XmdvTool^Q: Quality-Aware Interactive Data Exploration *

Elke Rundensteiner †, Matthew Ward, Zaixian Xie, Qingguang Cui,
Charudatta Wad, Di Yang, Shiping Huang
Worcester Polytechnic Institute

ABSTRACT

In this work, we describe our approach for making the interactive data exploration system, called XmdvTool, *quality-aware* to assure informed decision-making. XmdvTool^Q makes quality or lack thereof explicit for all stages of the data exploration process from raw data, to abstracted data, to the final visual displays, allowing users to query and navigate through data-, structure- and quality-spaces.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*quality measures*.

General Terms

Measurement, Management, Human Factor.

Keywords

Data Quality, Abstraction Quality, Display Quality.

1. INTRODUCTION

Exploratory data analysis is a critical task at the core of a large variety of application domains, including homeland security, bioinformatics and product quality control. Over the past ten years the XMDV team at WPI, composed of visualization, HCI and database experts, supported by a series of four NSF grants, has developed a tool suite called XmdvTool to facilitate activities related to interactive data analysis. During data exploration in XmdvTool, data is interactively selected from a database based on user interactions. Alternatively, the dataset may be first manipulated using some abstraction method, be it clustering or sampling, to construct layers of data abstraction, each successively more compact while ideally conveying the main features of the data. In XmdvTool^Q, users can interactively navigate the abstraction hierarchy to select data views of interest [3, 2]. The selected data is then mapped to graphical entities of the display based on chosen display view, be it a scatterplot or a parallel coordinate display [4, 6], and then rendered on the screen. User interactions can be used to modify and control any stage of this exploration process. Example interactions include progressive data filtering, zooming or distorting, and reduction of dimensions for high-dimensional data [8].

While being effective at supporting many data exploration tasks as illustrated by numerous users and case studies [5], our current research overcomes one major limitation experienced by XmdvTool and equally by most other state-of-the-art data exploration systems. Namely, operating under the assumption that data is perfect could potentially mislead the users and cause incorrect conclusions. Thus we make XmdvTool *quality-aware* by considering quality or lack thereof at all stages of data exploration. Types of quality we address in our work include:

- The quality of the data itself, which may vary based on source reliability and completeness.
- The quality of data transformations and the resulting information artifacts being generated.
- The quality of the query process, trading off the response time with displayed information content.
- The quality of different visual mappings and screen layout; resulting in visualizations with different perceptible structures and various degrees of clutter.

Our enhanced tools, henceforth called XmdvTool^Q, now enable effective exploration by explicitly exposing and integrating the quality of the data, thus equipping the human decision maker with insight into and control over the type of data and transformations they work with. This should result in more informed decision-making. We will demonstrate that our approach is unique, in that we employ a wide variety of quality-related techniques for data exploration [1, 7, 5], including: (a) Metrics for measuring quality, (b) Algorithms for quality optimization based on selected quality metrics, (c) Visualizations to convey quality information for both data and structure spaces, and (d) Interactive tools to control the optimization processes and specify preferences when trade-offs exist. Our techniques are general, and could easily be integrated into other existing visual exploration tools.

2. THE XMDVTOOL^Q SYSTEM

Enhancing XmdvTool with quality-awareness has resulted in changes to most components of our system (Figure 1). The Data Manager supports the modeling of data quality

*This work was supported under NSF grants IIS-9732897, IRIS 97-29878, IIS-0119276 and IIS-00414380.
†Email: rundenst@cs.wpi.edu

SIGMOD'07, June 11–14, 2007, Beijing, China.
ACM 978-1-59593-686-8/07/0006.

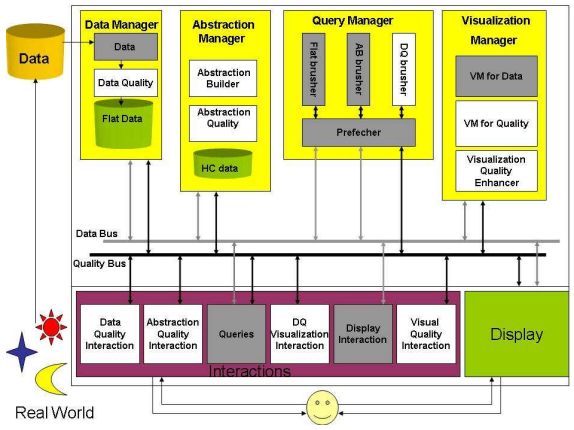

Figure 1: System Architecture of XmdvTool^Q

at multiple granularities of the original data sets (See Section 3), while the Abstraction Manager summarizes huge data sets by building progressively compact abstractions maintained in abstraction hierarchies, now augmented by their quality measures (See Section 4). Query manager supports efficient data extraction employing caching and prefetching technologies. Different retrieval functionality, called brushes, for queries on flat data, abstracted data, and data quality respectively is supported. Visualization Manager visualizes both data and data quality (Section 3), further taking quality of the actual display layout itself into account (Section 5). A rich suite of interactions is provided to users to navigate into and control any of the given spaces: from data, to abstraction, into quality spaces. The workflow in the system is controlled by a Data and a Quality Bus.

