Task Analysis for Information Visualization

Stacie L. Hibino

Bell Labs, Lucent Technologies
263 Shuman Boulevard, Naperville, IL 60566 USA
hibino@research.bell-labs.com, http://www.bell-labs.com/~hibino/

Abstract. Previous research in information visualization has primarily focused on providing novel views and frameworks to aid users in *exploring or accessing* data; very little work has been done to support users through the *full analysis process*—from the raw data to the final results. But what tasks *do* users perform when analyzing data using an information visualization (infoVis) environment? A task analysis of experts' use of an existing infoVis system was conducted to examine this question. Results indicate that users work on various tasks *outside* of data exploration—tasks such as conditioning and preparing data, collecting results, and gathering evidence for a presentation. This pilot study identifies key data analysis tasks that expert users perform when using an infoVis environment to analyze some real-life data.

1 Introduction

Recent advances in information visualization (infoVis) have led to novel visualizations and new paradigms to support users in accessing and exploring a wide variety of data. General infoVis frameworks [1, 13] and applications to fields such as temporal [8], software [9] and medical [10] data are just a few examples of the growing work in infoVis. Most of this work, however, has focused on enhancing data access and exploration; little, if any work has examined the larger problem of supporting users' data analysis process—the processes they use in transforming raw data into a presentation of results based on using infoVis as their key analysis tool.

The continual introduction of more powerful computers and the recent explosion in large data repositories offers new opportunities while posing new challenges to the infoVis community. In our research lab, where we have already developed a suite of infoVis tools capable of handling moderately sized data sets (easily accommodating data sets containing 100,000 data records), we are beginning to see evidence of these challenges as users complain about the effort required to prepare data for exploration or abstract key results from an infoVis analysis of a complex data set. Previous work in information workspaces [3, 7] offers glimpses of an organizing framework, but much of this work focuses on organizing *data and objects* (e.g., documents) rather than a user's *process*.

In order to understand how to better support users in the infoVis analysis process, we set out to identify the types of tasks they conduct during data analysis with an infoVis tool. In this paper, we describe the task analysis conducted to accomplish this goal. We had three driving questions in this endeavor:

- Do users conduct other tasks besides data exploration during data analysis?
- If so, what are these tasks?
- How important are these tasks to the data analysis process?

The task analysis was conducted on five infoVis experts using an existing infoVis environment (EDV: the Exploratory Data Visualizer [13]). Users were asked to analyze a disease data set used as part of an American Statistical Association (ASA) Data Exposition. Their goal at the end of their analysis sessions was to present key results of their analysis in their typical target presentation format.

Overview. This paper is divided into four additional sections. In the next section, I describe the experimental method. I then present and discuss the results, summarize some related work and finally, I provide conclusions and describe some future work.

2 Experimental Method

The goal of the task analysis was to study expert users analyzing real-life data with an existing infoVis system (EDV [13]) as their primary analysis tool—where users started with the raw data, analyzed it, and worked towards a presentation of results. In order to examine how users cope with a complex data analysis problem requiring multiple sessions, the users were asked to accomplish this over two separate sessions.

Participants. Five users (all male) participated in the study. They were invited to participate based on their expertise in using EDV as well as their experience in using information visualization (infoVis) environments for data analysis. Two users have a statistics background, while the other three hold advanced computer science degrees.

EDV. The Exploratory Data Visualizer (EDV) is an information visualization framework which provides linked data views [13]. EDV allows users to create and interact with any number of dynamically linked, user-specified data views including univariate (e.g., bar charts and histograms), bivariate (e.g., scatterplots and table views), multivariate, network, and text-based views (e.g., similar to a spreadsheet). Views are dynamically linked to one another so that a selection made in one view results in automatic highlighting of the corresponding selections in all other views.

Data Set. Users analyzed real-life tuberculosis (TB) disease data from the 1991 American Statistical Association Data Exposition [5]. The TB data included one primary file of over 11,000 records and five other supporting data files on information such as census data.

Procedure. The users' goal was to analyze the TB data and create a presentation of key results. They were given two 1-hour sessions, separated in time by at least one week, to accomplish this goal. They used EDV as their primary tool for analyzing the data, and were told that they could use any other tools that they typically use in conjunction with EDV to accomplish the analysis task. In the first session, users read a sheet describing this task analysis study and their goals, and a hard copy of the readme file included with the TB data set. They were asked to provide think-aloud verbal protocols of what they were doing as they were doing it, and then proceeded with the analysis task at hand. At the end of the second session, if extra time was available, a short informal post-interview was conducted. Information about the post-questionnaire was sent to users after they had completed their second session.

