Connecting the Dots with Related Notes

Yedendra B. Shrinivasan
Technische Universiteit Eindhoven
Den Dolech 2, HG 6.71
PO Box 513
5600 MB Eindhoven, The Netherlands
y.b.shrinivasan@tue.nl

David Gotz
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532 USA
dgotz@us.ibm.com

CHI 2009, April 4 – 9, 2009, Boston, MA, USA
ACM 978-1-60558-246-7/09/04.

Abstract
During visual analysis, users must often connect insights discovered at various points of time to understand implicit relations within their analysis. This process is often called "connecting the dots." In this paper, we describe an algorithm to recommend related notes from a user' past analysis based on his/her current line of inquiry during an interactive visual exploration process. We have implemented the related notes algorithm in HARVEST, a web based visual analytic system.

Keywords
Related notes, Reasoning process, Information visualization, Visual Analytics

ACM Classification Keywords
H.3.3 Information Search and Retrieval: *Retrieval models.*

Introduction
Interactive visualizations allow users to investigate various characteristics of a dataset and reason based on patterns, trends and outliers. During complex visual analyses, users must derive insights by connecting discoveries made at different stages of an investigation. However, during a long investigation process that can span hours, days or even weeks, it becomes difficult for users to recall the details of their past discoveries. Yet these details may form the key connections between their past work and current line of inquiry. We believe

that a user's limited recollection often leads them to overlook important connections. The challenge, therefore, is to develop techniques that assist in "connecting the dots" by uncovering connections to users' past work that would normally go unnoticed.

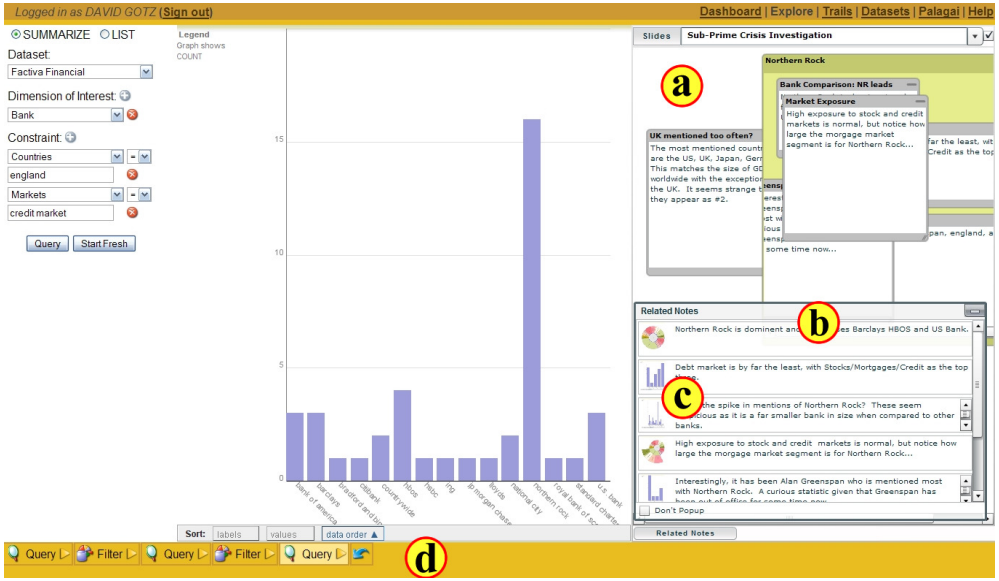

Figure 1. A user investigating a finance dataset in HARVEST, a web based visual analytics system. (a) Palagai, a note taking interface. (b) Related notes. (c) Thumbnail of the visualization displayed while a note was recorded. (d) The user's action trail.

Due to the volume of information discovered during a long analysis task, users often externalize interesting findings or new hypotheses using either annotations on top of visualizations or through bookmarks in electronic notes. These notes help users revisit and review their past analysis. However, as the number of notes and annotations grows larger, users again have difficultly recalling the details of each previous discovery. Therefore, surfacing notes from their past analysis that are relevant to their current line of inquiry can help them find interesting connections within their analysis.