3. DATA QUALITY

We model not just the data but also its quality characteristics using a multi-granularity model [7]. *Quality measures for data* consist of values for the quality of records, of dimensions, and of actual data values. For simplicity, all quality measures are normalized to the range of zero (lowest quality) to one (perfect quality). To illustrate our techniques, here we will use one popular multivariate visualization, called parallel coordinates, though within our system we support five such display methods. In this method depicted in Figure 2, each dimension corresponds to an axis, and the N axes are organized as vertical lines. A data element in an N-dimensional space manifests itself as a connected set of points, one on each axis. Thus a polyline is generated for representing each data point.

We propose two alternate approaches for integrating quality into existing multivariate visualizations [7]. The first, *visual encoding*, conveys quality measures using the graphical attributes of visual entities in the display (such as, hue, brightness, line thickness, etc.). For example, in parallel coordinates, we map record quality and dimension quality to the graphical attributes of polylines and vertical axes respectively, while the quality of a particular data value is indicated by visual attributes for its mapped subsegment of the polyline. The second approach, called *new-dimensions*, instead extends the original dataset by constructing new dimensions to record the value quality and record quality. Our methods are natural extensions of existing display techniques, thus all techniques for multivariate visualization and their interactions can still be applied to these quality-enhanced datasets. For instance, we can use N-dimensional brushing to create a quality brush to select data points with high quality.

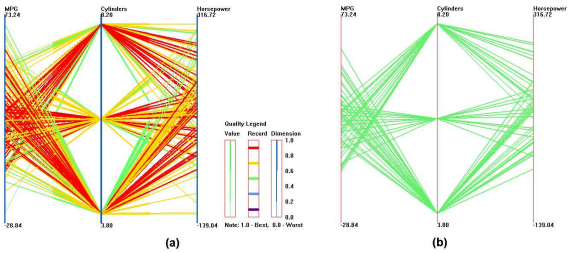

Figure 2: *Cars* dataset (a) conveying data quality by visual encoding method; (b) after filtering out data with low quality using quality brushing.

In Figure 2(a), the record quality is mapped to hue, while value and dimension qualities are both mapped to line width. In Figure 2(b), we used the *quality brush* to filter out those data points with low quality. This facilitates users to draw correct conclusions more quickly by working only with data with acceptable quality.

4. ABSTRACTION QUALITY

Data abstraction reduces a large dataset into one of moderate size, while maintaining dominant characteristics of the original dataset. We propose to quantify the capture of the main features of the dataset by the abstraction via what we term *abstraction quality* [1].

Without any knowledge about the quality or lack of quality of an abstraction we work with, the reliability of results gleaned from these abstractions could be in serious jeopardy. For this, we have designed several measures to assess the quality of an abstraction. For instance, Histogram Difference Measure (HDM) compares the normalized histograms of the original data and abstracted data to calculate their differences and thus quality of the later in terms of capturing the former [1]. These measures can be used to compare the quality of different abstractions when choosing between them, as well as the effectiveness of transformation algorithms in generating good abstractions.

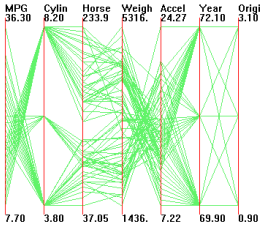

Figure 3: Lower abstraction quality (dataset: *Cars*)

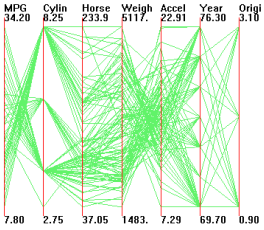

Figure 4: Higher abstraction quality (dataset: *Cars*)

Further, in our system, users can interactively select their desired performance versus quality trade off. This empowers the user to decide the desired level of quality they require; and the algorithm then will generate the abstraction

according to this quality specification. Our estimator provides a performance-quality trade-off chart that binds the given quality measures with the time required to construct abstractions of this level of quality. Abstractions are then visualized together with their quality, allowing a user to navigate in the quality space to adjust its visualization quality.