Data Collected. The following data was collected: observational data, files generated during the users' sessions (e.g., scripts written to transform the data, new files of transformed data, etc.), post-interview notes, and results of the post questionnaire. In addition, output of the users' screen was captured directly onto video, along with audio of their verbal protocols.

3 Results and Discussion

While users did take different approaches to their analyses, their core set of tasks and findings were very similar. However, none of them created an actual presentation of their results. Although they analyzed a real-life data set, they were not motivated to create a presentation and in most cases did not feel that they had enough time to do so. Instead, they either captured their key findings on paper or articulated them aloud during their analysis. Information about presentation-related tasks was gathered through informal post-interviews and through the post-questionnaire.

3.1 High-Level Analysis Tasks

During the task analysis, we observed several low-level user tasks. Through grouping similar tasks together, we identified seven categories of high-level tasks:

- *prepare:* data background and preparation tasks,
- *plan:* analysis planning and/or strategizing tasks,
- *explore:* data exploration tasks,
- *present:* presentation-related tasks,
- *overlay:* overlay and assessment tasks,
- *re-orient:* re-orientation tasks (when analysis requires more than one session), and
- *other:* statistics-based tasks.

The low-level tasks for each of the task categories are described in Section 3.2. When asked, users did not present an alternative list or any additional categories of high-level tasks. Users rated the importance of each of the low-level tasks to the analysis process based on a scale of 1 to 5. Figure 1 shows the average importance ratings for each of the high-level task categories, based on an average of all of the low-level task ratings in each category. Overall, the ratings are fairly high, with averages ranging from 3.4 to 4.5 on a 5-point scale. However, taking individual users into account, an analysis of variance indicates that the differences in importance ratings between categories is significant, leading to the following order of importance based on user ratings: plan > explore > prepare > present > statistics > overlay > re-orient tasks.

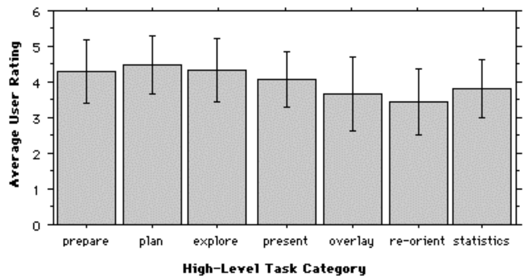

Fig. 1. Average importance ratings for task categories (1=unimportant, 5=very important; error bars indicate standard deviation)

3.2 Low-Level Analysis Tasks

The low-level tasks were identified through observational data, previous infoVis taxonomies [12], informal interviews and through the post-questionnaire. Tables 1 to 6 summarize the low-level tasks by category and include user examples. While there is some overlap between tasks and task categories, these summaries provide a working taxonomy of data analysis through information visualization.

Table 1. Prepare: data background and preparation tasks

Task Description	Example
Gather background information about data set at hand	Review TB `readme` file
Understand data sources	Note TB data file names, sizes, formats
Get clarification on data ambiguities	How is "race" different from "ethnicity?"
Collect additional data from other external sources	Can I get my almanac?
Reformat data for suitable input	Add header information to a data file for importing into EDV
Check data for potential data errors	Spot check raw data file
Check for missing data	Is there any missing data?
Transform the data	Split variables, rollup/aggregate data

Table 2. Plan: analysis planning and strategizing tasks

Task Description	Example
Hypothesize	There was a hypothesis that TB incidence increased in HIV infected groups
Make a strategy or plan for all or a part of your analysis	Decide what, how, and how much to investigate or explore
Identify data formats and variables required for desired views	We need sums of census data by state…

Table 3. Explore: data exploration tasks (incorporating tasks from [12])

