Analytic Trails: Supporting Provenance, Collaboration, and Reuse for Visual Data Analysis by Business Users

Jie Lu, Zhen Wen, Shimei Pan, Jennifer Lai

IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY, 10532, USA
{jielu, zhenwen, shimei, jlai}@us.ibm.com

Abstract. In this paper, we discuss the use of analytic trails to support the needs of business users when conducting visual data analysis, focusing particularly on the aspects of analytic provenance, asynchronous collaboration, and reuse of analyses. We present a prototype implementation of analytic trail technology as part of *Smarter Decisions* – a web-based visual analytic tool, with the goal of helping business users derive insights from structured and unstructured data. To understand the value and shortcomings of trails in supporting visual analytic tasks in business environments, we performed a user study with 21 participants. While the majority of participants found trails to be useful for capturing and understanding the provenance of an analysis, they viewed trails as more valuable for personal use rather than for communicating the analytic process to other people as part of a collaboration. Study results also indicate that rich search mechanisms for easily finding relevant trails (or portions of a trail) is critical to the successful adaptation and reuse of existing saved trails.

Keywords: Information visualization, Visual data analysis, Analytic provenance, Asynchronous collaboration, Analysis reuse.

1 Introduction

It is becoming increasingly common for business workers to need to analyze large amounts of data in order to derive the insights necessary for business decisions. Finding an effective way to turn data overload into information that can be used to make decisions quickly has become a high priority [22]. As a result, data analytic tools, particularly those that provide the ability to visualize data with charts, graphs, and maps (i.e. "visual analytic tools"), have attracted increasing attention in recent years [18, 26].

Despite this growing use and acceptance of visual data analysis, several problems exist with current visual analytic tools in business environments. First, the highly interactive and exploratory nature of visual analytic activities often makes it difficult for the user to capture the steps and metadata of the analytic process which are needed to facilitate effective re-visitation [10]. Without adequate support for capturing and retracing the provenance of an analytic process it is difficult and time-consuming to reconstruct or understand how a particular insight was discovered or why a decision was made. Second, with the rise of business globalization, people working together on

a task are often separated by time and distance, requiring them to work asynchronously. It is a challenge to use today's tools for effective collaborative visual analysis [14]. Third, the data sets that a business worker needs to analyze at different times often come from the same domain (e.g. sales figures) and require similar types of analysis. Because current tools provide limited support for reusing or adapting pre-existing analyses, the user mostly has to start from scratch each time s/he analyzes a new data set.

In an attempt to address these problems, we have developed the analytic trail technology as part of *Smarter Decisions*, an interactive web-based visual analytic tool built to enable users who are not visualization experts to interact visually with both structured (e.g. relational database, spreadsheets) and unstructured (e.g. paragraphs of text, blogs, articles) data. This technology automatically captures trails of the analytic steps taken by the user during visual data exploration and displays them as an interactive GUI component. Such trails can create a "corporate memory" of the decisions that were made. They can be rolled back at any time to view each step of the analysis, thereby increasing the transparency of decision making (e.g. who made the decision and why). Trails can be shared, allowing teams to collaborate in decision making. Saved trails can also be used as template or model to facilitate new analysis based on existing stored trails. Our goal in developing this technology was to provide support for three key needs of visual data analysis in business environments: analytic provenance, asynchronous collaboration, and analysis reuse.

To evaluate the effectiveness of the analytic trail technology at supporting this goal, we conducted a user study (N=21) of analytic trails as part of the *Smarter Decisions* tool. The study results help to shed light on situations where trails can provide the greatest benefits as well as the design considerations required to achieve these benefits.

In the following sections of the paper, we begin by describing related work. We then present insights gained from semi-structured interviews with six business analysts about their process and requirements when analyzing business data visually, which informed the design of the analytic trail technology. Next, we provide an overview of the *Smarter Decisions* tool, followed by detailed description of the analytic trail technology. Finally, we present results from the user study, and conclude with a discussion of the findings and directions for future work.

2 Related Work

Our work is related to several areas of research including visual analytic provenance, reuse, and asynchronous collaborative visual analytics. In this section, we review key papers in these areas upon which our research builds.

2.1 Analytic Provenance and Reuse

Research has shown that preserving a historical record of visual analytic activities (i.e. provenance) is an important requirement in many visual analytic applications [20, 25]. To capture visual analysis history, researchers have explored the use of various

history models, visual representations, and operations. Graph-based [16, 23] and tree-based [1, 2] history models have been developed for capturing complex non-linear analysis history. Taxonomies and classification schemes have been proposed to categorize actions in visual analysis [7, 10, 13, 27, 28]. Depending on the underlying history model, both non-linear [5, 17, 23, 24] and linear visual representations [13] have been used to visualize the history. Moreover, a set of operations (e.g. navigate, edit, search, annotate) have been supported to allow users to exploit the recorded history for re-visitation or reuse [1, 7, 11, 13, 15, 17, 21, 24].

The two pieces of work most closely related to our research are the graphical history tool for the Tableau database visualization system [13], and Aruvi, a prototype information visualization system developed for supporting the analytical reasoning process [24]. The Tableau graphical history tool [13] records user actions and visualization states as items that can be bookmarked, annotated, revisited, and exported. It was primarily designed to support re-visitation and communication of individual visualizations. Aruvi [24] captures the visualization states of the analytic process and presents them using a horizontal-vertical tree layout. The granularity of the history tracking was determined by application-specific heuristics (e.g., when the mouse pointer leaves a specific GUI panel). Its goal was to provide a high-level overview of all the exploration paths taken and to allow users to navigate back to any previous visualization state during the current analysis. By contrast, our analytic trail technology captures and allows for bookmarking of an entire analytic process (i.e., not just the final visualization state but all the steps that went into its derivation) and re-visitation of the process at any time. Furthermore, our trails can be edited to facilitate the re-purposing of existing analyses to new analytic tasks.

