Submission for SIGMOD'2002 Demonstration

XmdvTool: Visual Interactive Data Exploration and Trend Discovery of High-dimensional Data Sets *

Elke A. Rundensteiner, Matthew O. Ward, Jing Yang, and Punit R. Doshi
Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609, USA
tel.: (508) 831–5815, fax: (508) 831–5776
{rundenst,matt,yangjing,punitd}@cs.wpi.edu

1 Visual Data Exploration

Multi-dimensional data is being generated at an ever increasing rate in practically all modern applications. The development of techniques and tools to extract useful information out of such data is one of critical challenges to be tackled in the 21st century. Visualization is one popular technique for achieving effective data exploration by exploiting the visual perception abilities of domain experts. Visualization involves the graphical presentation of data and information for the purposes of communicating results, verifying hypotheses, and qualitative exploration.

In this demonstration, we present our solution to particular challenges we have been tackling in this area in the context of our XMDVtool project, a multi-year effort funded by NSF. These include multivariate data visualization to facilitate outlier and pattern discovery via a variety of displays, visual interaction tools, scalability of these visualization techniques to large data sets, interaction with commercial database technology and, more recently, extensions to handle data of very high-dimensionality.

2 *XmdvTool*: Multi-Dimensional Visualization and Exploration

***XmdvTool* Goals.** *XmdvTool* is a public-domain software package we have been developing at Worcester Polytechnic Institute for the interactive visual exploration of multi-variate data sets [4, 2]. XmdvTool supports an active process of discovery as opposed to passive display. The major hurdles we overcome are the problems of display clutter (too much data at once tends to confuse viewers and too many dimensions hinders the users from find-ind useful data features), intuitive navigation (what tasks comprise a typical exploration process, and how they can be made intuitive), and efficient data access for the above operations over large data sets (to allow for near real-time interactive exploration).

Multi-Faceted Visualization. *XmdvTool* incorporates several distinct display methods for multivariate data visualization that allow the users to view data from different perspectives. For example, Figures 1, 2, 3, and 4 show four displays of the Iris data set (Anonymous ftp from unix.hensa.ac.uk in /pub/statlib/datasets):

- **scatterplot matrices,** consisting of a grid of all pairwise scatterplots of the N-dimensional data (Fig. 1),
- **star glyphs**, displaying each N-dimensional data point as an N-sided polygon, with N evenly spaced rays whose lengths are driven by data values (Fig. 2),
- **parallel coordinates,** where each dimension is depicted as a vertical axis, and each data point manifests itself as a polyline across the axes (Fig. 3), and
- **dimensional stacking,** a recursive embedding of pairs of dimensions in which each dimension is discretized into a small number of bins, (Fig. 4).

The displays are tightly linked, such that visual interactions via one display can be seamlessly refined via other displays. XmdvTool supports a variety of advanced visual interaction tools, including brushing in screen space, data

* This work is supported in part by several grants from NSF, namely, NSF grant IIS-9732897, NSF CISE Instrumentation grant IRIS 97-29878, and NSF grant IIS-0119276.

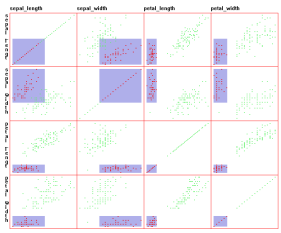

Figure 1: Iris data set in Flat Scatterplot Matrixes

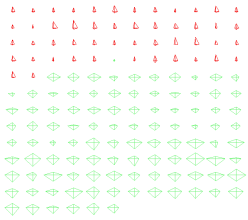

Figure 2: Iris data set in Flat Star Glyphs

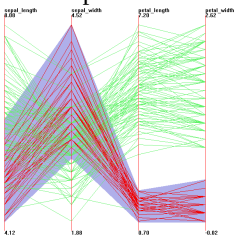

Figure 3: Iris data set in Flat Parallel Coordinates

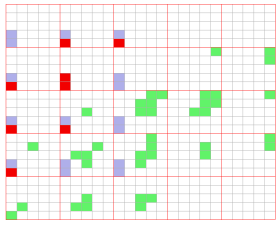

Figure 4: Iris data set in Flat Dimensional Stacking

space, and structure space, panning, zooming and distortion [4, 2].

Visual Exploration Scale-Up. Most conventional multivariate visualization techniques, including those four utilized in Figures 1 to 4, generally do not scale well with respect to the number of objects in the data set. For example, directly applying such parallel coordinates displays from Fig. 3 to large data sets would result in a visualization with an unacceptable level of clutter. As illustrated in Fig. 5, the display of a mass of overlapping lines already precludes the perception of relative densities present in the AAUP data set of only 1161 data points.

To overcome this limitation, XmdvTool adopts a hierarchical approach [3] that presents a multiresolutional view of the data. Such hierarchy is achieved by applying *hierarchical clustering* such as Birch (Wisconsin 1996) to structure the data set. Then we use a variation on parallel coordinates, which we call *hierarchical parallel coordinates*, to convey aggregation information for the resulting clusters. Lastly, we provide a suite of navigation and filtering tools to facilitate the systematic discovery of data trends and hidden patterns by seeing the desired focus region and level of detail. Fig. 6 depicts a pattern identified in the AAUP data set using the hierarchical approach that was not discernable in the flat display in Fig. 5.

