On a Mixed-Methods Evaluation of a Social-Agent Scenario Visualization

Thomas A. Alspaugh, Eric Baumer, and Bill Tomlinson
Department of Informatics
University of California, Irvine
{alspaugh,ebaumer,wmt}@ics.uci.edu

Abstract

Scenarios are a well-explored technique for working with and understanding a system's requirements. However, comprehending a large group of scenarios for a system can be difficult, especially for non-experts. Our previous work proposed that visualizing scenarios using social animated characters could assist this process. However, assessing the efficacy of visualization techniques can be challenging. This paper proposes that a mixed-method study combining qualitative and quantitative analysis can be effective for evaluating a social visualization of a group of scenarios. Specifically, we found that the quantitative data addressed focused hypotheses, while the qualitative data gave us insight into the nature of scenarios in requirements, the goals of scenario visualization, and how the technology can support these goals more effectively. Both forms of analysis can be valuable and mutually reinforcing in developing and evaluating effective social visualizations of scenarios, and by extension for other work in RE as well.

1. Introduction

Comparative evaluations in Requirements Engineering have become both more common and more expected in recent years. This trend reflects the growing maturity of the field and also the growing awareness of the value of such evaluations. In this paper we discuss a recent comparative evaluation of an approach for visualizing collections of scenarios using an automatically-produced animation of social interactions among autonomous characters [ATB06]. The evaluation gave us not only results reflecting the effectiveness of the specific technique being studied, but also a deeper understanding of the approach, the problems it is intended to address, and the use of qualitative and mixed-methods approaches for evaluating work in Requirements Engineering.

Scenarios and use cases are widely used in a number of ways during the development process, and by a variety of participants [Ale2004] [BFJZ92]. These participants frequently include stakeholders and users who are not experts in the use and analysis of scenarios. It is essential that these non-experts be able to understand the scenarios of usage that describe a software system so that they can participate fully in the process of defining the system's requirements. Scenarios are an effective communication medium between stakeholders and developers. Their narrative form and use of natural language take advantage of people's natural ability to understand stories. However, scenarios also involve challenges, especially when more than one scenario must be considered at the same time. Other researchers have noted that people are better at identifying errors of commission than errors of omission: they are more successful at finding individual statements that are incorrect, than at identifying missing information, non-local inconsistencies between two or more scenarios, unstated assumptions about the world or the system, or ambiguous or ill-defined terminology. Visually

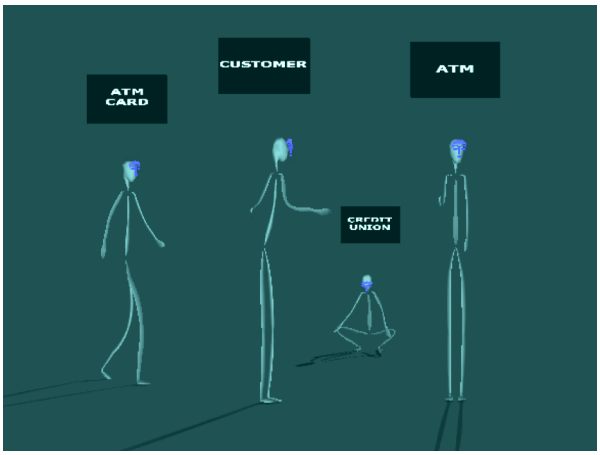

Figure 1: Several social animated characters enact a scenario.

modeling a collection of scenarios as social interactions (see Figure 1) provides an additional path to achieve better understanding of the systems described by those scenarios. If the modeling preserves the interconnections and dependencies of the scenarios, it can take advantage of people's basic competences with social interactions to give viewers insight into the interactions described by the scenarios. The research presented here uses the metaphor of social interaction to improve non-experts' comprehension of sets of scenarios. It is also possible that this technique would be helpful for experts trying to grasp fully the complexities of a large set of scenarios.

The project evaluated in the study combines two main elements: a scenario language with structure, semantics, and automated tool support (ScenarioML [Als06]); and an interactive graphical game engine featuring social autonomous characters and text-to-speech capabilities. The animated social interactions of the autonomous characters and the text that they speak are driven by the collection of scenarios used as input to the visualization.

This mapping from scenarios to animated social interactions results in the creation of an animated character for each actor and entity desired to be shown in the visualization. Each character is accompanied by an identifying label, and actors gesture and speak descriptions of their actions as a means of expressing the enactment of those actions. Interacting characters move to face each other during their interaction. So, for example, when software component A sends a message to component B in the scenario, the visualization would entail three animated characters – labeled "component A", "component B" and "message" – where the message character walks from A to B. Because the system is able to load arbitrary content from ScenarioML, visualization of another collection of scenarios is accomplished simply by giving the new scenarios as input to the visualization software.

