions in each level (example visualization shows a hierarchical file system of a large hard disk)

amount of patterns generated by an automatic data-mining algorithm in textual form it is almost impossible for the human to interprete and evaluate the pattern in detail and extract interesting knowledge and general characteristics. Visual data mining aims at integrating the human in the data-mining process, and applying human perceptual abilities to the analysis of large datasets available in today's computer systems. Presenting data in an interactive, graphical form often fosters new insights, encouraging the formation and validation of new hypotheses to the end of better problem-solving and gaining deeper domain knowledge.

Visual data exploration usually follows a three-step process: *Overview* first, *zoom and filter*, and then *details-on-demand* (which has been called the Information Seeking Mantra [32]). First, the data analyst needs to get an overview of the data. In the overview, the data analyst identifies interesting patterns or groups in the data and focuses on one or more of them. For analyzing the patterns, the data analyst needs to drill down and access details of the data. Visualization technology may be used for all three steps of the data exploration process. Visualization techniques are useful for showing an overview of the data, allowing the data analyst to identify interesting subsets. In this step, it is important to keep the overview visualization while focusing on the subset using another visualization technique. An alternative is to distort the overview visualization in order to focus on the interesting subsets. This can be performed by dedicating a larger percentage of the display to the interesting subsets while decreasing screen utilization for uninteresting data. To further explore the interesting subsets, the data analyst needs a drill-down capability in order to observe the details about the data. Note that visualization technology not only provides the base visualization techniques for all three steps but also bridges the gaps between the steps. Visual data mining can be seen as a hypothesis-generation process; the visualizations of the data allow the data analyst to gain insight into the data and come up with new hypotheses. The verification of the hypotheses can also be done via data visualization, but may also be accomplished by automatic techniques from statistics, pattern recognition, or machine learning. As a result, visual data mining usually allows faster data exploration and often provides better results, especially in cases where automatic data-mining algorithms fail. In addition, visual data exploration techniques provide a much higher degree of user satisfaction and confidence in the findings of the exploration. This fact leads to a high demand for visual exploration techniques and makes them indispensable in conjunction with automatic exploration techniques.

Visual data mining is based on an automatic part, the data-mining algorithm, and an interactive part, the visualization technique. There are three common approaches to integrate the human in the data exploration process to realize different kinds of visual data mining approaches (Fig. 43.3):

- **Preceding Visualization (PV)**: Data is visualized in some visual form before running a data-mining algorithm. By interaction with the raw data the data analyst has full control over the analysis in the search space. Interesting patterns are discovered by exploring the data.
- **Subsequent Visualization (SV)**: An automatic data-mining algorithm performs the data-mining task by extracting patterns from a given dataset. These patterns are visualized to make them interpretable for the data analyst. Subsequent visualizations enable the data analyst to specify feedbacks. Based on the visualization, the data analyst may want to return to the data-mining algorithm and use different input parameters to obtain better results.
- **Tightly Integrated Visualization (TIV)**: An automatic data-mining algorithm performs an analysis of the data but does not produce the final results. A visualization technique is used to present the intermediate results of the data exploration process. The combination of some automatic data-mining algorithms and visualization techniques enables specified user feedback for the next data-mining run. Then, the data analyst identifies the interesting patterns in the visualization of the intermediate results based on his domain knowledge. A motivation of this approach is to achieve independence of the data-mining algorithms from the application. A given automatic data-mining algorithm can be very useful in one domain but may have drawbacks in some other domain. Since there is no automatic data-mining algorithm (with one parameter setting) suitable for all application domains, tightly integrated

