When to Adapt: Detecting User's Confusion During Visualization Processing

Cristina Conati, Enamul Hoque, Dereck Toker, Ben Steichen

Department of Computer Science, University of British Columbia
{conati, enamul, dtoker, steichen}@cs.ubc.ca

Abstract. In this paper, we discuss an approach to collect data on instances of user confusion during visualization processing. The long-term goal is to use this data to train classifiers that can detect instances of user confusion in real time, as triggers for adaptive interventions aimed at alleviating the confusion.

1 Introduction

The benefits of user-adaptive interaction have been shown in a variety of tasks and applications such as operation of menu based interfaces, web search, desktop assistance, and human learning [19]. There are three key decisions that need to be made when designing a user-adaptive system: (1) *what to adapt to*, namely understanding which individual user features should be considered for adaptation, including stable, long-term user traits (e.g., cognitive abilities, expertise, personality), as well as transitory, short-term states (e.g., current task, cognitive load, attention); (2) *when to adapt*, namely understanding when it is appropriate and/or necessary to provide adaptive support to the user, by identifying those situations in which the benefits of providing adaptive interventions outweigh their cost; (3) *how to adapt,* namely understanding how adaptation should be provided.

In this paper, we discuss issues related to the *when to adapt* decision in the context of designing user-adaptive visualizations. While there has been extensive work in investigating how to detect when a user needs help in fields such as Intelligent Tutoring Systems [1] or Adaptive Games [23], this is not the case in visualization. To our knowledge, the work by Gotz & Wen [22] is so far the only one that actively monitors real-time user behavior in order to infer such needs for intervention. In their work, interface action data (e.g., mouse clicks) are constantly tracked in order to detect suboptimal usage patterns. Once these repetitive patterns (determined empirically a priori) are detected, the system then triggers adaptive help.

This approach, however, does not easily transfer to situations in which it is hard to define a priori a set of appropriate interaction behaviors to perform given tasks with a visualization, as well as their suboptimal counterparts. This is the case, for instance, for visualizations that support open ended or exploratory tasks, or when one wants to consider interaction data beyond mouse or keyboard events, such as gaze data. Gaze data has been shown to have a great potential for providing information on a user's task, expertise, and other cognitive measures relevant for adaptation [14, 15], but it is more erratic in nature than interface actions, and it is less well understood in terms of what constitutes a priori effective/ineffective interface actions.

In this paper, we explore an alternative approach that involves collecting ground truth labels for specific salient episodes during interaction with a visualization, that may indicate the need of adaptive interventions. The long-term goal is to use these labels to train classifiers on interaction data consisting of both action logs and eye-tracking data, and to leverage these classifiers to detect in real time, for a new user, when adaptive interventions may be needed.

Collecting ground truth labels for building classifiers on relevant user states or processes can be a challenging endeavour. Here we propose one possible approach to collect labels relevant for building user-adaptive visualizations, which we are currently testing in a user study. However, this paper's primary aim is to open the discussion on the issue of *when to adapt* in user-adaptive visualizations, as opposed to provide well-defined solutions.

In the rest of the paper, we first briefly describe ValueCharts, the visualization that used as a test-bed for this research, and the study we are running to test, among other things, the label collection method. Then, we discuss the labelling approach that we have developed and present some preliminary results on its effectiveness.

2 ValueCharts and User Study

A ValueChart is a set of visualizations and interactive techniques intended to support decision-makers in inspecting linear models of preferences and evaluation [4]. Linear models are popular decision-making tools designed to help the decision-maker perform preferential choice under conflicting objectives, i.e., select the best option out of a set of alternatives. However, as models and their domain of application grow in complexity, model analysis can become a very challenging task. ValueCharts are intended to help decision makers deal with this complexity, based on a design driven by a detailed task model for preferential choice [2]. They have been extensively evaluated and shown to be quite effective [3, 12,17].