2.2 Asynchronous Collaborative Visual Analytics

Researchers have studied designs to help users collaborate on visual analysis. Sense.us [14] and Many Eyes [4] are web sites that support asynchronous collaboration across a variety of visualization types through view sharing, discussion, graphical annotation, and social navigation. The grid-based web portal described in [15] allows asynchronous users to view, edit, and extend previous visual exploration sessions conducted by other users. Further design considerations for collaborative visual analytics are discussed in [12] and [3].

Existing techniques mostly focus on collaboration by means of static visualization snapshots (e.g. [4, 14]) or spreadsheets of visualization parameters (e.g. [15]). By contrast, our analytic trail technology allows the whole sequence of visual analytic activities encapsulated in a trail to be shared all together at once, and enables users to dynamically interact with, modify, or extend such a trail.

3 Business User Interviews

To inform the design of the analytic trail technology, we conducted semi-structured interviews with six business analysts, whose daily responsibilities include analyzing data visually to derive insights and make business decisions and/or recommendations.

All of the interviewees are considered experts in their respective domains, which include market research for emerging technologies; business unit market analysis; marketing consultation; strategic planning for sales & distribution; software mergers & acquisitions; and financial performance analysis. All analysts work for large, global enterprises. Although the areas of work are diverse, and the active life span of an analytic task ranges from hours to months, our interviews uncovered several common characteristics in how these business users perform their daily analytic tasks.

All six analysts use Microsoft Excel, especially the charting mechanisms within Excel for their data analytic tasks. Except for one user who took extensive training courses in Excel, the other analysts received little to no formal training and largely rely on self-training. Four analysts use additional internal or commercial business intelligence tools to aggregate the data retrieved from a data warehouse in order to generate data sets of a size manageable by Excel. Half of the interviewees analyze both structured, quantitative data and unstructured, qualitative data obtained from multiple sources, while the other half work only with structured data. Those who work with unstructured data manually add structure to the data by annotating and categorizing the textual content so that they can work with the data in Excel. In all cases, visualizations are used not only for detecting trends and outliers during analysis, but also for communicating the findings and the derived insights to their clients, colleagues, and management chain. Only a small number of visualization metaphors (e.g. bar chart, line graph, pie chart) are commonly used, especially when communicating analysis results, due to their simplicity and ease of being understood by business people.

Below we organize our interview findings as they relate to three key aspects that we focus on for data analysis in business environments: analytic provenance, collaboration, and reuse of pre-existing analyses.

3.1 Analytic Provenance

Analytic provenance refers to a historical record of an analytic process, which may include user analytic activities, the data being explored, as well as the insights uncovered during the analysis. All six business users preserve the insights derived from an analysis by manually associating notes and annotations with visualizations. However, this form of provenance is not always sufficient when the analysis needs to be revisited for various reasons. The majority of the users (four out of six) have a need to revisit an earlier analysis, sometimes conducted weeks or even months before, either to obtain the rationale of a decision/recommendation made based on the analysis, to refresh the analysis with updated data, or to document the steps taken in obtaining the final results. To recall the process of an earlier analysis, these users rely on their own memory or brief notes, which are often unreliable, especially for analytic processes that are "*explorative*," "*iterative*," and use "*a trial-and-error approach*." As a result, the analysts often have to manually redo the whole analysis. For these analysts re-visitation of prior analyses is a fairly common business activity, and the tools they are using do not provide sufficient support for this task.

3.2 Collaboration

All of the business users we interviewed team up with other people to perform data analysis for internal or external clients. As a result, they have a constant need for communicating their work with their colleagues. Because many of their colleagues work in different geographic locations and time zones, the collaboration is mostly asynchronous. They currently rely primarily on e-mail, for sending around copies of notes, tables, spreadsheets and in some cases PowerPoint presentations. One user explicitly expressed the desire for a tool to help him more easily share findings of visual analyses with his colleagues. He considered it "*a wasteful business process*" to "*export snapshots, cut and paste screenshots of Excel dashboard*," and felt that "*some sort of collaborative tool would be helpful for discussion of data during staff-to-staff interaction*."

3.3 Reuse of Analyses

Although each analytic task conducted by an analyst is usually with new/different data, the data sets often share many similarities. For example, the data sets with marketing information of different countries are likely to have a common list of marketing channels such as television, radio, newspaper, and magazine, as well as similar metrics to measure the effectiveness of the marketing activities through these channels. Similarly, for mergers & acquisitions, the data sets usually contain a common list of entities for company profiles, such as business focus, customers, partners, revenue, size, etc. As a result, the users can often transfer what they have learned and used in previous analyses to a new analysis. However, due to the lack of "replay" support that is easy to customize and use without requiring programming skills, the users often have to manually repeat each of the analytic steps they would like to reuse. When asked about the most difficult, tedious, frustrating, or unpleasant part of their work, five of the six analysts mentioned that they didn't like spending time conducting steps that were almost (but not exactly) the same at different times or across different data sets. An example mentioned was manually mapping the data to the visualization parameters or performing the same type of analysis for different companies or multiple geographic areas. In an attempt to address this issue, three users mentioned that they used Excel spreadsheets created for earlier analyses as "templates," and pasted new data on top of old data in the spreadsheets, so that previously defined functions and visual mappings from the data to the visualization parameters could be reused. When asked about why they didn't use Excel macros for their tasks, the analysts pointed to the lack of skills for creating and customizing macros as the main reason. Representative responses include "*I am not very good at writing macros*," "*I wish I had other people create macros for my purposes, but unfortunately we don't*."