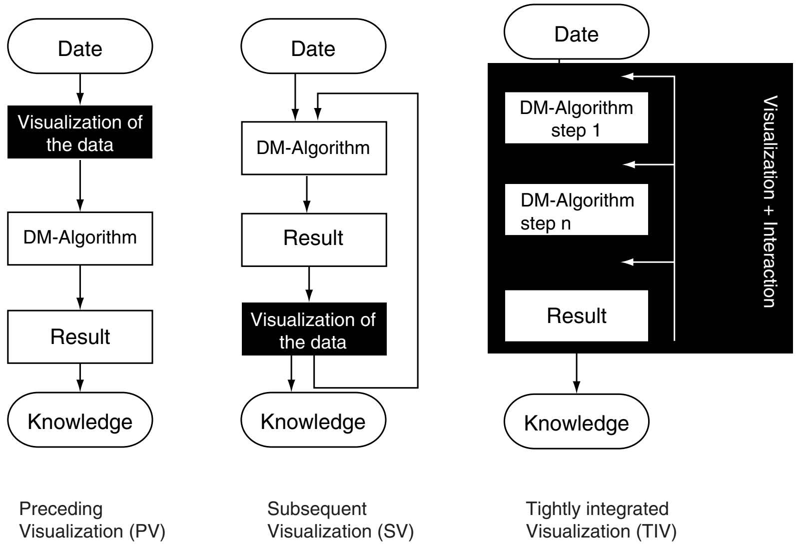


Figure 43.3 Overview of different approaches of human involvement.

visualization leads to a better understanding of the data and the extracted patterns.

In addition to the direct involvement of the human, the main advantages of visual data exploration over automatic data mining techniques are the following:

- Visual data exploration can easily deal with highly nonhomogeneous and noisy data.
- Visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.
- Visualization can provide a qualitative overview of the data, allowing data phenomena to be isolated for further quantitative analysis.

Visual data-mining techniques have proven to be of high value in exploratory data analysis and have a high potential for exploring large databases. Visual data exploration is especially useful when little is known about the data and the exploration goals are vague. Since the data analyst is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary. In the next sections, we show that the integration of the human in the data-mining process and applying human perceptual abilities to the analysis of large datasets can help to provide more effective results in important data-mining application domains, such as in the mining for association rules, clustering, classification, and text retrieval.

43.3 Association Rules

The goal of association rule generation is to find interesting patterns and trends in transaction databases. Association rules are statistical relations between two or more items in the dataset. In a supermarket basket application, associations express the relations between items that are bought together. It is, for example, interesting if we find out that in 70% of the cases when people buy bread, they also buy milk. Association rules tell us that the presence of some items in a transaction imply the presence of other items in the same transaction with a certain probability, called confidence. A second important parameter is the support of an association rule, which is defined as the percentage of transactions in which the items co-occur.

Let $I = \{i_1, \ldots i_n\}$ be a set of items and let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \in I$, $Y \in I$, and X, $Y \neq \emptyset$. The confidence c is defined as the percentage of transactions that contain Y, given X. The support is the percentage of transactions that contain both X and Y. For given support and confidence levels, there are efficient algorithms to determine all association rules [1]. A problem, however, is that the resulting set of association rules is usually very large, especially for low support and confidence levels. Using higher support and confidence levels may not be effective, since useful rules may then be overlooked.

Visualization techniques have been used to overcome this problem and to allow an interactive selection of good support and confidence levels. Fig. 43.4 shows SGI MineSets *Rule Visualizer* [17], which maps the left- and right-hand sides of the rules to the x- and y-axes of the plot and shows the confidence as the height of the bars and the support as the height of the discs. The color of the bars shows the interestingness of the rule. Using the visualization, the user is able to see groups of related rules and the impact of different confidence and support levels. The number of rules that can be visualized, however, is limited, and the visualization does not support combinations of items on the left- or right-hand side of the association rules. Fig. 43.5 shows two alternative visualizations called mosaic and double-decker plots [15]. The basic idea is to partition a rectangle on the y-axis according to one attribute and make the regions proportional to the sum of the corresponding data values. Compared to bar charts, mosaic plots use the height of the bars instead of the width to show the parameter value. Then each resulting area is split in the same way according to a second attribute. The coloring reflects the percentage of data items that fulfill a third attribute. The visualization shows the support and confidence values of all rules of the form $X_1X_2 \Rightarrow Y$. Mosaic plots are restricted to two attributes on the left side of the association rule. Double-decker plots can be used to show more than two attributes on the left side. The idea is to show a hierarchy of attributes on the bottom (Heineken, Coke, chicken, in the example shown in Fig. 43.5) corresponding to the left-hand side of the association rules; the

