University of Victoria | Graduate Studies

Notice of the Final Oral Examination for the Degree of Doctor of Philosophy

of

ALI SARVGHAD BATN MOGHADDAM

MSc (University of Malaya, 2008)
BSc (University Science Malaysia, 20068)

"Tracking and Visualizing Dimension Space Coverage for Exploratory Data Analysis"

Department of Computer Science

Thursday, July 20, 2016
9:00 A.M.
David Turpin Building
Room A144

Supervisory Committee:
Dr. Melanie Tory, Department of Computer Science, University of Victoria (Supervisor)
Dr. Yvonne Coady, Department of Computer Science, UVic (Member)
Dr. Valerie Irvine, Department of Curriculum and Instruction, UVic (Outside Member)

External Examiner:
Dr. Wesley Willett, Department of Computer Science, University of Calgary

Chair of Oral Examination:
Dr. Bob Kowalewski, Department of Physics and Astronomy, UVic

Dr. David Capson, Dean, Faculty of Graduate Studies

Abstract

In this dissertation, I investigate interactive visual history for collaborative exploratory data analysis (EDA). In particular, I examine use of analysis history for improving the awareness of the dimension space coverage 1 2 3 to better support data exploration. Commonly, interactive history tools facilitate data analysis by capturing and representing information about the analysis process. These tools can support a wide range of use-cases from simple undo and redo to complete reconstructions of the visualization pipeline. In the context of exploratory collaborative Visual Analytics (VA), history tools are commonly used for reviewing and reusing past states/actions and do not efficiently support other use-cases such as understanding the past analysis from the angle of dimension space coverage. However, such knowledge is essential for exploratory analysis which requires constant formulation of new questions about data. To carry out exploration, an analyst needs to understand "what has been done" versus "what is remaining" to explore. Lack of such insight can result in premature fixation on certain questions, compromising the coverage of the data set and breadth of exploration [76]. In addition, exploration of large data sets sometimes requires collaboration between a group of analysts who might be in different time/location settings. In this case, in addition to personal analysis history, each team member needs to understand what aspects of the problem his or her collaborators have explored. Such scenarios are common in domains such as science and business [32] where analysts explore large multi-dimensional data sets in search of relationships, patterns and trends. Currently, analysts typically rely on memory and/or externalization to keep track of investigated versus uninvestigated aspects of the problem. Although analysis history 4 mechanisms have the potential to assist analyst(s) with this problem, most common visual representations of history are geared towards reviewing & reusing the visualization pipeline or visualization states.

I started this research with an observational user study to gain a better understanding of analysts' history needs in the context of collaborative exploratory VA. This study showed that understanding the coverage of dimension space by using linear history 5 was cumbersome and inefficient. To address this problem, I investigated how alternate visual representations of analysis history could support this use-case. First, I designed and evaluated Footprint-I, a visual history tool that represented analysis from the angle of dimension space coverage (i.e. history of investigation of data dimensions; specifically, this approach revealed which dimensions had been previously investigated and in which combinations). I performed a user study that evaluated participants' ability to recall the scope of past analysis using my proposed design versus a linear representation of analysis history. I measured participants' task duration and accuracy in

answering questions about a past exploratory VA session. Findings of this study showed that participants with access to dimension space coverage information were both faster and more accurate in understanding dimension space coverage information. Next, I studied the effects of providing coverage information on collaboration. To investigate this question, I designed and implemented Footprint-II, the next version of Footprint-I. In this version, I redesigned the representation of dimension space coverage to be more usable and scalable. I conducted a user study that measured the effects of presenting history from the angle of dimension space coverage on task coordination (tacit breakdown of a common task between collaborators). I asked each participant to assume the role of a business data analyst and continue a exploratory analysis work which was started by a collaborator. The results of this study showed that providing dimension space coverage information helped participants to focus on dimensions that were not investigated in the initial analysis, hence improving tacit task coordination. Finally, I investigated the effects of providing live dimension space coverage information on VA outcomes. To this end, I designed and implemented a standalone prototype VA tool with a visual history module. I used scented widgets [72] to incorporate real-time dimension space coverage information into the GUI widgets. Results of a user study showed that providing live dimension space coverage information increased the number of top-level findings. Moreover, it expanded the breadth of exploration (without compromising the depth) and helped analysts to formulate and ask more questions about their data.

---

[1]In this dissertation, a dimension refers to a column in a tabular dataset where dimension's name is the column's header name. For instance, a financial dataset may include dimensions such as Sales, Profit and Inventory Cost.

[2]I define and use dimension space to refer to the set of all dimensions in a tabular dataset.

[3]I define dimension space coverage as the set of investigated data dimensions, either individually (e.g. a histogram showing distribution of Sales values) or collectively (e.g. a bar chart showing averages of Sales and Profit for different States).

[4]In the context of visual data analysis, history is usually comprised of recorded information about the analysis states/processes and/or outcomes. For more detailed description see Chapter 2.