We also discovered during the interviews that reuse was not restricted to a user's own analyses. One analyst mentioned that she sometimes studied the reports created by other people for their analyses to learn new ways of analyzing data in Excel, and applied them to her own analyses.

4 Smarter Decisions

In this section, we provide a brief overview of *Smarter Decisions*, a visual analytic tool within which the analytic trail technology was implemented. *Smarter Decisions* is an interactive web-based tool for visual analytics designed to help business users derive insights from large collections of both structured and unstructured data. Fig. 1 shows a screenshot of its main user interface for data analysis. The left hand side of the screen is the query panel where users can build a query using select-lists to retrieve the data (Fig. 1a). The middle portion of the screen is the visualization canvas where the retrieved data is visualized (Fig. 1b). Users explore and analyze data by issuing ad-hoc queries and interacting with the visualizations of the retrieved data (e.g. panning, sorting, filtering). *Smarter Decisions* currently includes several commonly used visualization metaphors such as bar chart, line graph, scatter plot, table, document list, and tag cloud. Based on the technology described in [8], *Smarter Decisions* assists users by automatically instantiating the data in the most appropriate visualization metaphor given the properties of the data, and provides alternate visualization choices on the right hand side of the screen (Fig. 1c). Users can switch to any of these alternates simply by clicking on them.

Fig. 1. Smarter Decisions user interface: (a) query panel, (b) visualization canvas, (c) alternate visualization choices, (d) thumbnail of the visualization for a step, (e) trail steps, (f) menu of operations for a step, (g) detail of the action performed during a step, (h) undo, (i) snapshot, (j) bookmark.

Smarter Decisions automatically captures the trail of the user actions taken during visual data exploration, such as issuing a query (Query), interacting with the visualization to filter to a subset of the data (Filter), changing to an alternate visualization (Change view), and displays the trail at the bottom of the screen (Fig. 1e, see Section 5.2 for a detailed description). The trail technology is the focus of this paper and is described in detail in the next section. Trails can be bookmarked and restored to replay the actions and data that went into the analytic and decision process, essentially creating a retraceable "memory" of what was done. Trails can be shared to allow for asynchronous collaboration and they can be modified and applied in a new analysis thus facilitating the reuse and/or sharing of an established method for analyzing a given data set. It is our belief that saved trails could also be used to assist with skill ramp-up, when a person is new to the department or organization, or for transfer of expertise when the expert is no longer available.

5 Analytic Trails

The analytic trail technology adopted the trail concept and the semantics-based action taxonomy [10] conceptualized in the HARVEST proof-of-concept system [9]. The design of the trail model, representation, and operations was inspired by prior research as discussed in Section 2, and informed by the interview findings described in Section 3. Compared with HARVEST and other visual analytic tools, the goal of the trail technology in *Smarter Decisions* is to provide an integrated solution to support the needs of visual data analysis in business environments, which include not only personal re-visitation and reuse, but also decision auditing, remote team collaboration, and expertise transfer. In this section, we describe the design considerations for building the trail model, its GUI representation, and the operations it supports to achieve the above goal.

5.1 Trail Model

The trail model defines the representation and organization of analytic provenance. To increase the transparency of provenance, user activities need to be recorded at a semantically meaningful level, easily understood by users. Because *Smarter Decisions* was developed as a visual analytic tool that could be used by average business users in different data domains, we adopted the semantics-based model proposed in [10] to capture the analytic process at a semantic level without using domain-specific heuristics. Low-level user interaction events such as clicks and drags are mapped to a set of semantic but generic user *actions* such as Query and Filter (see [10] for details), which are used as semantic building blocks for the trail. Each action includes a set of parameters to encode the information needed by the system to perform the action, such as the data set, data concepts/attributes, and data constraints. The system also dynamically maintains a summary of the user's task context, which is computed by aggregating the parameters of all the previous actions taken in the user's current line of inquiry. This summary provides the contextual information needed to execute the next action and transform the analytic process from one state to another.

A linear, logical sequence of user actions constitutes an *analytic trail*. A *trail graph* interconnects multiple trails to reflect a non-linear, progressive visual analysis workflow. Trails are connected when the user returns to an earlier state of a trail and creates a new branch of analysis from this state to result in another trail.

Based on the interview finding that the users often annotate visualizations for insight provenance, we added to the trail model a feature that allows text annotations to be associated with individual visualizations created during the analysis. The trail model is also equipped with access control to provide users the flexibility of making their analyses private/public, or sharing a trail with a group of collaborators, which enables the tool to support both personal and collaborative analytic tasks.

5.2 Trail Representation

A trail is represented as a linear sequence of steps. We decided to use a representation of the *active trail* to expose the trail model through the user interface. The active trail includes the sequence of user actions performed by a user during his/her current investigational thread. Bookmarked trails from earlier investigational threads are accessible from a trail library. The decision to only display the active trail was based on two main considerations. First, exposing the whole graph structure of the trail model would occupy too much screen space, distracting users from their primary visual analytic task. In contrast, a linear representation is compact and less obtrusive. Second, a graph-based display increases the complexity and difficulty of trail presentation and interaction, which may cause user confusion about how to interpret the display and how to interact with it. By comparison, a linear display is simpler and easier to understand, which reduces training time and potential cognitive burden on users when using the tool.

