

Visual Data Mining

Sunita Kanaujiya

Department of Information Technology, KIET, Ghaziabad
sunita_iet@yahoo.com

Abstract -Data Mining is a process of exploring and analyzing large quantities of data to discover useful knowledge (patterns and rules). Visual data mining is a novel approach to deal with the growing flood of information. Never before in history has data been generated at such high volumes as it is today. Exploring and analyzing the vast volumes of data is becoming increasingly difficult. Information visualization and visual data mining can help to deal with the flood of information. The aim is to combine traditional data mining algorithms with information visualization techniques to utilize the advantages of both approaches. The approach integrates the human mind's exploration abilities with the enormous processing power of computers to form a powerful knowledge discovery environment that capitalizes on the best of both worlds. The technology builds on visual and analytical processes developed in various disciplines including scientific visualization, data mining, statistics, and machine learning with custom extensions that handle very large, multidimensional, multivariate data sets. There are a large number of information visualization techniques which have been developed over the last decade to support the exploration of large data sets.

Keywords -Visual data mining, visualization , information visualization ,visualization technique.

I. INTRODUCTION

A visual data mining system must be syntactically simple to be useful. 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 means a customized data structure to facilitate fast and reliable searches. Simple to execute means a minimum number of steps needed to achieve the results. A reliable visual data mining system must provide estimated error or accuracy of the projected information for each step of the mining process.

A reusable visual data mining system must be adaptable to a variety of systems and 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 of existing knowledge cannot be planned. It may mean a portable system through telelinks or an embedded (local) system within the information domain. Finally, 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.

II. VISUAL DATA MINING TECHNIQUES

Information visualization focuses on data sets 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 data sets, 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 data sets. In the last decade, a large number of novel information visualization techniques have been developed, allowing visualizations of multidimensional data sets without inherent two or three- dimensional semantics. The techniques can be classified based on three criteria (see fig.1[6]):the data to be visualized, the visualization technique, and the interaction and distortion technique used.

The data type to be visualized [1] may be

- a. one-dimensional data, such as temporal data as used in ThemeRiver
- b. two-dimensional data, such as geographical maps as used in Polaris
- c. multidimensional data, such as relational tables as used in Polaris
- d. text and hypertext, such as news articles and Web documents as used in ThemeRiver
- e. hierarchies and graphs, such as telephone calls and Web documents as used in MGW and the Scalable Framework

The visualization technique used may be classified into

- a. standard 2D/3D displays, such as bar charts and x-y plots, as used in Polaris
- b. geometrically transformed displays, such as landscapes and parallel coordinates, as used in Scalable Framework
- c. icon-based displays, such as needle icons and star icons, as used in MGW
- d. dense pixel displays, such as the recursive pattern and circle segments techniques and the graph sketches as used in MGW
- e. stacked displays, such as treemaps or dimensional stacking

The third dimension of the classification is the interaction and distortion technique used. Interaction and distortion techniques allow users to directly interact with the visualizations. They may be classified into:

- a. Interactive Projection, as used in the GrandTour system
- b. Interactive Filtering, as used in Polaris

- c. Interactive Zooming, as used in MGV and the Scalable Framework
- d. Interactive Distortion, as used in the Scalable Framework

III. BENEFITS OF VISUAL DATA EXPLORATION

Visual data exploration aims at integrating the human in the data exploration process, applying its perceptual abilities to the large data sets available in today's computer systems. The basic idea of visual data exploration is to present the data in some visual form, allowing the human to get insight into the data, draw conclusions, and directly interact with the data. Visual data mining techniques have proven to be of high value in exploratory data analysis and they also have 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. The visualizations of the data allow the user to gain insight into the data and come up with new hypotheses. The verification of the hypotheses can also be done via visual data exploration, but it may also be accomplished by automatic techniques from statistics or machine learning. In addition to the direct involvement of the user, the main advantages of visual data exploration over automatic data mining techniques from statistics or machine learning are:

- a. Visual data exploration can easily deal with highly non homogeneous and noisy data,
- b. Visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.