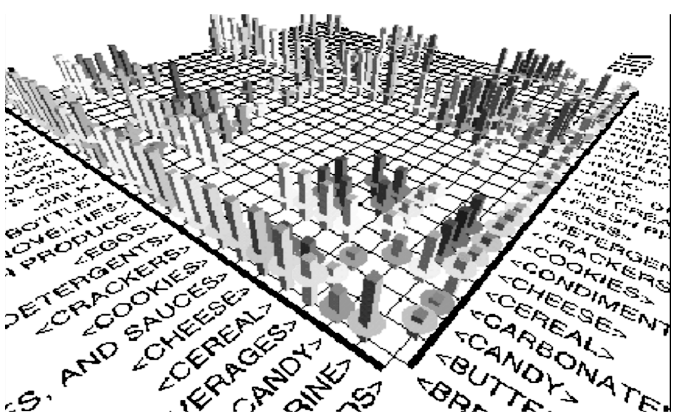


Figure 43.4. MineSet's Association Rule Visualizer [17] maps the left- and right-hand sides of the rules to the x- and y-axes of the plot and shows the confidence as the height of the bars and the support as the height of the discs; color of the bars shows the interestingness of the rule (example visualization shows market basket data for customer buying patterns) ©SGI

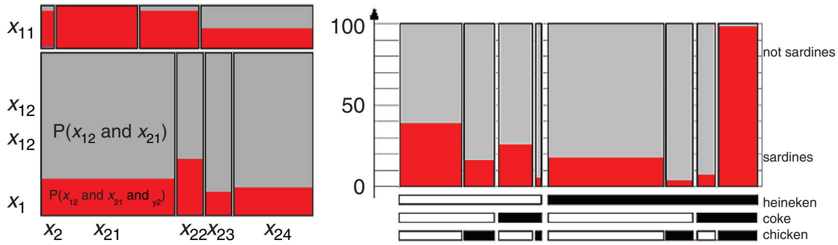


Figure 43.5 Association Rule Visualization [15] partitions a rectangle on the y-axis according to one attribute and makes the regions proportional to the sum of the corresponding data values. © ACM (a) Mosaic Plot: 2D mosaic plot of attributes Ax_1 and Ax_2; high-lighting show up in the mosaic plot as a third dimension (b) Double-Decker Plot: example visualization shows a hierarchy of supermarket basket items: Heineken, Coke, chicken and sardines.

bars on the top correspond to the number of items in the corresponding subset of the database and therefore visualize the support of the rule. The colored areas in the bars correspond to the percentage of data transactions that contain an additional item (sardines, in Fig. 43.5) and therefore correspond to the support. Other approaches to association rule visualization include graphs with nodes corresponding to items and arrows corresponding to implications as used in DBMiner [16] and association matrix visualizations to cluster-related rules [12].

43.4 Classification

Classification is the process of developing a classification model based on a training dataset with known class labels. To construct the classification model, the attributes of the training dataset are analyzed and an accurate description or model of the classes based on the attributes available in the dataset is developed. The class descriptions are used then to classify data for which the class labels are unknown. Classification is sometimes also called *supervised learning* because the training set is used to teach the system how to classify the data. There are many algorithms for solving classification talks. The most popular approaches are algorithms that inductively construct decision trees. Examples are ID3 [25], CART [7], ID5 [34,35], C4.5 [26], SLIQ [22], and SPRINT [30]. In addition, there are approaches that use neural networks, genetic algorithms, or Bayesian networks to solve the classification problem. Since most algorithms work as black-box approaches it is often difficult to understand and optimize the decision model. Problems such as over-fitting or tree pruning are difficult to tackle.

Visualization techniques can help to overcome these problems. The decision tree visualizer in SGI's MineSet system [17] shows an overview of the decision tree together with important parameters such as the attribute value distributions. The system allows an interactive selection of the attributes shown and helps the user understand the decision tree. A more sophisticated approach that also helps in decision tree construction is visual classification, as proposed by Ankerst et al. [3]. The basic idea is to show each attribute value by a colored pixel and arrange them in bars. The pixels of each attribute bar are sorted separately and the attribute with the purest value distribution is selected as the split attribute of the decision tree. The procedure is repeated until all leaves correspond to pure classes. An example of the decision tree resulting from this process is shown in Fig. 43.7. Compared to a standard visualization of a decision tree, additional information is provided that is helpful for explaining and analyzing the decision tree, namely

