Future Generation Computer Systems 55 (2016) 510–523

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A novel virtual node approach for interactive visual analytics of big datasets in parallel coordinates

Mao Lin Huang [a,b,*], Tze-Haw Huang [b], Xuyun Zhang [b]

[a] School of Computer Software Tianjin University, China 92 Weijin Rd., Nankai District, Tianjin, China

[b] School of Software, FEIT, University of Technology, Sydney P.O. Box 123, Broadway NSW 2007, Australia

HIGHLIGHTS

- Create "virtual nodes" that innovatively makes direct "mouse click" on data items possible in parallel coordinates visualization.
- Refine the classification of visual interactions into a four-layer model.
- The new approach can handle visualization and interaction with extremely large dataset.

ARTICLE INFO

Article history:
Received 18 June 2013
Received in revised form
3 July 2014
Accepted 10 February 2015
Available online 27 March 2015

Keywords:
Big data
Visual analytics
Parallel coordinates
Hierarchical clustering
Multidimensional data visualization
Data retrieval

ABSTRACT

Big data is a collection of large and complex datasets that commonly appear in multidimensional and multivariate data formats. It has been recognized as a big challenge in modern computing/information sciences to gain (or find out) due to its massive volume and complexity (e.g. its multivariate format). Accordingly, there is an urgent need to find new and effective techniques to deal with such huge datasets. Parallel coordinates is a well-established geometrical system for visualizing multidimensional data that has been extensively studied for decades. There is also a variety of associated interaction techniques currently used with this geometrical system. However, none of these existing techniques can achieve the functions that are covered by the *Select* layer of Yi's Seven-Layer Interaction Model. This is because it is theoretically impossible to find a *select* of data items via a mouse-click (or mouse-rollover) operation over a particular visual poly-line (a visual object) with no geometric region. In this paper, we present a novel technique that uses a set of virtual nodes to practically achieve the *Select* interaction which has hitherto proven to be such a challenging sphere in parallel coordinates visualization.

1. Introduction

Modern Information Visualization techniques, at their core, appear to have two main components: representation and interaction. The representation component is concerned with the mapping from data to advanced computer graphics and how to draw or render them on the display. On the other hand, the interaction component is concerned with the dialog between the user and the data stored on the system as the user explores the dataset for the purposes of uncovering insights. The interaction component's roots lie in the area of Computer–Human Interaction (CHI). Although discussed as two separate components, representation and interaction clearly are not mutually exclusive. While an information visualization system takes the role of providing advanced GUIs for supporting Computer–Human Interaction (CHI), it is supposed to facilitate CHI in both directions; i.e. (1) the input from human to computer (or data), and (2) the output from computer (or data) to human. However, in the past few decades, researchers in the InfoVis community have paid more attentions to the output part; which is concerned more about the visual representation of output data, such as the output of analysis results in order for users to better understand its contents, its attributes and relational structures. They have not paid enough attention to the human input part which is for instructing, monitoring and guidance in the whole area of direct data manipulation and analytical reasoning. The existing research works that have been in relation to the visual human input part have mainly focused on low-level zooming and navigation operations and they have not addressed the benefit of human involvement in the visual data manipulation and visual analytical reasoning processes. Up until now there has

* Corresponding author at: School of Computer Software Tianjin University, China 92 Weijin Rd., Nankai District, Tianjin, China.
E-mail address: Mao.Huang@uts.edu.au (M.L. Huang).

http://dx.doi.org/10.1016/j.future.2015.02.003
0167-739X/

been no standard visual interaction model which has been widely recognized and there have been many different definitions and interpretations of visual interaction in the visualization community, in terms of its goals, objectives and functionalities. In 2007, J.S. Yi et al. [1] proposed a complete seven-layer visual interaction model based on the existing context of interaction techniques in visualization. These layers are defined as below:

1. Select: mark something as interesting.
2. Explore: show me something else.
3. Reconfigure: show me a different arrangement.
4. Encode: show me a different representation.
5. Abstract/Elaborate: show me more or less detail.
6. Filter: show me something conditionally
7. Connect: show me related items.

This seven-layer model describes the existing visual interactions from the lower level to the higher level, where the 'select' operation is used for highlighting and manipulating a particular data item through the visualization. The 'explore' operation on the other hand is used to find out user-interested data items though visual navigation of the data source. The layers 3–5 are concerned with the design of views of display (visualization) for better understanding and highlighting one (or more) portion(s) (or pattern) of the visualization that are currently interest the user. The last two layers use the 'filtering' mechanism to display (or visualize) only the interesting or related data items in the visualization and remove other less interesting and related data items from the visualization.

To better understand and interpret the above layered structure of visual interactions, we would like to propose a new model by merging Yi's seven layers into four layers, based on the natural use of the operations and to classify them as follows:

1. Selection (or Locating) of Data Items.
2. Dynamic Viewing of Data (Structures).
3. Dynamic Scoping of Data (Structures).
4. Dynamic Scoping and Viewing of Data (Structures).