The social interactions thus modeled bring out the patterns of interaction between actors and entities, and the temporal pattern in which actions occur. Inconsistencies in those patterns within and between scenarios are perceived as inconsistent social interactions between the corresponding characters. Presenting these patterns in a social form provides an additional means of identifying missing information, unstated assumptions, and ill-defined terms by mapping them into a social space in which they may be more visible than in their original narrative form.

A video illustrating this work may be found at `http://orchid.calit2.uci.edu/~wmt/movies/softvis.mov`

There is a large body of work on mixed-methods research (for example, [Cre03]), and to evaluate this project we strove to follow that tradition. The initial goal was to obtain quantitative results that would support or refute several hypotheses about the effectiveness of the visualization approach. The study used questionnaires to elicit qualitative and quantitative responses on which a quantitative analysis could be performed. The questionnaire also included several questions that could provide more open-ended results if the more focused and targeted questions produced inconclusive results, and that could give us broader insights that could be useful in improving the visualization approach and guiding the longer-term direction of the research. The study design was not successful in setting up opportunities for triangulation, with which we could show that both the quantitative and qualitative results supported each other or that they converged. However, we found that both the quantitative and the qualitative results were valuable, and indeed the qualitative results have proved more valuable than the quantitative results that were our original goal.

Initial results of the study discussed here were presented as validation for the authors' previous work [ATB06]. The current paper evaluates the study in greater depth and discusses larger issues that it raised.

The remainder of the paper is organized as follows. Section 2 presents the study in more detail. Section 3 discusses some intriguing and unexpected qualitative results from the study. In Section 4 we reflect on some larger issues that the study raised. Section 5 concludes the paper with a summary of lessons learned and some future work.

2. The study

We hypothesized that the visualization described above would help non-experts understand a collection of scenarios. We tested this hypothesis by presenting two small scenario collections in two different ways – traditional text-based, and socially visualized – to two groups each consisting of eleven sophomore and junior computer science students at the University of California, Irvine. The students were appropriate subjects because they themselves are non-experts in understanding and analyzing collections of scenarios and the systems represented by them. We chose two scenario collections of approximately equal size and complexity, one describing a familiar system and context, an Automated Teller Machine (ATM), and the other describing a system and context unfamiliar to nearly all the students, the Traffic Information System (TIS) used in some private planes to help pilots avoid midair collisions. Both collections consisted of three scenarios expressed in approximately 800 words. The

ATM scenarios contained about 60 events in total and the TIS scenarios about 40. Both collections recorded observations of actual use of the systems, the ATM being those of the first author's credit union and the TIS being a Bendix/King KMD250 Multi-Function Display/GPS.

Both scenario collections were presented to the first group of eleven students in printed form only, while both collections was presented to the second group in printed form accompanied by an animated social interaction visualization. Each group was given 30 minutes to read the scenarios and/or view their visualizations. The two groups then switched presentation forms, with the first group viewing the visualizations of both collections and the second group reading them printed form. Thus each scenario collection was evaluated twice by both groups, but in opposite sequence (printed then visualization, or visualization then printed).

The visualizations were presented twice during the print plus visualization task, which took almost the entire 30 minutes

Each collection was seeded with approximately twenty faults covering a range of common types of scenario faults, including the types we expected our social visualization to be especially effective with. The faults included gaps, local inconsistencies, non-local inconsistencies, external inconsistencies, undefined items, and ambiguously-defined terms.

The subjects were given questionnaires to fill out during the 30-minute period. The questionnaire consisted of these questions:

"1. What problems did you find, and when (what minute) did you find each one?" [Space sufficient for 30 items was provided.]

"2. Have you ever used a system like the one the scenarios describe (ATM or TIS)?"

"3. What other comments or suggestions do you have about the scenarios, the way they were presented, or this study?"

At the end of the first 30 minutes, the two groups were given fresh copies of the questionnaire when they switched tasks. Thus each group answered the three questions for each collection and for each form of presentation.

The questionnaires were then initially analyzed to associate each noted problem, if possible, with a specific fault in the scenarios. These faults included both the intentionally seeded faults and a small number of additional "authentic" faults identified by the subjects. The fault identifications were then classified by the context in which they were discovered:

[PV1] during the first 30 minutes by a subject working with the printed scenarios and animated visualization;

[P1] during the first 30 minutes by a subject working with the printed scenarios only;

[PV2] during the second 30 minutes by a subject working with the printed scenarios and animated visualization (these subjects had already worked with the printed scenarios only);

[P2] during the second 30 minutes by a subject working with the printed scenarios only (these subjects had already worked with the printed scenarios and animated visualization).

We chose to give the subjects the printed scenarios to refer to as they viewed the visualization, because we thought it possible that many subjects would not effectively grasp the details of the visualization without a printed copy to refer to occasionally.