- Size of the nodes (number of training records corresponding to the node)

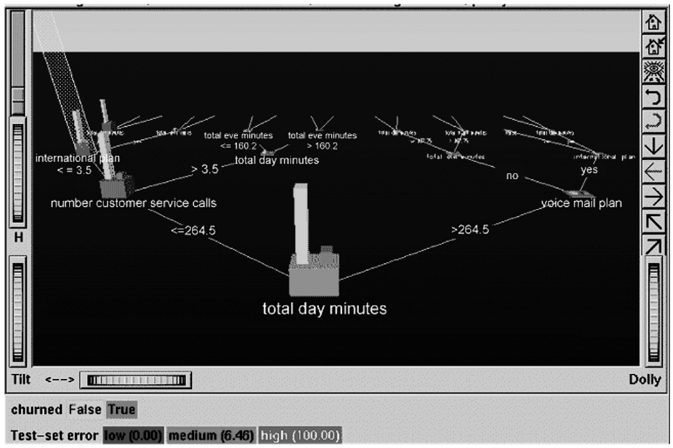


Figure 43.6 MineSets Decision Tree Visualizer [17] displays decision trees as 3D landscapes, each node contains bars whose height, color, and disk correspond to important parameters. © SGI

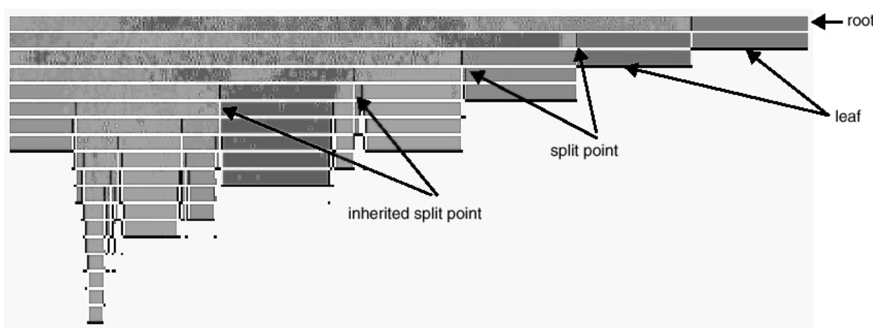


Figure 43.7 Visual Classification [3] shows each attribute value by a colored pixel and arranges them in bars (example shows a visualization of a decision trees for the DNA segment training data from the Statlog benchmark having 19 attributes). © ACM

- Quality of the split (purity of the resulting partitions)
- Class distribution (frequency and location of the training instances of all classes).

Some of this information might also be provided by annotating the standard visualization of a decision tree (for example, annotating the nodes with the number of records or the gini-index), but this approach clearly fails for more complex information such as the class distribution. In general, visualizations can help us to better understand the classification models and to easily interact with the classification algorithms in order to optimize the model generation and classification process.

43.5 Clustering

Clustering is the process of finding a partitioning of the dataset into homogeneous subsets called clusters. Unlike classification, clustering is *unsupervised learning*. This means that the classes are unknown and no training set with class labels is available. A wide range of clustering

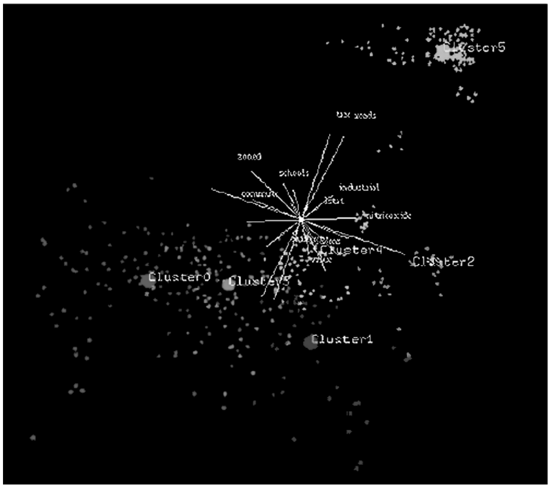
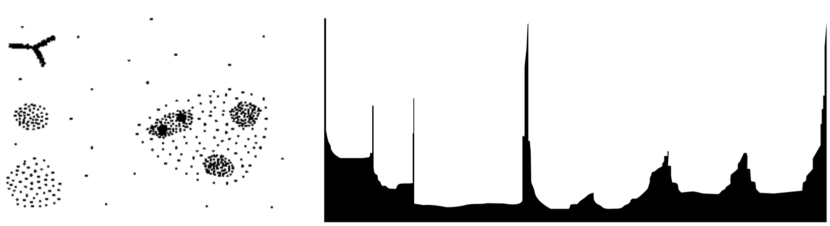