In this paper, we present a related notes recommendation system that retrieves notes from a user's past analysis based on their relevance to his/her current line of inquiry. We first describe how users interact with the related notes recommendation feature in HARVEST, a visual analytic system. We then present the details of our algorithm for recommending related notes to users. As future work we are planning user studies to measure the value of our approach.

Connecting the Dots in HARVEST

HARVEST is a web based visual analytics system that supports exploration of large unstructured datasets. An action tracking mechanism in HARVEST automatically captures user's analysis behavior as an *action trail* [3]. An action trail of a user's exploration process is shown in figure 1(d). Using this action trail interface, users can archive an action trail, as well as revisit and reuse past visualization states. HARVEST also has a note taking interface called *Palagai* that allows users to record notes and organize them into groups and slides (figure 1(a)).

When a user records a note, the Palagai interface augments it with a context model that represents the user's information interests. Then, as the user visually explores the dataset within HARVEST, Palagai dynamically derives a context model for the current line of inquiry and compares it with the context models associated with each of the user's notes. Based on this comparison, Palagai computes a relevance score for

each note and presents a ranked list of related notes to the user (figure 1(b)). A thumbnail of the visualization that was displayed while the user originally recorded each note is also shown (figure 1(c)). With the Palagai interface, users can either explicitly request related notes at anytime or have the system automatically recommend them after each user action.

Our algorithm therefore dynamically surfaces the most relevant notes from earlier stages in an analysis. We believe that this approach helps users maintain awareness of relevant information and assists in connection discovery during the reasoning process.

In the following sections, we describe how we model a user's exploration process within a visual analytics system using action trails. Then we discuss our design considerations for the context model and present how it is used to recommend related notes to the user.

Modeling the Exploration Process in a Visual Analytics System

During exploratory data analysis, analysts must derive insights from large and complex datasets. In visual analysis, users employ interactive visualizations to perceive trends, patterns and outliers as they explore their data. The exploration process is facilitated by a set of actions exposed through the user interface of a visual analysis tool. An action corresponds to an atomic and semantic analytic step performed by an analyst with a visual analytic system [3]. Some examples of actions are filtering, clustering, selecting objects through direct manipulation, sorting, drilling-down, zoom, pan, recording a note, bookmarking, undo-redo, and revisiting a past visualization state.

Gotz and Zhou [3] classify action types in a visual analytic system into three broad categories: exploration actions; insight actions; and meta actions. An exploration action alters the visualization specifications in a visual analytics system and creates a new visualization state. A sequence of exploration actions can lead to new findings or validate existing findings. Users then make use of insight actions to record annotations, bookmark visualizations or organize notes recorded during an analysis. They can make use of meta actions to revisit, undo, redo, delete or edit a past exploration or insight action. These meta actions structure their lines of inquiry. A sequence of exploration, insight and meta actions forms an *action trail*, an activity model that captures a user's analytical behavior within a visual analysis tool. We use action trails to derive context models for notes and visualization states created during a visual analysis.

Context Model

We argue that data characteristics investigated in each exploration action of a user's action trail correspond to the user's information interest. The data characteristics specified through exploration actions are called *action concepts*. For each new visualization state and note created during a visual analysis, a set of related action concepts can be derived from the action trail. We then use this set of related action concepts to represent the context in which the user recorded his/her note. We support our argument for a context model based on action concepts with the following use case.

Use Case

Figure 2 shows a portion of an action trail for an analyst investigating products sales data. She starts her analysis by focusing on sales that are more than

$50000 (figure 2.1). She compares sales of each product using a scatterplot visualization and bookmarks it (figure 2.2). Then, she studies quarterly sales of the products by aggregating the sales represented on the y-axis of the scatterplot based on a quarterly time period (figure 2.3). Next, she uses a treemap to visualize the sale figures in various regions (figure 2.4). Further, she clusters the products by their category to get an overview of the sales performance by product category in various regions (figure 2.5). This view triggers her to reconsider the products sales comparison that she investigated some time back. She therefore revisits the comparison view she bookmarked earlier. Then she narrows down to the east and south regions (figure 2.6). This revisit and reuse of a visualization state creates a branch in her action trail.