Each step in the trail consists of a semantic user action defined in the trail model (e.g. Query, Filter), and the visualization displayed as a result of the action. Information about each step is displayed on two levels. At the higher level, a step is depicted as a button with an icon and a text label to indicate the type of action performed at that step (Fig. 1e). Hovering the mouse over a step shows a small thumbnail of the associated visualization for the step (Fig. 1d). This high-level display enables users to quickly obtain an "at-a-glance" pictorial summary of the analytic process. At a more detailed level, clicking on a step reveals information about the action performed during the step in the form of parameter name-value pairs (Fig. 1g), and a menu of the operations that can be performed in this step (Fig. 1f, which we describe in the next section). The reason to use parameters instead of a natural language-style summary for describing a user action is two-fold: 1) to avoid any misinterpretation caused by ambiguities in natural language, and 2) to allow users to more easily modify the action to reuse its logic within a new context.

5.3 Trail Operations

Smarter Decisions supports operations both at the trail level and at the level of individual trail steps. At the trail level, users can click the bookmark button located at

the bottom right of the interface (Fig. 1j) to save the sequence of actions included in the current trail. A unique URL is assigned to each bookmarked trail. Clicking on the URL (e.g., within an email, a blog, or the trail library) results in the trail being restored within the *Smarter Decisions* interface, which replaces any existing trail display at the interface. Once a trail is restored, it becomes the current active trail, which means that it is fully interactive and can also be extended with new user actions to continue the analysis. This mechanism enables users to work collaboratively on an analysis and to adapt an existing trail for new analysis. By default, a bookmarked trail is private so that only its creator is able to access and restore it. A user can change a bookmarked trail s/he created from private to public to allow any other user to access it. Alternatively a bookmarked trail can be shared with specific users, identified by name. At any time, the creator of a trail can delete it from the trail library.

At the level of individual trail steps, users can perform operations such as removing the step from the active trail ("Delete"), removing all the subsequent steps of this step from the active trail ("Undo to here"), and revisiting the step ("Revisit this step") by selecting from the menu associated with each trail step (Fig. 1f). Single and multiple step deletions enable users to remove unwanted actions, especially those performed during the exploration phase of an analysis, and to keep a record of only the sequence that leads to the desired analysis outcome. For convenience, a single-step undo button is displayed at the end of the representation of the active trail (Fig. 1h). Clicking on it results in the last step being deleted. Revisiting a step restores the application to the state that was reached as a result of the user action recorded in that step. The restored information includes an aggregation of the parameters of all the actions performed up to this point of the trail, the visualization displayed, and any user-provided annotations associated with the visualization. Step re-visitation provides a mechanism for users to quickly examine an earlier state of an analysis for understanding the logic and reviewing the result. If a new user action is performed as part of revisiting a previous step (e.g., the parameters are changed, or a filter is applied to the visualized data), a new investigational thread is started, which creates a new trail in the trail graph by branching out from the current active trail. Then the newly created trail becomes the new active trail. This mechanism provides users with the capability of reusing the same steps of a recorded analytic trail for a new analysis without having to manually repeat them. For example, if the user who performed the analysis shown in Fig. 1 wants to investigate the mortgage market as discussed in the articles that mention the state of California, he can reuse the query step in the existing analysis by revisiting this step and execute a filter to just the mortgage market.

The parameter values of categorical, numerical, or keyword constraints for any trail step are editable (Fig. 1g), allowing users to apply the corresponding user action in a different but related context for new analysis. This functionality was designed to enable an analysis to be adapted for use in similar but different analytic tasks. For example, the user can adapt the analysis shown in Fig. 1 for a new investigation in another state, e.g., Texas, by changing the value of the constraint (Fig. 1g) from "California" to "Texas."

Finally, a snapshot button (Fig. 1i), located at the bottom right of the interface next to the bookmark button (Fig. 1j), enables users to export the current visualization to an image that can be embedded in reports and presentations.

6 User Study

We conducted a user study with two primary goals: 1) to evaluate the quality of the support provided by the trail technology with regard to our three focus areas of analytic provenance, asynchronous collaboration, and reuse of analyses, and 2) to gather user feedback on how the design could be improved to better assist users with their visual analytic tasks.

6.1 Study Design

We evaluated the analytic trail technology in the context of the *Smarter Decisions* visual analytic tool. The objective of the user study was *not* to see if use of *Smarter Decisions* with analytic trails was better or faster than a baseline condition, but to examine to what extent the system met (or failed to meet) the needs of business users (e.g. provenance, collaboration, and reuse) and to what degree the features of trails were discoverable by users with little to no prior training. We also felt that a set of baseline metrics would have been difficult to generalize since most of this user population use Excel as a tool, and have no current equivalent to trails.

The data set used for the study was created from the InfoVis publication data [6], which was chosen because it is publicly available and the concepts in this data set are easily understood by the users that would participate in the study. It contains the metadata from 614 papers published between 1974 and 2004, including the title, authors, abstract, topic, references, length, source, and year for each paper.

There were a total of four tasks in the study. The first two tasks were designed to evaluate the tool's support for analytic provenance by asking the users to validate a set of statements based on two existing trails, one for each task. The statements were about research topics, authors, and citations of the InfoVis publications contained in the data set. For example, one of the statements for task 1 was "There are four researchers in total who have published papers on the topic of 'visualizing large data sets' based on this data set," and a statement from task 2 was "The paper titled 'The information visualizer, an information workspace' is the most cited paper in the data set." Task 1 also contained false statements, while all of the statements provided for task 2 were true. Each trail consisted of five or six steps. The users were encouraged to inspect the trails, but were not allowed to extend or modify them.