Figure 43.8 Visualization based on a projection into 3D space [39]: 3D cluster-guided projection, where the 3D subspace is determined by centroids of 4 clusters 0, 1, 3, 5. © ACM

algorithms have been proposed in the literature, including density-based methods such as kernel density estimation [29] and linkage-based methods [6]. Most algorithms use assumptions about the properties of the clusters that are either used as defaults or have to be given as input parameters. Depending on the parameter values, the user gets differing clustering results. In 2D or 3D space, the impact of different algorithms and parameter settings can easily be explored using simple visualizations of the resulting clusters (for example, x-y plots), but in higher-dimensional space the impact is much more difficult to understand. Some higher-dimensional techniques try to determine 2D or 3D projections of the data that retain the properties of the high-dimensional clusters as much as possible [39]. Fig. 43.8 shows a 3D projection of a dataset consisting of five clusters.

While this approach works well with low- to medium-dimensional datasets, it is difficult to apply to large high-dimensional datasets, especially if the clusters are not clearly separated and the dataset also contains noise (data that does not belong to any cluster). In this case, more sophisticated visualization techniques are needed to guide the clustering process, select the right clustering model, and adjust the parameter values appropriately. An example of a system that uses visualization techniques to help in high-dimensional clustering is OPTICS [2]. The idea of OPTICS (*Ordering Points To Identify the Clustering Structure*) is to create a 1D ordering of the database representing its density-based clustering structure. Fig. 43.9 shows a 2D example dataset together with its reachability distance plot. Intuitively, points within a cluster are close in the generated 1D ordering and their reachability distance shown in Fig. 43.9 is similar. Jumping to another cluster results in higher reachability distances. The idea works for data of arbitrary dimension. The reachability plot provides a visualization of the inherent clustering structure and is therefore valuable for



(a) ExampleDataSet

(b) Reachability Plot - objects are on the x-axis with their reachability values on the y-axis

Figure 43.9 OPTICS Visual Clustering [2]. © ACM understanding the clustering and guiding the clustering process.

Another interesting approach is the *HD-Eye* system [14]. The *HD-Eye* system considers the clustering problem a partitioning problem and supports a tight integration of advanced clustering algorithms and state-of-the-art visualization techniques, allowing the user to directly interact in the crucial steps of the clustering process. The crucial steps are the selection of dimensions to be considered, the selection of the clustering paradigm, and the partitioning of the dataset. Novel visualization techniques are employed to help the user identify the most interesting projections and subsets as well as the best separators for partitioning the data. Fig. 43.10 shows an example screenshot of the *HD-Eye* system with its basic visual components for cluster separation. The separator tree represents the clustering model produced so far in the clustering process. The *abstract iconic displays* (top-right and bottom-middle in Fig. 43.10) visualize the partitioning potential of a large number of projections. The properties are based on histogram information of the point density in the projected space. The number of data points belonging to the maximum corresponds to the color of the icon. The color follows a given color table ranging from dark colors for large maxima to bright colors for small maxima. The measure of how well a maximum is separated from the others corresponds to the shape of the icon, and the degree of separation varies from sharp spikes for well separated maxima to blunt spikes for badly separated maxima. The *color- and curve-based point density displays* present the density of the data and allow a better understanding of the data distribution, which is crucial for an effective partitioning of the data. The visualizations are used to decide which dimensions are used for the partitioning. In addition, the partitioning can be specified interactively