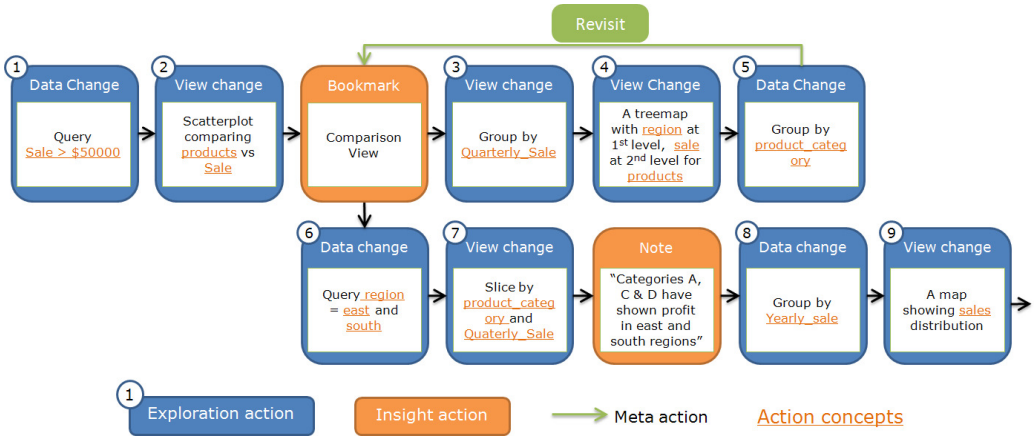

Figure 2. Part of an action trail for an analyst investigating product sales. Exploration actions are represented with a blue box; Insight actions such as bookmarking and note taking are represented using an orange box; Meta actions such as revisit are represented using a green line with an arrow.

She further slices the products in the x-axis of the scatterplot by their category; and slices sales in the y-axis of the scatterplot by quarterly period (figure 2.7). This slicing creates a scatterplot matrix showing sales of various product categories in different quarters of the year. She finds out that product categories A, C and D have shown profit consistently in the east and south regions. She records this finding using a note. Then, she continues her analysis by studying yearly sales (figure 2.8) and sales distribution across regions using a map (figure 2.9).

Related Action Concepts as Context

In the products sales use case, the user started her analysis with general sales data and moved on to investigate quarterly and yearly sales trends. Region was another aspect considered in the investigation; she focused on all regions, then narrowed down to the east and south regions, and finally moved on to see the actual geographical sales distribution. She also investigated the sales of individual products as well as product categories (groups of products).

The action concepts associated with this action trail (e.g., the east region and product category) correspond to the user's information interests. However, some of the action concepts were more predominant at certain times than others. For instance, she was interested only in sales of more than $50000 throughout the investigation. However, she shifted her focus among other action concepts such as quarterly sales, product categories, and regions. These concepts were more predominant during some parts of the analysis than others. This reflects that during an investigation, the relative importance of action concepts can change after each exploration action. Therefore, associated with

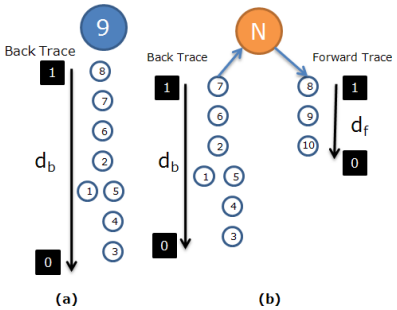

Figure 3. (a) Back trace of exploration actions for the visualization state 9 in figure 2. (b) Back trace and forward trace of exploration actions for the note in figure 2. d_b and d_f are the normalized weight for each exploration action in the back trace and forward trace respectively.

each visualization state and note is a weighted set of related action concepts that correspond to the user's context at the time of discovery.

Importance Score for Action Concepts

In our algorithm, we assign an importance score to each action concept in a user's trail to represent its degree of salience to a specific a visualization state or a note. Our approach is motivated by the spreading-activation construct that is used in many theories for retrieving information from long term memory [1, 2].