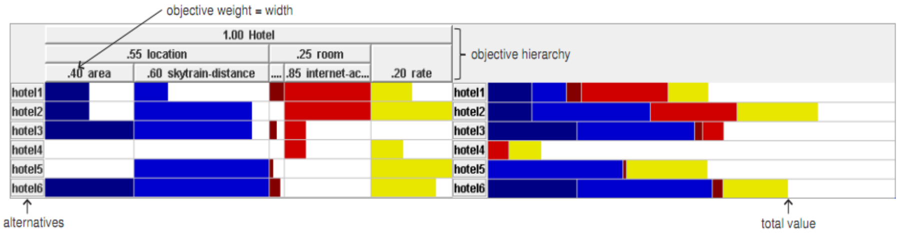


Figure 1: Sample Value Chart

Figure 1 shows an example of Value Chart for the simple preferential choice of selecting an hotel when traveling to a new city, out of six available alternatives. For the sake of simplicity, here we just describe the key features of ValueCharts. The relevant hotel attributes or *objectives* (e.g., area, skytrain distance, internet access, etc.) are arranged hierarchically and are represented in the top-left quadrant of the figure, forming the columns in the ValueChart display. The width of each column indicates the relative *weight* assigned to the corresponding objective (e.g., sky-train distance is much more important than area). The available alternatives (hotels here) are represented as the rows in the display. The cells in each row specify how the corresponding alternative fares with respect to each objective (i.e., the *value* of that objective for that alternative), indicated by the amount of filled color in the cell. So for instance, *hotel1* is far from the sky-train, but it has excellent internet access. In the rightmost quadrant, all values for each alternative are accumulated and presented as horizontal stacked bars, displaying the overall value of each alternative (e.g., in Figure 1, *hotel2* is the best alternative). Several interactive techniques are available in ValueCharts to support the inspection of the preference model. For instance, users can inspect the specific domain value of each objective (e.g., actual distance from the sky-train of *hotel1*); sensitivity analysis of objectives' weight is enabled by allowing the user to change the width of the corresponding column.

In the contest of an on-going project on devising theories and techniques for user adaptive visualizations, we are currently running a user study designed to evaluate the impact of a variety of user traits (e.g., perceptual speed, visual/verbal working memory, visualizations expertise, locus of control, etc.) on the effectiveness of two different versions of ValueCharts. The first version uses an horizontal layout to show the components of the decision making problem (see Figure 1), while the second version displays the same information by using a vertical layout (see Figure 2). We are comparing these two layouts because previous studies with ValueCharts suggest that they may not be equivalent with respect to the user's performance and preference. We test the impact of the aforementioned user traits because they were shown to have an effect during interaction with other visualizations [e.g., 6,16,18].

During the study we conduct, participants use each of the two Value Chart versions in two phases. The first phase (also known as *structured phase*) involves performing a selection of specific tasks in one of four available domains. The tasks are mainly related to retrieving information on the available decision alternatives (e.g., "how far is *hotel1* from the sky-train?", "How many hotels have better internet access than *hotel3*?", "List the 3 highest valued hotels"). The second phase (also known as *open-ended phase*) involves having a participant select a new domain and exploring it until the participant can identify a preferred alternative. Throughout the two phases, we track participants' gaze with a Tobii T120 desktop-mounted eye-tracker, similar to the study described in [16], because that study showed that gaze data can provide useful information on a user's individual differences and on the user's tasks [14, 15].

While one goal of this study is to ascertain whether the tested set of individual differences affect a user's performance with the two ValueChart layouts (i.e., help with the decision of *what to adapt to* in the context of using ValueCharts), we also wanted to leverage the study to provide data toward the question on *when to adapt*. Namely, we wanted to see whether we could find ways to collect information on salient points of the interactions that may benefit from adaptive interventions. The next section describes the approcah that we tested in this study.