Task 3 was designed to evaluate the tool's support for cases where multiple people collaborate asynchronously to complete an analysis. Specifically, the users were asked to complete a partially finished analysis by extending an analytic trail which had been started by an imaginary colleague. This colleague had selected a topic for analysis ("dynamic queries"), identified the two researchers with the most papers on this topic, and the papers that each of them published on this topic. The study participants were asked to continue the analysis in order to determine the number of papers each of these two researchers had published across all topics, as well as naming two other topics each researcher had been working on.

Finally, task 4 focused on analysis reuse. For this task participants were asked to conduct a new analysis by either reusing and modifying any trail from the trail library, or starting the analysis from scratch. More specifically, participants were asked to

conduct an analysis of the papers published in 1996 to find answers to questions about papers on the topic of "internet" and with the keyword "representation" in their titles or abstracts. A trail that recorded an analysis of the papers published in 1995 with different topic and keyword constraints was included in the trail library, along with trails used for the other tasks of the study.

The trails used for the tasks mostly consisted of Query and Filter steps. The Query steps were used to obtain aggregated (e.g. count) or detailed (e.g. topic, title, author) information of the InfoVis publication data given specific constraints (e.g. a particular topic or year of publication). The Filter steps were used to apply additional constraints to drill down to a subset of the requested data.

It should be noted that the tasks described above didn't cover all forms of use cases for visual data analysis, in particular the synthesis case that brings together multiple threads of analysis conducted by multiple users. Such a use case is often limited to deep analysis conducted by government agencies or scientists. Business users have (relatively) simpler questions that need answering. We believe that the use cases in the study (e.g., auditing an earlier analysis, continuing an analysis started by a colleague, finding an existing analysis from the library to reuse) are representative of the visual data analysis tasks conducted in business environments.

Prior to the main evaluation, we tested the design of the study including the tutorial and tasks with two users and made revisions accordingly. For the main study, we recruited twenty-one users from a large corporate research firm, sixteen of whom were male and five were female. Their ages varied from mid 20s to early 50s, with an average age of mid 30s. All of the users had some experience with general tools that include visualizations (e.g. Google Maps, PowerPoint charts), but were not visualization experts. Before performing the study tasks, each user was given a tutorial on the *Smarter Decisions* tool, during which s/he was instructed to interact with a trail made available for training purposes. All participants received identical training, with the training material being read by the experimenter to ensure conformity. After the training session the participants were given time to ask as many questions as they wanted to ensure understanding. For each task in the study, the users were given the task description and the questions they needed to answer in print form. Access to the trail library containing the trails needed for the tasks was included as part of the user interface for the tool.

We recorded the task completion time for each task, which was counted from the time when the user started interacting with *Smarter Decisions* to the moment when s/he indicated that s/he had finished answering all questions for the task. We collected subjective feedback from the users through questionnaires at the end of every task. The questionnaires included a set of questions for which the answers were measured using a Likert scale that ranged from 1 for "strongly disagree" to 7 for "strongly agree" (see Table 1), and open-ended questions such as "what was difficult about using trails" to further collect user comments and suggestions.

6.2 Study Results

In this section we report on objective and subjective data collected from the study. Section 7 provides further discussion of the study findings and their implications.

The average training session across all the users lasted 8.89 minutes. 81% of the users spent less than 10 minutes on training (the variable being the number of follow-up questions each participant asked). One user required over 20 minutes of training, in which he asked many detailed questions while exploring various components of the interface. By comparison, the other users asked fewer questions and largely familiarized themselves with the interface during task performance.

In spite of this very brief training time, all participants were able to complete their tasks. The longest study completion time was 40.07 minutes, which was about 10 minutes per task. We observed during the study that longer training mostly yielded shorter total task completion times. Fig. 2 depicts the relation between training time and total completion time for each participant. The correlation coefficient between the two time variables is –0.50 with a *p*-value of 0.02 (–0.67, *p*-value < 0.001 if excluding the participant mentioned above with the longest training time), indicating a weak negative linear correlation.

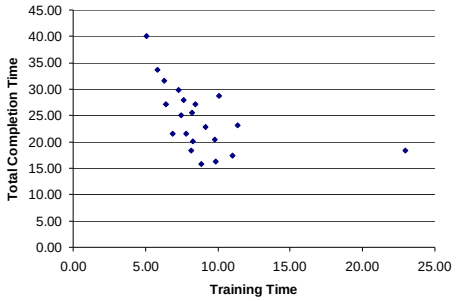

Fig. 2. The relations between the training time and the total completion time (in minutes).

Among the four tasks, task 1 and task 2 required similar completion times (mean 5.79, std. dev. 2.20 vs. mean 5.21, std. dev. 1.75 in minutes for tasks 1 and 2 respectively, with no statistically significant difference between them) since they were similar to each other by design (i.e., validating three statements about paper authors or citations based on an existing trail). Task 3 took the longest time to complete (mean 9.48, std. dev. 2.57 in minutes). This was because in comparison with the other tasks, each user needed to perform more analytic steps and answer more questions in order to complete task 3. Task 4 contained the least number of questions, yielding the shortest completion time (mean 3.91, std. dev. 1.87 in minutes).