In these theories, knowledge is encoded as a network structure, consisting of nodes representing concepts and links representing associations among concepts. During a retrieval process, this network structure is used to identify knowledge relevant to a current focus of attention and facilitate processing of associated items. The two basic points emphasized in these approaches are (1) activation is modeled as a spreading function, and (2) activation decays exponentially with the distance it spreads over a network structure [2]. Following these guidelines, we consider the following aspects to determine the set of related action concepts and their importance score for a visualization state or a note:

Back Trace and Forward trace

We apply spreading activation to our action trail representation to trace related action concepts for a visualization state or a note. A trace spreads through the branching structure of an action trail to reflect that a visualization state or note can be created by a confluence of different lines of inquiry. Figure 3(a) shows a back trace of exploration actions for a visualization state using the structure of the analyst's action trail shown in figure 2.

A user can record a note in three ways. Firstly, she can record a note when a sequence of exploration actions leads to a finding. Hence, a back trace of exploration actions will give related action concepts. Secondly, she can make some assertions or hypotheses which she wants to explore and confirm during an investigation. This note influences subsequent actions. Hence, a forward trace of the exploration actions will give related action concepts. Thirdly, she can collect some relevant information from outside a visual analytics system (e.g., a snippet from the internet). In this case, either a sequence of exploration actions might have triggered her to look for some external information or she may be preparing for an investigation by gathering some external information. Hence, in this case, both back trace and forward trace is required to derive related action concepts (figure 3(b)).

Trace Length

Defining the boundary of a trace is semantic, subjective, and difficult to algorithmically determine from an action trail. Hence we apply a threshold where a trace is extended either until *n* unique action concepts are extracted, or when the start or end of an action trail is reached. After experimenting with various values, we use a threshold of n=10 in our current prototype.

Recency

Proximity of an exploration action to a visualization state or a note in an action trail is used to weigh an action concept. In Figure 3, d_b and d_f are the normalized weight for each exploration action in the

back trace and forward trace respectively based on the length of the trace. This normalization compensates for the variation in length for each trace.

Specificity

During an exploration process, analysts may be focused on all values of an attribute (e.g., sales in all regions) or she may focus on specific values of those attributes (e.g., sales in the east and south regions). Hence, if an action concept references specific values within the dataset, then it is given more weight than those which reference generic characteristics. In our current prototype, specific concepts are given a specificity weight s_j that is twice the weight of generic concepts.

Importance Score

Therefore, based on the factors above, the importance score for an action concept c_i (Ic_i) is as follows:

$$Ic_j = s_j * \left(w_b * \sum_{i=1}^{b} d_i + w_f * \sum_{K=1}^{f} d_k \right),$$

Where s_j is the specificity weight of the action concept c_j; b and f are the length of the back and forward traces respectively; d_i and d_k are the normalized weight based on recency of an exploration action for back trace and forward trace respectively; d_i and $d_k = 0$, if c_j is not specified in an exploration action; w_b and w_f are the weights for back trace and forward trace; w_f is zero during a trace for a visualization state.

Recommending Related Notes

For each note, related action concepts are extracted and an importance score for each action concept is computed based on the structure of the note's action trail. As the exploration process evolves, the set of related action concepts for each note and their importance scores are updated. In addition, after each newly performed user action, a set of related action concepts is extracted from the user's new action trail and their importance score is computed. The relevance score RN_k for every note K in a user's notes is computed as follows:

$$RN_k = \sum_{i=1}^{m} \left(A(Ic_i) * N_k(Ic_i) \right),$$

$N_k(Ic_i) = 0$, when c_i is not a related action concept for Note K.

Based on the relevance scores, a ranked list of related notes for the current line of inquiry is recommended to the user (as shown in figure 1(b)).

Conclusion and Future Work

In this paper, we presented a related notes recommendation system that retrieves notes from a user's past analysis based on their relevance to his/her current line of inquiry during a visual analysis. These related notes help to "connect the dots" by uncovering connections to users' past work that would normally go unnoticed. We have implemented this related notes algorithm in the HARVEST visual analytic system. In future work, we plan to conduct case studies to understand the implications of our approach.

Reference

[1] Anderson, J. R., Pirolli, P. L., 1984. Spread of activation. Journal of Experimental Psychology: Learning, Memory and Cognition 10, 791-798.

[2] Collins, A. M., and Loftus, E. F., 1975. A spreading-activation theory of semantic processing. Psychological Review, 82, 407-428.

[3] Gotz, D. and Zhou, M., 2008. Characterizing Users' Visual Analytic Activity for Insight Provenance. IEEE VAST'08, Columbus, Ohio.