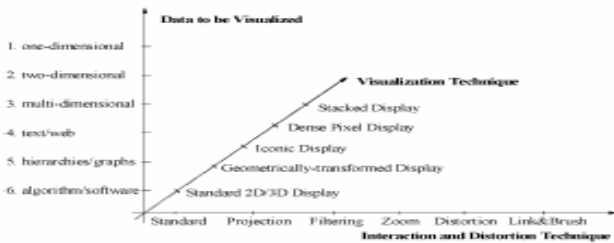


Fig. 1. Classification of information visualization techniques.

IV. DATA TYPE TO BE VISUALIZED

In information visualization, the data usually consists of a large number of records, each consisting of a number of variables or dimensions. Each record corresponds to an observation, measurement, transaction, etc. Examples are

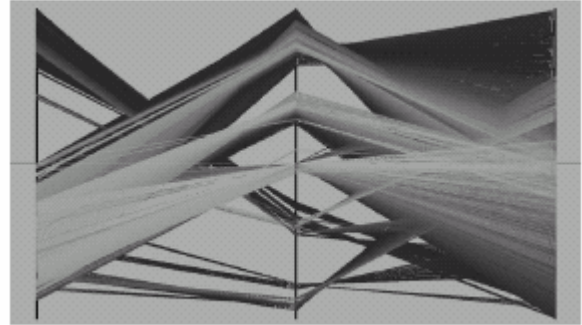


Fig. 2. Parallel coordinate visualization

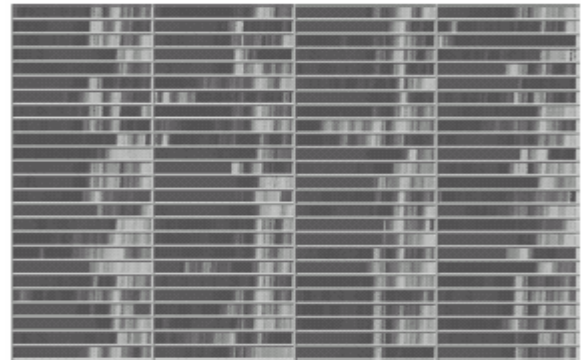


Fig. 3. Dense pixel displays: recursive pattern technique

customer properties, e-commerce transactions, and physical experiments. The number of attributes can differ from data set to data set: One particular physical experiment, for example, can be described by five variables, while another may need hundreds of variables. We call the number of variables the dimensionality of the data set. Data sets may be one-dimensional, two-dimensional, multidimensional, or may have more complex data types, such as text/hypertext or hierarchies/graphs.

A. ONE-DIMENSIONAL DATA

One-dimensional data usually has one dense dimension. A typical example of one-dimensional data is temporal data. Note that, with each point of time, one or multiple data values may be associated. Examples are time series of stock prices (see Fig. 3 and Fig. 4 for an example)

B. TWO-DIMENSIONAL DATA

Two-dimensional data has two distinct dimensions. A typical example is geographical data, where the two distinct dimensions are longitude and latitude. X-Y-plots are a typical method for showing two-dimensional data and maps are a special type of x-y-plots for showing two dimensional geographical data. Although it seems easy to deal with temporal or geographic data, caution is advised. If the number of records to be visualized is large, temporal axes and maps quickly get cluttered—and may not help to understand the data.

C. MULTIDIMENSIONAL DATA

Many data sets consists of more than three attributes and, therefore, they do not allow a simple visualization as twodimensional or three-dimensional plots. Examples of multidimensional

(or multivariate) data are tables from relational databases, which often have tens to hundreds of columns (or attributes). Since there is no simple mapping of the attributes to the two dimensions of the screen, more sophisticated visualization techniques are needed. An example of a technique which allows the visualization of multidimensional data is the Parallel Coordinate Technique [16](see Fig. 2, which is also used in the Scalable Framework (see Fig. 12 in [10]) . Parallel Coordinates display each multidimensional data item as a polygonal line which intersects the horizontal dimension axes at the position corresponding to the data value for the *corresponding dimension*.

D. TEXT AND HYPERTEXT

Not all data types can be described in terms of dimensionality. In the age of the world wide web, one important data type is text and hypertext as well as multimedia web page contents. These data types differ in that they cannot be easily described by numbers and, therefore, most of the standard visualization techniques cannot be applied. In most cases, a transformation of the data into description

vectors is necessary first before visualization techniques can be used. An example for a simple transformation is word counting (see ThemeRiver[7]), which is often combined with a principal component analysis or multidimensional scaling (for example, see [17]).