3 Collecting labels of user confusion

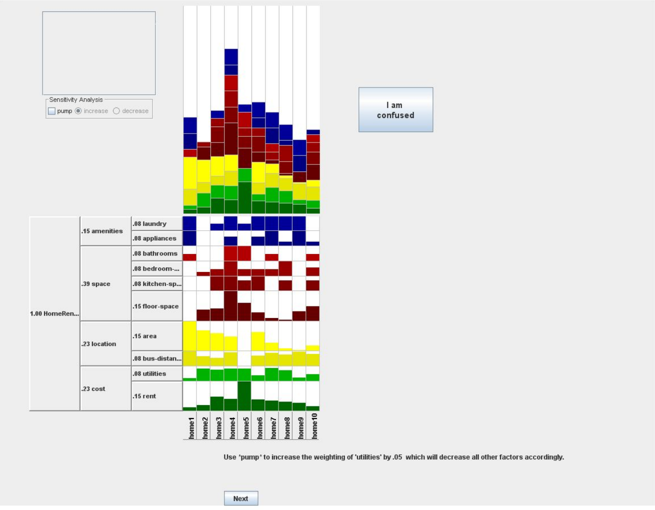
The aim of providing real-time adaptive interventions is to help a user overcome situations that may generate a sub-optimal experience with an interface. For instance, adaptive interventions in an educational application can be generated when the user makes a mistakes or otherwise shows that she is not learning from the interaction [1]. Intuitively, adaptive interventions to improve a user's experience with a visualization would be suitable when the user is not processing the visualization appropriately, for instance when the user is uncertain or confused about where to look or how to interpret the visualization. Thus, in our ValueChart study we tried to devise a way to capture instances of user confusion during interaction, with the long-term goal of building a classifier user model for confusion detection, trained on these instances and on the related action and gaze patterns.

There have been a variety of methods proposed in the literature to capture confusion, mostly in the context of emotion modeling during the interaction with educational software. *Concurrent verbal protocols* involve having participants verbalize their thinking or feelings during interactions [e.g., 13]. We discarded this approach because of existing research indicating that concurrent protocols may alter a user's gaze patterns in unexpected ways, thus generating gaze data not representative of the user's attention patterns during a more naturalistic interaction [e.g., 9] *Retrospective verbal protocols* involve having participants look at a replay of the target interaction and try to verbalize their thinking or feelings at that time. While this approach has shown good results for collecting labels of emotion valence/arousal [11] or on the occurrence of one specific emotion (not including confusion) triggered on purpose via selected movie clips [10], it showed to be quite unreliable when subjects had to identify their naturally occurring emotions (including confusion), during interaction with an educational system [8]. We actually tried this approach in the study described in [16], quite similar in design to the study described here, and also found it inadequate. During pilot phases of that study, we first tried to ask subjects to generate retrospective protocols after each individual task, but because there are many rather short tasks, subjects quickly grew tired and the process interfered with the primary task. We then resorted to ask subjects to generate retrospective protocols at the very end of their study session, but at that point subjects had a difficult time re-generating their relevant states over the course of the complete interaction.

An alternative to verbal protocols is to *obtain labels from subjects via interface input*, e.g., buttons, pop-up windows, or other affordances that allow participants to select the relevant labels when the related episodes occur during interaction. This method has been successfully used to elicit information on user motivation and emotions during interaction with educational systems [7]. However, it has two main drawbacks. The first is that it can be hard to strike a balance between leaving it to the user to provide the information (e.g., via an ever-present interface button), which may results in not collecting a sufficient number of labels, vs forcing the user to provide as many labels as needed (e.g., via pop-up windows that cannot be dismissed), which may disrupt the interaction. The second drawback is that this approach does not provide as much information on the episodes of interest as verbal protocols do[1].

[1] An additional drawback when the user's gaze is tracked is the that presence of the interface affordance that allows for label provision changes the user's gaze patterns that would happen when the affordance is not present. The changes, however, are predictable and can be dealt with during gaze data processing.

As far as the first problem is concerned, in this study we decided to be conservative in terms of intrusiveness and we rely on the user's willingness to provide the labels via a *confusion button* placed on the side of the currently displayed ValueChart (see Figure 2 for an example with a vertical ValueChart). At the beginning of each study session, the experimenter reads the participant a script that includes the following instructions to elicit usage of the confusion button:


Sensitivity Analysis
□ pump ● increase ○ decrease

Use 'pump' to increase the weighting of 'utilities' by .05 which will decrease all other factors accordingly.

Figure 2: Vertical value chart with confusion button

"...We ask that you press this button any time you feel even slightly confused while performing the task. For instance, if you feel that you want to ask the experimenter a question about something, click the confusion button. If you are confused about the interface, click the confusion button, if you are confused about the wording of a question, click the confusion button. If you find a glitch or typo that confuses you, click the confusion button. These are just a few examples, to show that confusion can occur in many unforeseeable ways, and we want consider any type of confusion as being an OK reason to click the confusion button. Note that the system will not be able to give you any help to resolve your confusion. Pressing the confusion button will simply allow us to record moments in which the tasks or the visualizations make you confused. It is very important for the objectives of the experiment that we collect this information, so please take the time to press the confusion button when appropriate."