Fig. 3 shows the subjective data measured using a Likert scale. The results indicate that the users on average felt positive about being able to easily understand and use the concept of analytic trails and associated steps for quickly understanding and validating analysis results (Q1-Q4). Two users explicitly pointed out that they liked being able to validate analysis results by revisiting the trail saved for the analysis so they didn't need to start from scratch. 71% of the users responded positively that trails would improve the transparency of decision making if used in their teams or in their company (Q5). Also 19% of the users felt that they could see the potential usefulness of trails in increasing the transparency of decision making, but since they didn't do visual data analysis as part of their current assignment for the work, they chose to remain neutral on this subject.

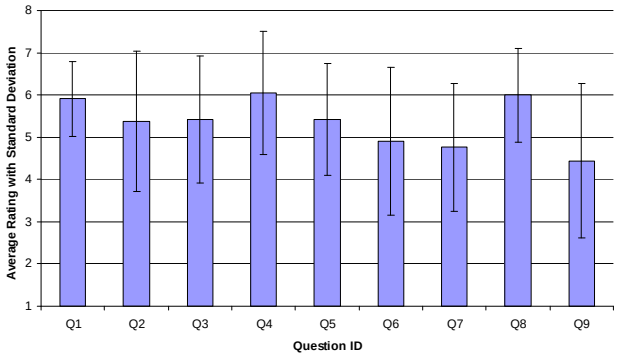

Fig. 3. The average ratings (with +/– 1 std. dev. as error bars) for the Likert-scale questions (1: strongly disagree; 2: disagree; 3: somewhat disagree; 4: undecided; 5: somewhat agree; 6: agree; 7: strongly agree).

Table 1. The Likert-scale questions included in the questionnaires following the tasks.

ID	Question
Q1	Easy to understand the concept of trail and associated steps
Q2	Easy to understand a previous analysis based on its trail
Q3	Easy to validate analysis results by revisiting trail steps
Q4	Faster to validate analysis results using trails than using other sources
Q5	Trails would improve the transparency of decision making
Q6	Easy to determine how to extend an trail to complete the analysis
Q7	Helpful to be able to extend a saved trail to complete an analysis
Q8	Helpful to adapt an existing trail and re-apply it to a new analysis
Q9	Easy to find a trail most relevant to task at hand in the library

Compared with the almost universally positive opinions about the usefulness of trails for understanding the provenance of existing analyses (task 1 and task 2), the users were less enthusiastic about extending a saved trail to complete an analysis started by someone else (task 3). Only 52% of the users agreed or strongly agreed that it was easy for them to determine how to extend a trail to complete the analysis (Q6). Three users who expressed negative opinions on this subject wanted to distinguish their new analytic steps from existing ones, but the current design lacked support for this feature. Only 38% of the users agreed or strongly agreed that they found it helpful to be able to extend a saved trail to complete an analysis (Q7). Four users commented that they preferred to follow their own logic and thought process to perform an analysis, and didn't like to start from a trail created by somebody else.

For task 4, one third of the users chose to start from scratch rather than reusing a trail from the trail library, citing the relative ease of performing a new analysis to complete the task, and the estimated degree of difficulty in finding a relevant trail from the trail library as the two main reasons for such a decision. Among the participants who selected a trail from the trail library to reuse and adapt it to complete the required analysis, on average they agreed that it was helpful to adapt an existing trail and re-apply it to a new analysis (Q8). However, only 43% of them successfully

found the most relevant trail for the task from the trail library, and agreed or strongly agreed that they could easily find a trail in the library most relevant to their task at hand (Q9). The other 57% reused a sub-optimal trail instead.

For the open-ended questions, when asked about what was easy about using trails, the users mentioned that the high-level information of a trail was easy to understand, the details of trail steps were easy to access and understand, and a trail was easy to navigate. These aspects made it easy to revisit and follow the logical path of an analysis and validate the results along the trail. Regarding what was hard about using trails, there were three common complaints. First, seven users felt that the high-level information about trail steps (i.e., descriptions of action types in the step buttons and small thumbnails of visualizations when the step buttons are moused over, as illustrated in Fig. 1d-e) was too abstract and desired more details at a glance. Second, seven users expected the visualization associated with a step to be shown by clicking on the button of this step instead of being required to select the "Revisit this step" option from the step's menu. Third, because the current design of trail representation only displays one active trail at a time, three users felt that it would be difficult to use trails for analyses involving multiple active investigational threads in parallel.

7 Discussion

We developed the analytic trail technology for our *Smarter Decisions* visual analytic tool with the goal of increasing the transparency of analytic provenance as well as supporting asynchronous collaboration and reuse of visual data analyses. During the user study, we received valuable feedback on the effectiveness of our development and how the technology could be further improved to achieve this goal. Here we discuss the feedback and the remaining challenges.

7.1 Analytic Provenance

Study results indicate that the users were receptive of the analytic trail technology and positive about the value of trails for increasing the transparency of analytic provenance. Trails were shown to be effective in helping the users understand the logic of existing analyses as needed to validate the results/statements generated. However, the feedback from the users suggests that the current design of trail GUI representation could be improved to strengthen the benefits of trails for analytic provenance. Particularly, several users commented that the design of the representation at the trail level needs to provide an effective summary for them to quickly understand the analytic process. In some cases, using action type descriptions (e.g. Query, Filter) and small thumbnails of visualizations to describe trail steps was not sufficiently informative for the users to quickly understand the analytic process that was undertaken. The users were required to perform multiple mouse clicks to get to the details of the trail steps one step at a time, and rely on their memory to piece together the logic of the analytic process. In other cases, multiple semantic actions might correspond to one logical action in a user's thought process, but individual steps displayed in a linear sequence didn't reflect the logical relation or groupings of

the steps. As a result, for a long trail with a large number of steps, the trail representation became too low-level, making it difficult for the users to grasp the logical flow of the analysis. It is a challenge to find a single level of granularity for trail representation that works well in all the cases, especially without the help of domain heuristics.