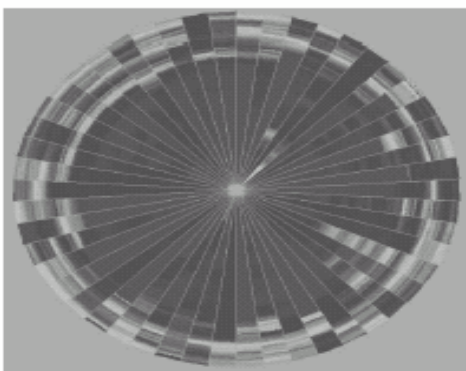


Fig. 4. Dense pixel displays: circle segments technique

E. HIERARCHIES AND GRAPHS

Data records often have some relationship to other pieces of information. Graphs are widely used to represent such interdependencies. A graph consists of a set of objects, called nodes, and connections between these objects, called edges. Examples are the e-mail interrelationships among people, their shopping behavior, the file structure of the hard disk, or the hyperlinks in the world wide web. There

are a number of specific visualization techniques that deal with hierarchical and graphical data. A nice overview of hierachical information visualization techniques can be found in [18], an overview of web visualization techniques at [19], and an overview book on all aspects related to graph drawing is [20].

F. ALGORITHMS AND SOFTWARE

Another class of data are algorithms and software. Coping with large software projects is a challenge. The goal of visualization is to support software development by helping to understand algorithms, e.g., by showing the flow of information in a program, to enhance the understanding of written code, e.g., by representing the structure of thousands of source code lines as graphs, and to support the programmer in debugging the code, i.e., by visualizing errors. There are a large number of tools and systems which support these tasks. An nice overview can be found in [21].

V. VISUALIZATION TECHNIQUES

There is a large number of visualization techniques which can be used for visualizing the data. In addition to standard 2D/3D-techniques, such as x-y (x-y-z) plots, bar charts, line graphs, etc., there are a number of more sophisticated visualization techniques. The classes correspond to basic visualization principles which may be combined in order to implement a specific visualization system.

A. GEOMETRICALLY TRANSFORMED DISPLAYS

Geometrically transformed display techniques aim at finding ^ointeresting^o transformations of multidimensional data sets. The class of geometric display techniques includes techniques from exploratory statistics, such as scatterplot matrices and techniques which can be subsumed under the term ^oprojection pursuit^o . Other geometric projection techniques include Prosection Views, Hyperslice , and the well-known Parallel Coordinates visualization technique . see Fig. 2).

B. ICONIC DISPLAYS

Another class of visual data exploration techniques are the iconic display techniques. The idea is to map the attribute values of a multidimensional data item to the features of an icon. Icons can be arbitrarily defined: They may be little faces, needle icons as used in MGV , star icons , stick figure icons, color icons , and TileBars. The visualization is generated by mapping the attribute values of each data record to the features of the icons. In the case of the stick figure technique, for example, two dimensions are mapped to the display dimensions and the remaining dimensions are mapped to the angles and/or limb length of the stick figure icon.

C. DENSE PIXEL DISPLAYS

The basic idea of dense pixel techniques is to map each dimension value to a colored pixel and group the pixels belonging to each dimension into adjacent areas [11]. Since, in

general, dense pixel displays use one pixel per data value, the techniques allow the visualization of the largest amount of data possible on current displays (up to about 1,000,000 data values). Dense pixel techniques use different arrangements for different purposes. By arranging the pixels in an appropriate way, the resulting visualization provides detailed information on local correlations, dependencies, and hot spots.

Well-known examples are the recursive pattern technique and the circle segments technique. The recursive pattern technique is based on a generic recursive back-and-forth arrangement of the pixels and is particularly aimed at representing datasets with a natural order according to one attribute (e.g., time series data). The idea of the circle segments technique is to represent the data in a circle which is

divided into segments, one for each attribute. Within the segments, each attribute value is again visualized by a single colored pixel. The arrangement of the pixels starts at the center of the circle and continues to the outside by plotting on a line orthogonal to the segment halving line in a back and forth manner. The rationale of this approach is that, close to the center, all attributes are close to each other, enhancing the visual comparison of their values. Fig. 4 shows an example circle segment visualization of the same data (50 stocks) as shown in Fig. 3).