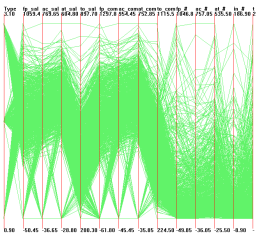

Figure 5: Parallel Coordinates display of AAUP data set: 14-dim. data set with 1161 records.

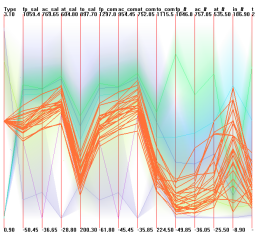

Figure 6: Hierarchical Parallel Coordinates display of AUUP: Highlighted cluster displayed in more detail than other clusters.

Efficient Data Acess Support. Efficient database transcriptions of the operations defined in the visualization context, such as visual hierarchical drill-down and roll-up, are critical to allow for the near real-time behavior required for interactive tools. Typically, such operations have a recursive definition and require expensive computation for each step of the recursion. We have shown [5] that the hierarchical exploration can be reduced to a two-dimensional exploration by identifying a particular class of recursive unions of joins and divisions that operate on hierarchies. Then, using adequate pre-computation (i.e., organizing the hierarchical structure as what we called a MinMax tree [5]), this recursive processing is shown to be reducible to range queries. MinMax has allowed us to achieve performance levels required for interactive visualization even when connecting to large persistent data sets on Oracle.

High-Dimensionality Reduction. When visualizing data sets with large number of dimensions, existing multidimensional visualization techniques become seriously cluttered and thus uneffective. One important solution to this problem is to reduce the dimensionality of the data while maintaining the relationships between data points. Current dimensionality reduction approaches, such as Principal Component Analysis and Multidimensional Scaling, have the major drawback that the generated dimensions no longer signify any clear meaning for the users. We are exploring a new approach to dimensionality reduction, that we call *visual hierarchical dimension reduction*. For this, we construct hierarchical dimension cluster trees based on clustering the dimensions, instead

of as done traditionally the data points. Thereafter, we explore how to construct low dimensional spaces guided by user interaction of the hierarchical dimension cluster tree.

Prefetching-Driven Caching. Exploration via visual interaction tools typically results in predictable traversal patterns of the data sets and thus effectively continuous query refinements. Hence customized caching and prefetching techniques have been shown to be effective in our tool [5]. In particular, we are employing semantic caching principles [1, 5] to maintain in the local client buffer not only the relevant results of the previously executed queries but also their query specifications, called *semantic descriptors*. When a request R is issued, these query specifications are used to determine what objects satisfying R are in the cache and what others need to be fetched from the server.

To further reduce the response time, we have designed a *speculative prefetcher* that brings data into memory when the system is idle. The prefetcher is based on the property of exploratory systems that queries remains "local", i.e., given the set of currently selected objects we have a small number of choices of which objects can be selected next. The property provides therefore "implicit hints" to the system. XmdvTool incorporates a suite of distinct prefetching techniques that are applied based on both the analysis of current user interactions as well as archived user session history. For the later, XmdvTool incorporates a session collection feature where entire user sessions are recorded and can be replayed at a later time (for analysis and also for experinmentation purposes).

3 *XmdvTool* Implementation

XmdvTool 5.0 is implemented in C/C++ with TclTk and OpenGL primitives. Interaction to commercial databases, in particular Oracle 8i, are written in C with ProcC* embedded SQL primitives. The XmdvTool Home Page at *http://davis.wpi.edu/~xmdv* provides for information related to the software, including multi-platform downloads of yearly releases of our software.

4 Demonstration of Scenarios

Our demonstration will show the system in action with multiple real-data sets from distinct application areas, including census and traffic accident data sets. Features to be illustrated include:

- Interactive data exploration processes that illustrate the linking and consistency among all the flat and hierarchical displays in the system.
- Real-time drill-down and roll-up navigation through hierarchical displays, and their effectiveness in helping us to find patterns and outliers in large data sets.
- Dimension reduction with visual interaction tools that find patterns from high-dimensional data sets.
- Illustration of performance of exploration tasks with different system settings, including client caching off and on, and prefetching strategies off and on.
- Automated data exploration simulations that intercept the interactive interface by running real user traces collected by our session capture tool.

References

[1] S. Dar, M. J. Franklin, B. T. Jónsson, D. Srivastava, and M. Tan. Semantic data caching and replacement. In *VLDB 1996*, pages 330–341.

[2] Y. Fua, M. Ward, and E. Rundensteiner. Structure-based brushes: A mechanism for navigating hierarchically organized data and information spaces. *IEEE Visualization and Computer Graphics, 2000*, pages 150–159.

[3] Y. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. *IEEE Visualization, Oct. 1999*, pages 43–50.

[4] J. Yang, M. O. Ward and E. A. Rundensteiner. Hierarchical exploration of large multivariate data sets. *Dagstuhl '00: Scientific Visualization*, May 2001, to appear.

[5] I. D. Stroe, E. A. Rundensteiner, and M. O. Ward. Scalable visual hierarchy exploration. In *Database and Expert Systems Applications, Greenwich, UK*, Sept. 2000, pages 784–793.