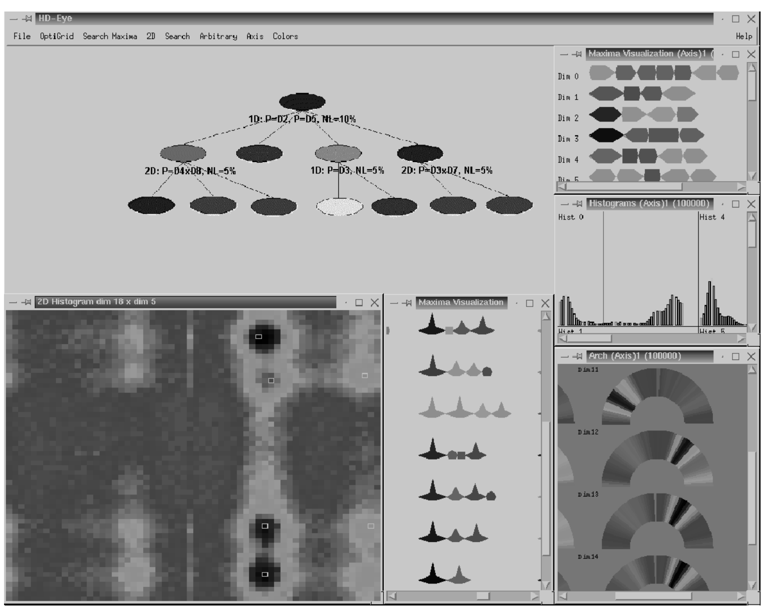


Figure 43.10 *HD-Eye* screenshot [14] showing different visualizations of projections and the separator tree. Clockwise from the top: separator tree, iconic representation of 1D projections, 1D projection histogram, 1D color-based density plots, iconic representation of multidimensional projections and color-based 2D density plot (example visualization shows a large molecular-biology dataset) © IEEE

directly within the visualizations, allowing the user to define nonlinear partitionings.

43.6 Text

With the growing importance of electronic media for storing and exchanging text documents, there is also a growing interest in tools that can help us find and sort information included in the text documents. Text documents are semistructured data, in that they are neither completely unstructured nor completely structured. For example, a document may contain some structured fields, such as title, authors, publication date, length, and category, as well as largely unstructured text components, such as abstract and content. Text mining is a process in finding for patterns in text databases, and may be defined as the process of analyzing text to extract information from it. Text mining recognizes that complete understanding of natural-language text, a long-standing goal of computer science, is not immediately attainable and focuses on extracting a small amount of information from text with high reliability. The goals of the text-mining process are automatic document clusterization/categorization, assignment of keywords to text documents, topic identification and tracking in ordered (time) sequences of text documents, searching documents based on the content categories and not only keywords, generation and analysis of user profiles based on the usage of text databases, and other related problems. A wide range of automatic text-mining algorithms have been proposed in the literature over the last few decades [10,11].

An interesting visual data-mining approach is ThemeRiver [13]. The ThemeRiver visualization depicts thematic variations over time within a large collection of documents. The thematic changes are shown in the context of a timeline and corresponding external events. The document collection's timeline, selected thematic content, and thematic strength are indicated by the river's directed flow, composition, and changing width, respectively. The directed flow from left to right is interpreted as movement through time, and the horizontal distance between two points on the river defines a time

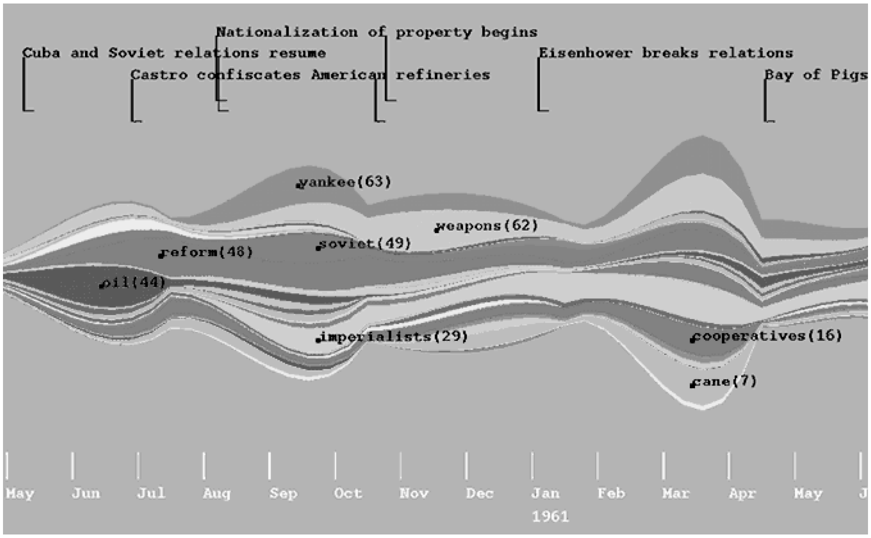


Figure 43.11 ThemeRiver [13]: visualization of thematic changes in documents (example visualization shows Castro data from November 1959 through June 1961). © IEEE.