D. STACKED DISPLAYS

Stacked display techniques are tailored to present data partitioned in a hierarchical fashion. In the case of multidimensional data, the data dimensions to be used for partitioning the data and building the hierarchy have to be selected appropriately.

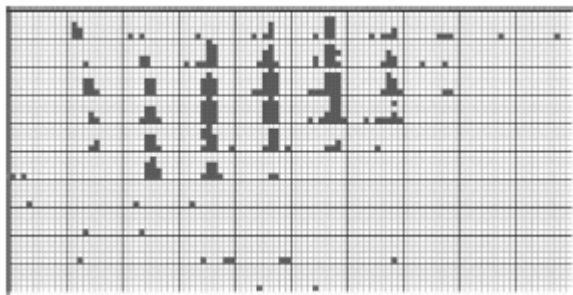


Fig. 5. Dimensional stacking visualization of oil mining data (used by permission of M. Ward, Worcester Polytechnic

An example of a stacked display technique is Dimensional Stacking. The basic idea is to embed one coordinate system inside another coordinate system, i.e., two attributes form the outer coordinate system, two other attributes are embedded into the outer coordinate system, and so on. The display is generated by dividing the outmost level coordinate systems into rectangular cells and, within the cells, the next two attributes are used to span the second level coordinate system. This process may be repeated one more time.

VI. INTERACTION AND DISTORTION TECHNIQUES

Interaction techniques allow the data analyst to directly interact with the visualizations and dynamically change the visualizations according to the exploration objectives and they also make it possible to relate and combine multiple independent visualizations.

Distortion techniques help in the data exploration process by providing means for focusing on details while preserving an overview of the data. The basic idea of distortion techniques is to show portions of the data with a high level of detail, while others are shown with a lower level of detail.

A. DYNAMIC PROJECTIONS

The basic idea of dynamic projections is to dynamically change the projections in order to explore a multidimensional data set. A classic example is the GrandTour system [15], which tries to show all interesting two-dimensional projections of a multidimensional data set as a series of scatter plots.

Note that the number of possible projections is exponential in the number of dimensions, i.e., it is intractable for a large dimensionality. The sequence of projections shown can be random, manual, precomputed, or data driven. Systems supporting dynamic projection techniques are XGobi, XLispStat, and ExplorN

B. INTERACTIVE FILTERING

Therefore, a number of interaction techniques have been developed to improve interactive filtering in data exploration. An example of an interactive tool which can be used for interactive filtering is Magic Lenses. The basic idea of Magic Lenses is to use a tool like a magnifying glass to support filtering the data directly in the visualization. The data under the magnifying glass is processed by the filter and the result is displayed differently than the remaining data set. Magic Lenses show a modified view of the selected region, while the rest of the visualization remains unaffected. Note that several lenses with different filters may be used; if the filters overlap, all filters are combined.

C. INTERACTIVE ZOOMING

Zooming is a well-known technique which is widely used in a number of applications. In dealing with large amounts of data, it is important to present the data in a highly compressed form to provide an overview of the data, but, at the same time, allow a variable display of the data on different resolutions. Zooming not only means to display the data objects larger, but also means that the data representation automatically changes to present more details on higher zoom levels. An interesting example applying the zooming idea to large tabular data sets is the TableLens approach. The basic idea of TableLens is to represent each numerical value by a small bar. The initial view allows the user to detect patterns, correlations, and outliers in the data set.

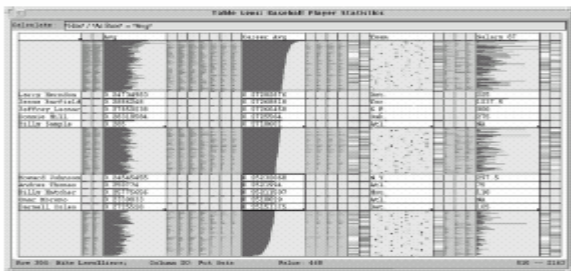


Fig. 6. Table Lenses (used by permission of R. Rao, Xerox PARC & ACM).