A better solution may be to dynamically adjust the granularity of trail representation based on the characteristics of the analysis. For example, when the analysis includes a small number of steps with simple logic, details of the steps can be made visible at the trail level. For a complex analysis with many steps, individual steps that correspond to one logical construct of the analysis can be "chunked" together. Such logical chunking can be performed manually by the user who conducts the analysis, or automatically by the system based on machine learning and mining from user analytic behaviors. User-provided descriptions can be associated with each chunk to improve understanding. This solution requires the trail model to be augmented to support hierarchical organization of actions, and interaction mechanisms to support multi-level zoom in/out for the trail display at the interface.

In addition to determining the right level of granularity intelligently for trail representation, the trail technology should also support more intuitive interaction mechanisms for trail step re-visitation (e.g., clicking on the button of a step to revisit instead of selecting the revisit option from the step's menu).

7.2 Collaboration

Interestingly, the use of trails in asynchronous collaboration around visual data analysis was not as well-received as we had hoped. The log and questionnaires from the study revealed two primary explanations for the relatively lukewarm response to this feature. First, the participants were not sufficiently motivated to understand all the details of an analysis conducted by their (imaginary) collaborator in order to complete the required task (task 3). This could be due to the fact that the task was conducted in the laboratory setting with an imaginary collaborator instead of the users' actual working environment with real persons as collaborators. Some users felt that they only needed to know the outcome of an analysis and didn't really care about how the analysis was conducted. Similarly, these users only felt compelled to share their analysis results but not the process. One user mentioned that he was not sure if he would ever share the trail of his analysis with others since his collaboration with others were loosely coupled, for which sharing analysis results would be sufficient. Therefore, he would want to save the trail of his analysis for personal use, but not for collaboration. However, the current design of the trail technology makes it difficult for the users to obtain or share the information about "what" (analysis results) without sharing the detail about "how" (analytic process). Second, some users didn't want to follow other people's thought process in order to perform an analysis collaboratively. They preferred following their own logic and didn't want to mix the record of their analytic steps with the record of the collaborators'. For these participants, trails were viewed as more valuable for recording, navigating, and adapting an analytic process for personal use, rather than for communicating with other people as part of a collaboration.

The above findings suggest that the most appropriate structure and granularity of trail representation depend on not only the characteristics of the analysis (e.g. complexity) as discussed in the previous section, but also the purpose for analytic provenance. For example, representation at the level of individual analytic steps may work well for personal use, which includes viewing the details of the analysis, but a level based on logical grouping, or authorship of trail steps, may be more appropriate for loosely-coupled collaboration. Also if the goal of the review of individual steps is to audit the decision, or to increase the transparency of what data was included in a decision, then the details could be made available on demand. Furthermore, the trail representation should make it easy to navigate between multiple related trails created by different users as part of a collaborative analysis.

7.3 Reuse of Analyses

We designed task 4 of the user study to focus on evaluating the effectiveness of trails with regard to facilitating reuse of analyses. We were surprised that as many as seven users didn't even try to browse the trail library to find a trail they could reuse for new analysis. When asked about why they didn't reuse a trail, some users said that they felt they could perform the task easily from scratch without the need to reuse an existing trail. More complex tasks might provide greater incentive for the users to reuse trails of existing analyses, and longer-term use of our tool could reveal greater benefits of trail reuse, which we weren't able to test in the laboratory setting where we limited the complexity of the tasks so as not to overwhelm the study participants, all of whom were first-time users of our tool. Eight users didn't correctly identify the most relevant trail for the task when browsing the small library of six trails. Instead of selecting a well-suited trail we had placed in the library as a public trail, they chose a trail they had used in one of the previous tasks. These users felt that without built-in search support for the trail library, it was difficult to determine the relevance of a trail for a new task based on its brief description in the library, and too time-consuming to examine all the trails to find the most relevant one, especially for trails created by others. They expected this problem to become more serious as the trail library grew. This study result points to the need to provide rich search support for trails in the library (by author, by keyword, by date, etc.), or trail reuse will likely be limited to cases where the users know in advance which trail to reuse.

We also observed that the users who reused a trail for task 4 first spent time inspecting the trail to find the steps relevant to the task at hand, then deleted all of the unwanted steps before modifying and adapting the relevant steps to suit their needs. These users expressed a common desire for the support to easily find the relevant parts of the trail and easily manipulate trail steps, such as moving (e.g. by dragging and dropping) a trail step from one position of the trail to another and making a copy of one or more trail steps. Therefore, in addition to finding a relevant trail, effective reuse of analyses requires system support for easily locating the parts of the trail to be reused and adapting them for new analysis. How to design and implement such new features without making the interface overly complex and reducing its usability is a challenging problem, which is next on our research agenda.

To further support reuse of analyses, we also plan to develop new functionality that generalizes trails to create trail templates that can be easily customized and applied by different people on different data and analytic tasks, with the goal of facilitating the use and sharing of best practices and helping more effective skill/expertise transfer.