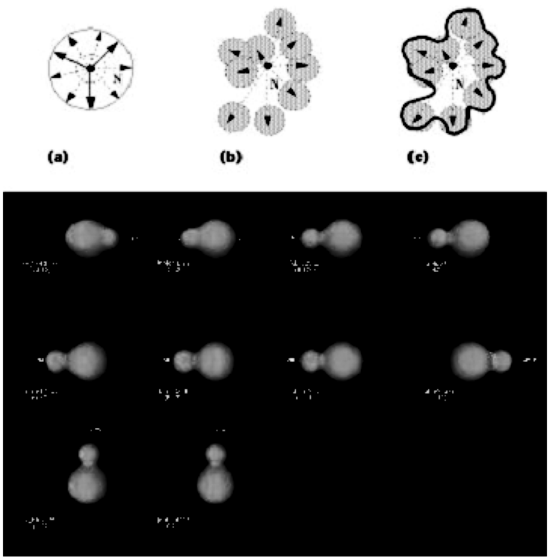


Figure 43.12 Shape-based Visual Interface for Text Retrieval [27]: shape-based visualization of query results (example visualization shows the result for the key words lion, sheep, mouse, and wolf). © ACM

interval. At any point in time, the vertical distance, or width, of the river indicates the collective strength of the selected themes. Colored "currents" flowing within the river represent individual themes. A current's vertical width narrows or broadens to indicate decreases or increases in the strength of the individual theme.

Another interesting approach is the shape-based visual interface for text retrieval [27].
Q1 This exploratorion system uses procedurally generated shapes coupled with an underlying text retrieval engine. Traditional text-based queries and summarization are enhanced with a visual interface based on 3D shapes (glyphs). The interface allows visualization of multidimensional relationships among documents and perception of more information than with conventional text-based interfaces.

43.7 Conclusion

The exploration of large datasets is an important but difficult problem. Information visualization techniques can be useful in solving this problem. Visual data exploration has a high potential, and many applications such as fraud detection and data mining can use information visualization technology for improved data analysis.

Avenues for future work include the tight integration of visualization techniques with traditional techniques from such disciplines as statistics, machine learning, operations research, and simulation. Integration of visualization techniques and these more established methods would combine fast automatic data-mining algorithms with the intuitive power of the human mind, improving the quality and speed of the data-mining process. Visual data-mining techniques also need to be tightly integrated with the systems used to manage the vast amounts of relational and semistructured information, including database management and data warehouse systems. The ultimate goal is to bring the power of visualization technology to every desktop to allow a better, faster, and more intuitive exploration of very large data resources. This will not only be valuable in an economic sense but will also stimulate and delight the user.

References

1. R. Agarwal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996.
2. M. Ankerst, M. Breunig, H. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. *Proc. ACM SIGMOD '99, Int. Conf on Management of Data, Philadelphia, PA*, pages 49–60, 1999.
3. M. Ankerst, M. Ester, and H. Kriegel. Towards an effective cooperation of the computer and the user for classification. *SIGKDD Int. Conf. On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA*, pages 179–188, 2000.
4. M. Ankerst, D. A. Keim, and H.-P. Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In *Proc. Visualization '95, Atlanta, GA*, pages 279–286, 1995.
5. M. Ankerst, D. A. Keim, and H.-P. Kriegel. Circle segments: A technique for visually exploring large multidimensional data sets. In *Visualization '96, Hot Topic Session, San Francisco, CA*, 1996.