Selection (or Locating) interaction, this is similar to *Select* interaction; layer 1, defined by J.S. Yi, et al. [1], which provides users with the ability to lock up a particular data item(s) of interest for keeping track or data access (or data retrieval). When too many data items are presented on a view, it is difficult for users to follow items of interest. By visually highlighting items of interest, users can easily keep track of them even in large displays of hundreds of items. This type of the interaction does not require the change of views and data scopes. As shown in Fig. 1, we can select the JGAA home page by highlighting a rectangular node labeled 'JGAA_paper' in the graph visualization through a mouse rollover interaction.

Dynamic Viewing (DV) interaction, this merges layers 3, 4 and 5: Reconfigure, Encode and Abstract/Elaborate of the interaction defined in J.S. Yi's model [1], thereby allowing users to change the data representation for achieving better readability or understanding of the data and its relational structures. Examples include the reordering of axes in Parallel Coordinators visualization and navigation in graph visualization by using a Hyperbolic Tree [2] or a Fish Eye Browser [3]. DV interaction also includes the change of visual encoding; that is to use an alternative visualization method to present the same complete set of data.

Dynamic Scoping (DS) interaction, this merges layers 2, 6 and 7: Explore, Filter and Connect of the interaction defined in J.S. Yi's model [1], which allows users to visualize a portion of the complete data set through the filtering of less important, less interesting and lower relevance data items. Examples include the navigation method used in DA-TU: an On-Line Visualization [4] system and in other dynamic visualization techniques.

Dynamic Scoping and Viewing of Data (DSV) interaction, this allows users to dynamically select a portion of the complete dataset for visualization, and at the same time it also allows users to dynamically change the views, representation styles or representation arrangements for better understanding of the data.

1.1. Interactions in parallel coordinates

Big data is a collection of large and complex datasets that commonly appear in multidimensional and multivariate data formats that become difficult to mine and present meaning knowledge owing to their volume and complexity (e.g. its multivariate). Thus, there is an urgent need to find more effective techniques to deal with such huge data sets.

Parallel coordinates [5] is a common way of visualizing and analyzing multidimensional datasets. It has been extensively studied for decades. To show a set of (value) points in an n-dimensional space, a backdrop is drawn consisting of N parallel lines, typically vertical and equally spaced. A (value)point in n-dimensional space is represented as a *polyline* with vertices on the parallel axes; the position of the vertex on the ith axis corresponds to the attribute value in ith coordinate. An example of the parallel coordinate visualization is illustrated in Fig. 2.

In the literature review of the parallel coordinates visualization, we also see that there is a variety of associated interaction techniques that have been developed to support CHI through this type of the geometry (or visual interfaces). The current available interaction methods in parallel coordinates can be classified into two main categories namely, (1) *brushing*, and (2) *filtering*.

1. *Brushing*—In 1994, Ward [7] first proposed the concept of 'n-dimensional brushing' that can be used to highlight the n-dimensional data items which fall within a user-specified subspace (or sub-region) in either scatterplots or parallel-coordinates geometry. By using brushing interaction, a subset of data items (polylines) within specified value ranges of one or more dimensions can be highlighted (or focused) for viewing the structure of data patterns. This allows users to gain insights into spatial relationships of the n-dimension. Lately, several alternative brushing methods have been proposed in parallel coordinates visualization. For example, Hauser et al. [8] in 2002 proposed the concept of angular brushing as an extension of Ward's standard brushing to facilitate the data subsets grouping and highlighting by using angular constraint. Both techniques integrated the composite brushing and focus + context technique to further improve the visual exploration in parallel coordinates.

In 2003, Yang et al. [9] contributed an automatic and manual brushing mechanism to the parallel coordinate geometry called Visual Hierarchical Dimensionality Reduction (VHDR). VHDR has been integrated into XmdvTool [7] since version 6.0 which is a well-known visualization toolkit. The interaction technique offered by Yang is capable of exploring a big dataset in a more interactive manner with greater flexibility to dynamically change the view. VHDR first constructs a hierarchy of a dimensional tree grouped by similarities and further allows the user to interactively select the interested dimensional cluster for analysis. The hierarchical approach is the commonality between our work and Yang's work, but VHDR targets the level of selective detail at the dimensional scope (vectors) whereas we provide the hierarchical interaction at the data scope through the virtual nodes (data points within a vector).

In 2008, Zhou et al. [10] extended the 'brushing' method for visual clustering in parallel coordinates geometry that allows users to specify the transfer functions in order to control the density value of the lines using alpha blending. They consider the visual clustering as a variation of brushing. One significant drawback of applying alpha blending in brushing is the fact that the low density patterns will become illegible due to high transparency. This may not be desirable if the low density pattern is statistically significant.

Overall, while the 'brushing' technique performed well in highlighting (focusing) and grouping a subset of polylines (or data items) for visual analysis in parallel coordinates geometry, it could not help users to select (or locate) a particular data item (one polyline). Therefore, it covers the functionalities from layers 2–5 of Yi's