Task Description	Example
Get an overview of the data	I always like to use something to get some idea of the whole data set…
Investigate data to test hypotheses (top down approach)	There was a hypothesis that…. We can look at that actually
Explore data in search of trends or exceptions (bottom-up approach)	Now let's look at [the] race [variable]
"Query" or filter the database	We go to race=2 [African American], we see that they get TB around …
Identify curiosities to investigate further	So that's sort of interesting… I wonder, let's…
Zoom in on items of interest	Let's look at [just] that peak of youngsters
Remove uninteresting items	I'm just going to eliminate those early years; concentrate on data where there at least seems to be stable reporting going on
Identify data clusters	Alright, let's try clustering…cluster view
Identify relationships between variables	So in terms of age, whites seem to get TB more when they're older in comparison to the other races…
Explain view/visualization	There are two possible answers [explanations] here. One is that…
Identify a trend or exception	An interesting gap in the [age] data here… There's a gap in around 12 year olds.
Verify a trend or exception	Examine alternative view to verify a trend
Drill-down for more details	[looking at text records] You can actually count the number of 15 year olds...

Table 4. Present: presentation-related tasks

Task Description	Example
Gather evidence to answer a hypothesis or driving question	Well, in terms of the urban area hypothesis, it looks like it might be reasonably, likely; District of Columbia, which is an urban area …
Record or keep track of trends and results tested and found	No significant effect per time-of-year
Articulate importance of a result (rank it or identify it as "interesting")	How does this result rank in comparison to the others?
Articulate/summarize all and/or key results	I'm going to present a summary of my results in written form
Decide what to include in presentation of results	What are the top 2-3 interesting results I want to show?
Create presentation of results	Paste screen dumps into an electronic presentation and annotate
Give presentation of results	Communicate results to others

Table 5. Overlay: overlay and assessment tasks

Task Description	Example
Take notes	Write down code info: 1=male; 2=female
Window management	Move and resize windows
Assess your strategy	Is this the right strategy to take?
Assess your observations	Does this observation make sense?
Assess your assumptions about data formats	Is my data in the right format to accomplish this part of the analysis?
Assess your progress	I'm wondering if there's anything else I haven't considered which I should look at
Estimate cost-benefit ratio of additional data collection or conditioning	I should really recode those… but I couldn't be bothered with that

Table 6. Re-Orient: re-orientation tasks

Task Description	Example
Review goal(s),	Review TB `readme` file
Review data and formats,	`head tb.txt`
Review notes	Flip through written notes
Review progress	What I remember doing last time was…
Identify starting point for current session	So what I wanted to try doing today was…

Due to space limitations, we cannot discuss each of the low-level tasks in detail, but we do highlight some of the more interesting observations here. A couple of interesting data exploration tasks include explaining a visualization and verifying a trend or exception. These tasks, which have not previously been reported in other infoVis taxonomies, indicate that expert users do not just stop at the identification of a trend or exception, and that they consider their analysis to be as much of an *investigation* as an exploration.

Different users have different types of target presentations. Users in this study listed a variety of typical target presentations ranging from static screen dumps to interactive web pages [6] and live EDV demonstrations. For small data sets and in situations where only a few results are identified, presentations are much easier to create and require fewer tasks. Large complex data sets, however, require additional work to keep track of, rank, and decide on presentation contents of results. This is especially the case when users may be sorting through a series of 20 to 30 results.

The take notes and window management tasks are overlay tasks that cut across the other task categories. For example, users could take notes while preparing data, planning their analysis, or exploring the data. The assessment tasks identify the metacognitive activity exhibited by users during their analysis sessions. That is, users asked themselves the types of sample questions listed with each task in Table 5.

Statistics-Based Tasks. Although users did not conduct any statistical tests during their analysis of the TB data, several users mentioned situations where they either would follow-up with a statistical test or where they thought it would be nice to have system support to conduct a particular statistical test.

Additional Tasks. Three users listed five tasks between them that they felt were not included in the task list presented in the post-questionnaire. Three of the five tasks listed were very similar to existing tasks while the other two included:

- given a relationship or feature of interest, explore what other factors may contribute to it, and
- sort and order results so that related results and their impact on each other can easily be accomplished.

4 Related Work

Several infoVis taxonomies have been proposed (e.g., [12, 2]), but these typically focus on categorizing aspects of infoVis limited to accessing and exploring data—aspects such as data types, visualization types and exploration tasks. The results of the task analysis presented in this paper indicates that such taxonomies only address a part of the problem, especially when considering the use of infoVis for data *analysis* rather than only data *access*.