Figures 3 and 4 represent abstractions with different abstraction quality levels. Abstraction in Figure 3 looses some small clusters whereas abstraction in Figure 4 captures them.

5. VISUAL QUALITY

Visual clutter denotes a disordered collection of graphical entities in a visualization. Clutter can obscure the structure present in the data. Even in a small dataset, clutter can make it hard for the viewer to find patterns and reveal relationships. We identify dimension order as one (of many) display attributes that can significantly affect a visualization's expressiveness. By varying the dimension order in a display, it is possible to reduce clutter, such as the crossing of lines back and forth or the overwriting of data on top of each other in the display, without reducing information content or modifying the data. Thus, we define visual clutter as any aspect of the visualization that interferes with the viewer's understanding of the data.

Clutter reduction is a display-dependent task. For each display technique, we first determine what constitutes clutter in terms of display properties and then we design a metric to measure visual clutter in this display. Finally, we propose algorithms for clutter-based dimension reordering to improve the visual quality.

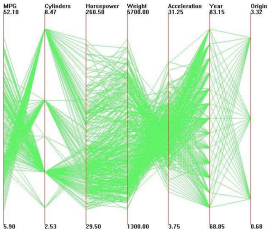

Figure 5: Before Clutter based reordering (dataset: *Cars*)

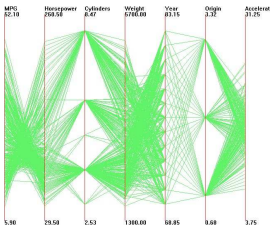

Figure 6: After Clutter based reordering (dataset: *Cars*)

In the parallel coordinates display, for instance, as the axes order is changed, the polylines representing data points take on very distinct shapes. (See Figure 6). Parallel coordinates make inter-dimensional relationships between neighboring dimensions easy to see, while not disclosing much about relationships between non-adjacent dimensions. The existence of many outliers, data points that don't belong to any of the clusters between two dimensions, tends to indicate that there is little relationship between the two. Thus, clutter in parallel coordinates can be defined as the proportion of outliers divided by the total number of data points. Since our goal is to disclose more relationships and patterns between adjacent dimensions, we propose algorithms that rearrange the dimensions to minimize the outliers between neighboring dimensions. In Figure 5 data is displayed with default dimension ordering. Figure 6 displays the data after being processed with clutter-based ordering. It can be seen that in Figure 6 the data points are better separated and thus it is easier to view patterns.

6. SOFTWARE DEMONSTRATION

XmdvTool is made available as freeware tool to the community, with successive versions being released regularly at `http://davis.wpi.edu/~xmdv`. In our demonstration, we will illustrate the quality-aware enhancement of all stages of XmdvTool and let the audience firsthand experience its impact on the data exploration process. Further, we demonstrate that this quality-awareness support is compatible with existing visualizations, and how best to communicate quality-related information to the analysts. The demonstration will be based on 3 real data sets, namely, the Census data, the Skyserver data, and the Cars dataset. Demonstrations include, but are not limited to:

- Changes in interpretability and thus informed decision making based on whether and if so how the the quality of the explored data itself is conveyed on the display.
- Ranking of different data abstractions based on their quality measure, allowing the users to trade-off the desired degree of abstraction quality versus response time and sizes of retrieved data subsets.
- Novel interaction tools that support exploration not only over quality-enhanced data or structure-spaces but also directly in the quality-space, resulting in more effective pattern discovery support.
- Metrics and algorithms for measuring display quality in terms of perceived clutter, revealing previously clouded information in the data.

7. REFERENCES

[1] Q. Cui, M. Ward, E. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. *IEEE TVCG*, 12(5):709–716, 2006.

[2] P. Doshi, G. Rosario, E. Rundensteiner, and M. Ward. A strategy selection framework for adaptive prefetching in data visualization. *SSDBM*, pages 107–116, 2003.

[3] Y. Fua, M. Ward, and E. Rundensteiner. Structure-based brushes: A mechanism for navigating hierarchically organized data and information spaces. *IEEE TVCG*, 6(2):150–159, 2000.

[4] A. Inselberg. The plane with parallel coordinates. *Special Issue on Computational Geometry, The Visual Computer*, 1:69–97, 1985.

[5] W. Peng, M. Ward, and E. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. *IEEE InfoVis*, 2004.

[6] E. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 411(85):664–675, 1990.

[7] Z. Xie, S. Huang, M. Ward, and E. Rundensteiner. Exploratory visualization of multivariate data with variable quality. *IEEE VAST*, pages 183–190, 2006.

[8] J. Yang, A. Patro, S. Huang, N. Mehta, M. Ward, and E. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. *IEEE InfoVis*, pages 73–80, 2004.