Preliminary data from the study participants that we ran so far (eight) indicate that the current set up is quite effective in eliciting presses of the confusion button. Five of the eight participants pressed the confusion button at least once. The total number of clicks by the 8 participants is 22, with an average of 2.75. presses per participant (Std dev =3.32). If this trend continues through all of the 30 participants planned to evaluate the impact of individual differences on ValueChart effectiveness, we will run additional subjects with the sole purpose of collecting enough episodes of user confusion for training a classifier that detects confusion from users' action and gaze data

To address the second problem in obtaining labels from user interface input, (i.e., lack of details on what may be causing a user to be confused), we combine the confusion button approach with a form of *focused retrospective verbal protocol*. Namely, we show to each participant video replays of interaction segments centered around presses of the confusion button, and for each of these segments, we ask the participant to explain why the confusion button was pressed. The elicited participant speech is then audio recorded. These verbal protocol sessions happen after each pair of structured and open-ended tasks performed with one of the two ValueChart versions, thus each participant undergoes at most two of these sessions· Pilots of this approach showed that participants are quite capable of generating explanations of confusion that happened during a structured task, after the subsequent open-ended tasks.

We expect that the information collected via focused retrospective verbal protocols will help us qualify the labels obtained via button presses in terms of *the reasons* for confusion, a fundamental piece of information to identify potential adaptive interventions that can help users resolve their confusion. For instance, one of the study's pilot subjects pressed the confusion button 3 times, always during a structured task with a Vertical ValueChart. For the first confusion button event, the participant said that she was confused because of the alternatives' names being vertical, which caused her difficulties in reading them. If this reason for confusion could be automatically identified, it might be alleviated by enabling a functionality that allows the user to mouse over the names to see them displayed more clearly (e.g., horizontally). The two subsequent confusion button events were both explained by the participant as due to difficulty during two different instances of a structured task that requires comparing alternatives with respect to a higher-level dimension in the objective hierarchy (e.g., location or room quality in our hotel selection example). This task requires to visually aggregate the values of the objectives under the target dimension and then comparing them. It could be facilitated, if confusion is detected, by visual props that help identify the aggregated blocks of values and draw the comparison. Another pilot subject clicked the confusion button twice, and for both occurrences the given reason was that the participant had not be able to tell the difference between the overall values of two alternatives, because they were placed in non-contiguous rows and were too similar to tell which one was greater/lower. A suitable adaptive intervention for this type of confusion might be a visual prop that, as before helps draw a comparison, but focusing on discriminating between small differences.

Discussion and Conclusions

In this paper, we discussed an approach to collect data on instances of user confusion during visualization processing, with the long-term goal to use this data to train classifiers that can detect instances of user confusion in real time. There are several open questions on this approach, that we would like to discuss at the workshop including: (i) how much data will be required to reliably identify user's confusion? (ii) Will it be possible to identify a taxonomy of confusion types, along with a mapping between elements of this taxonomy and types of adaptive interventions adequate to alleviate them? (iii) Which other user states in addition to confusion, or which other interaction episodes could serve as triggers for adaptive interventions? (iv) Which other approaches could be explored to collect data on the relevant user states?