We hypothesized that this analysis would indicate that the visualization was more effective than the printed form for the task of identifying faults, and that the relative effectiveness would vary across the types of faults, being higher for non-local inconsistencies and lower for local inconsistencies for example. We also hypothesized that a further analysis in more detail (not yet completed) would show the variation of relative effectiveness for specific faults. This further analysis involves finding cases in which a subject did not initially identify a fault during the first 30 minutes but did identify it during the second 30 minutes using the other presentation form.

Subjects working first with this prototype of the visualization identified somewhat fewer problems than those working first from the printed scenarios alone, both for the familiar system (ATM, 14% fewer) and the unfamiliar system (TIS, 9% fewer). It was interesting that the subjects who worked first with the visualization and printed scenarios (PV1), and second with the printed scenarios alone (P2), identified a substantial number of additional problems in the second half-hour (P2): ATM 105% in addition and TIS 129% in addition. This was in contrast to the subjects who worked first with the printed scenarios (P1) and then also with the visualization (PV2); these subjects found relatively few additional problems (ATM 23%, TIS 22%). Overall, the PV1+P2 subjects identified 43% more ATM problems and 71% more TIS problems. This suggests that the visualization may significantly augment the effectiveness of the printed scenarios, especially for an unfamiliar system or domain such as TIS. A further study will be necessary comparing PV1+P2 to an equivalent time spent only on the printed scenarios.

3. Some intriguing qualitative results

The qualitative analyses led to several realizations about the process of creating viable social software visualizations. First, it became clear that the visualization prompted a much broader dynamic range of emotion from participants than the text-based scenarios did. While examining the scenarios in text form, the participants sat quietly and read diligently over the scenarios looking for inconsistencies. However, during the visualization session, several emotions and expressive facial gestures were exhibited. Participants looked confused when first presented with the visualization. Several of them smiled when they realized the connection between the scenarios they'd read and the characters on the screen. One gave a look of either intense concentration or disgust as she attempted to decipher the relationships among the animated characters. Several participants laughed at the characters' pronunciation of the word "espanol".

Several questions arise from these emotional responses. Why does the visualization lead to greater emotion? How can the negative responses (disgust, confusion) be reduced, and the positive responses (smiling, laughing) be enhanced? More specifically, how can this emotion be parlayed into more useful behavioral patterns such as attention, concentration, engagement, and understanding? While the qualitative analysis did not immediately provide answers to these questions, it nevertheless started the research team down some paths that appear to be fruitful.

A second realization that arose from the qualitative analysis of the pilot study involved the interactional scaffolding (that is, the setting and other framing elements of the experience) for the participants' engagement with visualized scenarios. With text-based scenarios, the interactional scaffolding is fairly standard – most people who might engage with software scenarios are already familiar with the process of reading books and binders of text. However, scenario visualization draws on a different set of interactional experiences from the participants – movies, cartoons, video games. The interactional scaffolding for social software visualization needs to help people make the right connections between their media expertise (e.g., looking for plot inconsistencies, understanding sequences of social interactions) and the scenarios with which they are interacting.

To give a very simple example, at the beginning of the pilot study with the visualization, nearly all of the students were sitting too far from the screen to read the text labels above the characters. While this seems like a silly mistake, it is a mistake that would have been auto-correcting in a traditional text-based interaction. People would automatically move the book closer to their eyes until the text was legible. Not so with the scenario visualization. In order for people to process the visualization in a way that will result in greater understanding of the constituent scenarios, the set-up of the interaction needs to be crafted with care, from the placement of seats, to the volume of the speakers, to the contrast of the monitors, to the number of times the scenarios are repeated, to the mechanism by which the viewers interact with the scenarios. All of these elements demonstrate a need for basic principles of human-computer interaction design in social software visualization, whereas the corresponding principles in text-based interaction may be taken for granted.

4. Discussion

Overall, the quantitative results provided ambiguous results except for the points discussed in Section 3. The numbers were roughly the same for the visualization and the printed portions of the study. In addition, the numbers were small in absolute terms (most faults were identified by 0 to 3 subjects) so that the significance of each individual result was low. We speculate that a deeper analysis, involving relating the results for individual anonymous subjects, may produce more significant results or provide more insight; such an analysis is possible future work.

By comparison, the qualitative results have proved more useful for understanding how the visualization is effective or not in its present form, and the directions in which it could be improved. The qualitative results from the questionnaires proved less helpful in this respect than the observations of the subjects by the second and third authors over the course of the study (the first author, who was a class instructor for the subjects, was not directly involved in the conduct of the study in order to reduce any bias this would produce).