While no infoVis environment currently addresses all of the types of tasks identified through this task analysis, some work has touched on some of the issues identified here. For example, the SAGE system [11] is a knowledge-based presentation system that potentially reduces the users' task load on analysis planning as well as presentation-related tasks. A second example is an infoVis spreadsheet [4] that provides a framework for organizing data conditioning as well as data exploration based on graphical transformations.

Information workspaces (e.g., [3, 7]) have typically focused on organizing data rather than processes. However, one can imagine using a rooms [7] or book metaphor [3] for organizing an infoVis analysis. For example, rooms or books could be used as logical separators for the different types of tasks (e.g., data preparation, analysis planning, data exploration, etc.) or they could be used to separate the analysis along themes or threads (e.g., separate rooms could be dedicated to investigations of different hypotheses). The challenge in using either of these metaphors, however, is in understanding how, if, and when process support can be bridged across rooms.

5 Conclusion and Future Work

Users do perform many other tasks beyond data exploration when using an infoVis environment for analyzing data. In this pilot study, I identified six other categories of analysis tasks besides data exploration: data background and preparation, analysis planning and strategizing, presentation-related, overlay and assessment, re-orientation, and statistics-based tasks. Moreover, not only do users conduct these other tasks, they also rate them highly in terms of their importance to the analysis process. In particular, they rate planning and strategizing tasks significantly higher than exploration tasks.

We are currently in the process of performing a detailed analysis of the video data and verbal protocols to identify how often users performed the various tasks as well as how much time they spent on each of them. In the mean time, we note that users indicated on the post-questionnaire that they typically spend, on average, about 25% of their analysis time on data exploration and at most 40% of their time; thereby spending over half of their analysis time on tasks *other than data exploration*.

In the future, we plan to investigate and prioritize the importance of system support of each of the tasks. Our long-term goal is to work towards a more integrated infoVis framework, one that provides better support to users through the full data analysis process.

Acknowledgments. Special thanks to expert users who participated in the study, to Graham Wills for EDV, and to Beki Grinter and Ken Cox for reviewing earlier drafts of this paper.

References

1. Ahlberg, C., & Shneiderman, B. (1994). Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *CHI'94 Conf. Proc*. ACM Press, 313-317.
2. Card, S. and J. Mackinlay. (1997). The Structure of the Information Visualization Design Space. *IEEE Proceedings of Information Visualization'97*, 92-99.
3. Card, S., Robertson, G., and W. York. (1996). The WebBook and the Web Forager: an information workspace for the World-Wide Web. *CHI '96. Conf. Proceedings*, 111-119.
4. Chi, E.H., Riedl, J., Barry, P., and J. Konstan. (1998). Principles for Information Visualization Spreadsheets. *IEEE Computer Graphics & Applications, 18(4)*, 30-38.
5. 1991 ASA Data Exposition, Disease Data. Available at: http://www.stat.cmu.edu/disease/.
6. Eick, S., Mockus, A., Graves, T. and Karr, A. (1998). A Web Laboratory for Software Data Analysis. *World Wide Web Journal, 12*, 55-60.
7. Henderson, J. & S. Card. (1986). Rooms: The use of multiple virtual workspaces to reduce space contention in window-based graphical user interfaces. *ACM Transactions on Graphics, 5(3),* 211-241.
8. Hibino, S. and Rundensteiner, E. (1996). MMVIS: Design and Implementation of a Multimedia Visual Information Seeking Environment. *ACM Multimedia'96 Conf. Proc.* NY:ACM Press, 75-86.
9. Jerding, D.F., Stasko, J.T. and Ball, T. (1997). Visualizing interactions in program executions. *ICSE'97 Conference Proceedings.* NY:ACM Press, 360-370.
10. North, C., Shneiderman, B. and Plaisant, C. (1997). Visual Information Seeking in Digital Image Libraries: The Visible Human Explorer. *Information in Images* (G. Becker, Ed.), Thomson Technology Labs (http://www.thomtech.com/mmedia/tmr97/chap4.htm).
11. Roth, S., Kolojejchick, Mattis, J. and J. Goldstein. Interactive graphic design using automatic presentation knowledge. *CHI'94 Conference Proceedings*, 318-322.
12. Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *IEEE Proceedings of Visual Languages 1996*, 336-343.
13. Wills, G. (1995). Visual Exploration of Large Structured Datasets. *New Techniques and Trends in Statistics*. IOS Press, 237-246.