In order to explore a region of interest, the user can zoom in, with the result that the affected rows (or columns) are displayed in more detail, possibly even in textual form. Fig. 6 shows an example of a baseball database with a few rows being selected in full detail.

D. INTERACTIVE DISTORTION

Interactive distortion techniques support the data exploration process by preserving an overview of the data during drill-down operations. The basic idea is to show portions of the data with a high level of detail while others are shown with a lower level of detail. Popular distortion techniques are hyperbolic and spherical distortions, which are often used on hierarchies or graphs, but may be also applied to any other visualization technique. Examples of distortion techniques include Bifocal Displays, perspective Wall, Graphical Fisheye Views, Hyperbolic Visualization, and Hyperbox.

E. INTERACTIVE LINKING AND BRUSHING

The idea of linking and brushing is to combine different visualization methods to overcome the shortcomings of single techniques. Scatterplots of different projections, for example, may be combined by coloring and linking subsets of points in all projections. In a similar fashion, linking and brushing can be applied to visualizations generated by all visualization techniques described above. As a result, the brushed points are highlighted in all visualizations, making it possible to detect dependencies and correlations. Interactive changes made in one visualization are automatically reflected in the other visualizations. Typical examples of visualization techniques which are combined by linking and brushing are multiple scatterplots, bar charts, parallel coordinates, pixel displays, and maps.

CONCLUSION

The exploration of large data sets is an important but difficult problem. Information visualization techniques may help to solve the problem. Visual data exploration has high potential and many applications, such as fraud detection and data mining, will use information visualization technology for an improved data analysis. Future work will involve 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 quickly combine automatic data mining algorithms with the intuitive power of the human mind, improving the quality and speed of the visual 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] B. Shneiderman, TMThe Eye Have It: A Task by Data Type
- [2] D. Keim, TMVisual Exploration of Large Databases
- [3] L. Nowell, S. Havre, B. Hetzler, and P. Whitney, TMThemeriver: Visualizing Thematic Changes in Large Document Collections. *IEEE Trans. Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9-20, Jan.-Mar. 2002.
- [4] D. Tang, C. Stolte, and P. Hanrahan, TMPolaris: A System for Query, Analysis and Visualization of Multidimensional Relational Databases, ^o *IEEE Trans. Visualization and Computer Graphics*, vol. 8, no. 1, pp. 52-65, Jan.-Mar. 2002.
- [5] J. Abello and J. Korn, TMMGV: A System for Visualizing Massive Multidigraphs, ^o *IEEE Trans. Visualization and Computer Graphics*, vol. 8, no. 1, pp. 21-38, Jan.-Mar. 2002.
- [6] D. Keim, TMDesigning Pixel-Oriented Visualization Techniques: Theory and Applications, ^o *IEEE Trans. Visualization and Computer Graphics*, vol. 6, no. 1, pp. 59-78, Jan.-Mar. 2000.
- [7] B. Johnson and B. Shneiderman, TMTreemaps: A Space-Filling Approach to the Visualization of Hierarchical Information, ^o *Proc. Visualization '91 Conf.*, pp. 284-291, 1991.
- [8] M.O. Ward, TMXmdvtool: Integrating Multiple Methods for Visualizing Multivariate Data, ^o *Proc. Visualization 94*, pp. 326-336, 1994.
- [9] C. Chen, *Information Visualisation and Virtual Environments*. London: Springer-Verlag, 1999.
- [10] M. Dodge, TMWeb Visualization, ^o http://www.geog.ucl.ac.uk/casa/martin/geography_of_cyberspace.html, Oct. 2001.
- [11] G.W. Furnas and A. Buja, TMProsections Views: Dimensional Inference through Sections and Projections, ^o *J. Computational and Graphical Statistics*, vol. 3, no. 4, pp. 323-353, 1994.
- [12] R. Spence, L. Tweedie, H. Dawkes, and H. Su, TMVisualization for Functional Design, ^o *Proc. Int'l Symp. Information Visualization*