6. H. H. Bock. *Automatic Classification*. Vandenhoeck and Ruprecht, Göttingen, 1974.
7. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.
8. S. Card, J. Mackinlay, and B. Shneiderman. *Readings in Information Visualization*. Morgan Kaufmann, 1999.
9. H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal Amer. Statistical Association*, 68:361–368, 1973.
10. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.
11. D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT-Press, 2001.
12. M. Hao, M. Hsu, U. Dayal, S. F. Wei, T. Sprenger, and T. Holenstein. Market basket analysis visualization on a spherical surface. *Visual Data Exploration and Analysis Conference, San Jose, CA*, 2001.
13. S. Havre, B. Hetzler, L. Nowell, and P. Whitney. Themeriver: Visualizing thematic changes in large document collections. *Transactions on Visualization and Computer Graphics*, 2001.
14. A. Hinneburg, D. Keim, and M. Wawryniuk. HD-Eye: Visual Mining of High-dimensional Data. *IEEE Computer Graphics and Applications*, 19(5), 1999.
15. H. Hofmann, A. Siebes, and A. Wilhelm. Visualizing association rules with interactive mosaic plots. *SIGKDD Int. Conf. On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA*, 2000.
16. D. T. Inc. Dbminer. *http://www.dbminer.com*, 2001.
17. S. G. Inc. Mineset. *http://www.sgi.com/software/mineset*, 2001.
18. A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *Proc. Visualization 90, San Francisco, CA*, pages 361–370, 1990.
19. B. Johnson and B. Shneiderman. Treemaps: A space-filling approach to the visualization of hierarchical information. In *Proc. Visualization '91 Conf*, pages 284–291, 1991.
20. D. Keim. Visual exploration of large databases. *Communications of the ACM*, 44(8):38–44, 2001.
21. D. Keim and M. Ward. *Visual Data Mining Techniques, Book Chapter in: Intelligent Data Analysis, an Introduction by D. Hand and M. Berthold*. Springer Verlag, 2 edition, 2002.
22. M. Mehta, R. Agrawal, and J. Rissanen. SLIQ: A fast scalable classifier for data mining. *Conf. on Extending Database Technology (EDBT), Avignon, France*, 1996.
23. R. M. Pickett. *Visual Analyses of Texture in the Detection and Recognition of Objects*. Academic Press, New York, 1970.
24. R. M. Pickett and G. G. Grinstein. Iconographic displays for visualizing multidimensional data. In *Proc. IEEE Conf. on Systems, Man and Cybernetics, IEEE Press, Piscataway, NJ*, pages 514–519, 1988.
25. J. R. Quinlan. Induction of decision trees. *Machine Learning*, pages 81–106, 1986.
26. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, Los Altos, CA, 1993.
27. R. M. Rohrer, J. L. Sibert, and D. S. Ebert. A shape-based visual interface for text retrieval. *IEEE Computer Graphics and Applications*, 19(5):40–47, 1999.
28. H. Schumann and W. Müller. *Visualisierung: Grundlagen und allgemeine Methoden*. Springer, 2000.
29. D. W. Scott. *Multivariate Density Estimation*. Wiley and Sons, 1992.
30. J. Shafer, R. Agrawal, and M. Mehta. SPRINT: A scalable parallel classifier for data mining. *Conf. on Very Large Databases*, 1996.
31. B. Shneiderman. Tree visualization with treemaps: A 2D space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992.
32. B. Shneiderman. The eye have it: A task by data type taxonomy for information visualizations. In *Visual Languages*, 1996.
33. B. Spence. *Information Visualization*. Pearson Education Higher Education publishers, UK, 2000.
34. P. E. Utgoff. Incremental induction of decision trees. *Machine Learning*, 4:161–186, 1989.
35. P. E. Utgoff, N. C. Berkman, and J. A. Clouse. Decision tree induction based on efficient tree restructuring. *Machine Learning*, 29:5–44, 1997.
36. J. Walter and H. Ritter. On interactive visualization of high-dimensional data using the hyperbolic plane. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 123–131, 2002.
37. M. O. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proc. Visualization 94, Washington, DC*, pages 326–336, 1994.
38. C. Ware. *Information Visualization: Perception for Design*. Morgen Kaufman, 2000.
39. L. Yan. Interactive exploration of very large relational data sets through 3d dynamic projections. *SIGKDD Int. Conf. On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA*, 2000.

AUTHOR QUERIES

Q1 Au: exploratorion?