References

1. B. P. Woolf. Building Intelligent Interactive Tutors, Student-Centered Strategies for Revolutionizing E-Learning, Published by Elsevier & Morgan Kaufmann, (2008).
2. Bautista J. and Carenini G., An Integrated Task-Based Framework for the Design Evaluation of Visualizations to Support Preferential Choice," Proc. Advanced Visual Interfaces (AVI 2006), (2006), 217-224.
3. Bautista J., Carenini G., An Empirical Evaluation of Interactive Visualization Techniques for Preferential Choice. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, (AVI08), Naples, Italy, (2008).
4. Carenini G. and Loyd J. ValueCharts: Analyzing Linear Models Expressing Preferences and Evaluations, Proc. The International Working Conference on Advanced Visual Interfaces (AVI 2004), (2004), 150-157.
5. Conati C. and Maclaren H.. Modeling User Affect from Causes and Effects. In *Proceedings of UMAP 2009*, First and Seventeenth International Conference on User Modeling, Adaptation and Personalization, Springer, (2008).
6. Conati, C., and Maclaren, H., Exploring the Role of Individual Differences in Information Visualization. In Proceedings of the working conference on Advanced visual interfaces (AVI '08). ACM, New York, NY, USA, (2008), 199-206.
7. deVincente, A., & Pain, H.. Motivation Self-Report in ITS. AIED'99, 9th International Conference on Artificial Intelligence in Education, Le Mans, France, (1999), 651-659.
8. D'Mello, S.K., S. K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., and Graesser, A. C. Automatic Detection of Learner's Affect from Conversational Cues. User Modeling and User-Adapted Interaction, 18(1-2), (2008) , 45-80.
9. Kim B., Dong Y., Kim S., and Lee K. Development of integrated analysis system and tool of perception, recognition, and behavior for web usability test: with emphasis on eye-tracking, mouse-tracking, and retrospective think aloud. In *Proceedings of the 2nd international conference on Usability and internationalization (UI-HCII'07)*. Springer-Verlag, Berlin, Heidelberg, (2007), 113-121.
10. Lisetti, C. L., & Nasoz, F.. Using Non-invasive Wearable Computers to Recognize Human Emotions from Physiological Signals. EURASIP Journal on Applied Signal Processing,11, (2004), 1672-1687.
11. Peter, C., & Herbon, A.. Emotion Representation and Physi-ology Assignments in Digital Systems. Interacting with Computers,18(2), (2006), 139-170.
12. Pommeranz A., Broekens J., Wiggers P., Brinkman W.P., and Jonker C.M., "Designing Interfaces for Explicit Preference Elicitation: a User-Centered Investigation of Preference Representation and Elicitation Process," J. User Modeling and User-Adapted Interaction, vol. 22, no. 4-5, (2012), 357-397.
13. Sidney K. D'Mello, Scotty D. Craig, Jeremiah Sullins, Arthur C. Graesser: Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. I. J. Artificial Intelli-gence in Education 16(1), (2006), 3-28.
14. Steichen, B., Carenini G., and Conati C. Adaptive Information Visualization: Using Gaze Data to Infer Visualization Types, Tasks, and User Characteristics. In

Proceedings of IUI 2013, International Conference on Intelligent User Interfaces, ACM, (to appear), (2013).

15. Toker D., Conati C., Steichen B., and Carenini G.. Individual User Characteristics and Information Visualization: Connecting the Dots through Eye Tracking. In *Proceedings of CHI 2013, ACM SIGCHI Conference on Human Factors in Computing Systems*, (to appear), (2013).
16. Toker, D., Conati, C., Carenini, G., and Haraty, M.. Towards Adaptive Information Visualization: On the Influence of User Characteristics. In *Proceedings of UMAP 2012, the 20th International Conference on User Modeling, Adaptation, and Personalization*. Springer LNCS 7379, (2012), 274-285.
17. Wongsuphasawat K., Plaisant C., Taieb-Maimon M., and Shneiderman B., “Querying Event Sequences by Exact Match or Similarity Search: Design and Empirical Evaluation,” J. Interacting with Computers, 24(2), (2012), 55-68.
18. Ziemkiewicz, C., et al. How Locus of Control Influences Compatibility with Visualization Style. In *Proc. IEEE VAST*, (2011), 81-90.
19. Jameson, A. "Adaptive Interfaces and Agents" in Human-Computer Interface Handbook, eds J.A. Jacko and A. Sears, 305-330, 2003.
20. Grawemeyer, B. Evaluation of ERST – an external representation selection tutor. In Proceedings of the 4th international conference on Diagrammatic Representation and Inference, 154-167, 2006.
21. Green, T. M. & Fisher, B. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In IEEE Visual Analytics Science and Technology, 203-210, 2010.
22. Gotz D., & Wen, Z.. Behavior Driven Visualization Recommendation. ACM Int. Conf. on Intelligent User Interfaces, 315-324, 2009.
23. Peirce, N., Conlan, O., Wade, V. Adaptive Educational Games: Providing Non-invasive Personalised Learning Experiences. In Second IEEE International Conference on Digital Games and Intelligent Toys Based Education, 2008.