8 Conclusion

In this paper, we present the analytic trail technology in the context of a visual analytic tool designed to empower business users to derive insights from large amounts of data. Informed by the findings from the interviews with several business users about their visual analytic activities, we present the design of the trail model, its GUI representation, and the operations it supports with the goal of providing a mechanism to increase the transparency of analytic provenance as well as support asynchronous collaboration and reuse of visual data analyses in business environments. To facilitate analytic provenance, the trail model represents user analytic activities with semantic actions (e.g. Query, Filter, Change view) and captures a linear, logical sequence of actions into a trail. Multiple trails of an analysis are organized into a graph-based structure to reflect a non-linear, progressive visual analysis workflow. The active trail which corresponds to the sequence of actions performed during a user's current investigational thread is displayed at the GUI. The trail GUI representation was designed to help users easily navigate a trail and obtain information about the encapsulated user actions and visualization results for understanding the provenance of the analysis. Trail operations such as bookmarking, sharing, revisiting, and editing are provided with the goal of facilitating re-visitation, asynchronous communication, and reuse of analyses.

We designed and conducted a user study to evaluate the effectiveness of the analytic trail technology at supporting its goal of provenance, reuse and collaboration. The results indicate that most participants found trails to be useful for capturing and understanding the provenance of an analysis. However, with the current design, trails were considered to be more valuable in recording, navigating, and adapting an analytic process for personal use, rather than for communicating the analytic process to other people as part of a collaboration. The results also indicate that search support for easily finding relevant trails or relevant parts of a trail is critical to support the goal of adaptation and reuse of analyses. These findings suggest areas where trails provide the greatest value and point out directions for future research in the area of capturing analytic processes.

References

1. L. Bavoil, S. Callahan, P. Crossno, J. Freire, C. Scheidegger, C. Silva, and H. Vo. Vistrails: Enabling interactive multiple-view visualizations. In *IEEE Vis*, 2005.
2. K. Brodlie, L. Brankin, G. Banecki, A. Gay, A. Poon, and H. Wright. Grasparc – a problem solving environment integrating computation and visualization. In *IEEE Vis*, 1993.

3. N. Chinchor and W. Pike. Science of analytical reporting. In *Information Visualization*, 8:286–293, 2009.
4. C. Danis, F. Viegas, M. Wattenberg, and J. Kriss. Your place or mine? visualization as a community component. In *CHI*, 2008.
5. M. Derthick and S. Roth. Example based generation of custom data analysis appliances. In *IUI*, 2001.
6. J.-D. Fekete, G. Grinstein and C. Plaisant. *IEEE InfoVis* 2004 contest data set. http://www.cs.umd.edu/hcil/iv04contest, 2004.
7. H. Goodell, C. Chiang, C. Kelleher, A. Baumann, and G. Grinstein. Collecting and harnessing rich session histories. In *IV*, 2006.
8. D. Gotz and Z. Wen. Behavior-driven visualization recommendation. In *IUI*, 2009.
9. D. Gotz, Z. Wen, J. Lu, P. Kissa, M. Zhou, N. Cao, W. Qian, and S. Liu. HARVEST – Visualization and analysis for the masses. In *IEEE InfoVis Poster*, 2008.
10. D. Gotz and M. Zhou. Characterizing users' visual analytic activity for insight provenance. In *IEEE VAST*, 2008.
11. D. Groth and K. Streefkerk. Provenance and annotation for visual exploration systems. In *IEEE Trans on Vis. and Comp. Graphics*, 12(6), 2006.
12. J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. In *Information Visualization*, 7(1):49–62, 2008.
13. J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. In *IEEE Trans on Vis. and Comp. Graphics*, 14(6):1189–1196, 2008.
14. J. Heer, F. Viegas, and M. Wattenberg. Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *CHI*, 2007.
15. T. Jankun-Kelly, O. Kreylos, K. Ma, B. Hamann, K. Joy, J. Shalf, and E. Bethel. Deploying web-based visual exploration tools on the grid. In *IEEE Computer Graphics and Applications*, 23(2):40–50, 2003.
16. T. Jankun-Kelly, K. Ma, and M. Gertz. A model for the visualization exploration process. In *IEEE Vis*, 2002.
17. N. Kadivar, V. Chen, D. Dunsmuir, E. Lee, C. Qian, J. Dill, C. Shaw, and R. Woodbury. Capturing and supporting the analysis process. In *IEEE VAST*, 2009.
18. D. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *IEEE InfoVis*, 2006.
19. A. Kobsa. An empirical comparison of three commercial information visualization systems. In *IEEE InfoVis*, 2001.
20. M. Kreuseler, T. Nocke, and H. Schumann. A history mechanism for visual data mining. In *IEEE InfoVis*, 2004.
21. D. Kurlander and S. Feiner. Editable graphical histories. In *IEEE Workshop on Visual Language*, pages 127–134, 1998.
22. M. Lissack. Of chaos and complexity: Managerial insights from a new science. In *Management Decision*, 35:205–218, 1997.
23. K. Ma. Image graphs – a novel approach to visual data exploration. In *IEEE Vis*, 1999.
24. Y. Shrinivasan and J. van Wijk. Supporting the analytical reasoning process in information visualization. In *CHI*, 2008.
25. J. Thomas and K. Cook, editors. Illuminating the path: The research and development agenda for visual analytics. *IEEE Press*, 2005.
26. P. Wong and J. Thomas. Visual analytics. In *IEEE Computer Graphics and Applications*, 24:20–21, 2004.
27. L. Xiao, J. Gerth, and P. Hanrahan. Enhancing visual analysis of network traffic using a knowledge representation. In *IEEE VAST*, 2007.
28. J. Yi, Y. Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. In *IEEE Trans on Vis. and Comp. Graphics*, 13(6), 2007.