Perhaps most interestingly, the study raised issues of what goals a scenario visualization should be directed towards, and in a larger sense what benefits can be hoped for or expected in scenario-based requirements engineering. An evaluation of the visualization approach discussed here, or in fact any scenario technique, must eventually be connected to these goals and benefits. This study focused on defect identification in part because this was believed to be quantifiable and measurable, but defect identification is only a part of what people do with scenarios. We are continuing to investigate what goals and benefits are significant and how each of these can be evaluated and if possible quantified.

Any sort of quantitative analysis is predicated on qualitative judgments. For example, in the study presented here, one of the metrics used to evaluate the system was the number of problems (such as inconsistencies or omissions) the participants recognized. This allowed for a number of different calculations, including percentage of problems recognized by any single participant, the percentage of participants that recognized any single problem, correlations between whether different pairs or groups of problems were all recognized, and so forth. However, all this quantitative analysis is predicated upon qualitative judgments. First, what constituted recognizing a problem? For example, in a scenario about getting cash from an ATM, one participant noted "How does the user select the language?" Indeed, in the scenario to which the participant was referring, the ATM's language selection prompt had been omitted, but should the response be noted as recognizing a problem? It could be argued that, because the participant phrased the response in the form of a question, this should not be counted as recognizing a problem, because the participant did not exactly identify the nature of the problem. On the other hand, the participant's question indicates that s/he certainly noticed there was a problem, and while s/he might not have known the proper approach to remedy the problem, the purpose of the system was to expose problems in the scenarios. In this instance, the ultimate decision was to count it as recognition of a problem. However, the example still serves to demonstrate the ways in which qualitative decisions are made in order to make data amenable for a quantitative analysis.

Similar transformations are made during statistical analyses that attempt to demonstrate statistically significant differences. However, as opposed to the move from qualitative to quantitative described above, these transformations are from quantitative data into qualitative comparisons between groups. The purpose of statistics is to distinguish signal from noise, for example, to determine if differences between groups are likely to happen due to random chance or if it is more likely that there is a significant difference between two groups. This, however, begs the question of what is significant. Statistical methods provide many different ways to establish significant differences; one which is quite common is called the p-value, which corresponds to the percent chance that the observed differences between two groups was due to random chance. Generally, the goal is to produce a p-value < 0.05. Many researchers use this as a hard requirement to establish significance, which leads to the practice of researchers making statements such as "our results were almost significant," or "there were differences between the group, but those differences were not significant." Here, the quantitative evaluations are transformed into qualitative judgments about the significance of the difference between two sample groups.

It is important at this point to note the central role of calculation in any scientific venture. Latour [2] describes ways in which calculations serve as an obligatory passage point that shapes the sciences. By making qualitative judgments such as those described above, one can transform any data into a numerical form. Once in that form, mathematics does not care if what is being represented is a participant's questions about problems in a scenario or data of a completely different kind. Mathematical tools can be used to compare two things in a quantifiable way that in actuality are quite different, such as two participants' experiences with a scenario visualization, or even two different participants' experiences with two entirely different systems. However, this use of calculation as the lingua franca is not without its drawbacks. For example, in fields such as Requirements Engineering, where there is an increasing push toward quantitative evaluation, a lack thereof may be seen as a significant weakness, when in fact coercing certain qualitative data into a quantitative form may make it less informative and less useful for the audience. This is not to say that quantitative evaluations are unnecessary or not useful, but rather that a quantitative presentation may not be the most effective way to exhibit some studies' results, and that the presentation of quantitative data with the omission of any qualitative data can potentially make the results of a study confusing and misleading.

5. Lessons learned and future work

During the course of the study and in evaluating it in retrospect, we found that a mixed-method study can provide better results than a quantitative or qualitative study. Although we were not able to triangulate across the two kinds of results, we were able to obtain valuable results of each kind from the data for a single study. Overall, we found that where the quantitative results were not ambiguous they provided focused support for or against specific hypotheses; by contrast, the qualitative data was more effective in providing insight and direction. It is likely that a better study design will allow us to triangulate between the results and to derive more focused support from the qualitative data.

The experience of conducting a comparative evaluation provided value beyond the potential validation of our research work. In order to set up the study, and as part of the process of evaluating the study

retrospectively, we were forced to give careful thought to the goals and context of our prototype visualization and of our research in this area. This careful thought has allowed us to better understand what we have already done, and more effectively plan the future direction of this research project and related research work. We believe that every requirements researcher can benefit similarly from considering his or her work in terms of how it can be evaluated, and from planning and performing comparative evaluation studies.

The central position taken by this paper is that quantitative and qualitative evaluations of Requirements Engineering techniques are valuable separately, and are also mutually reinforcing. The paper has presented a study of a social scenario visualization technique that included both qualitative and quantitative evaluations, and demonstrated the value derive from both forms of assessment. While this paper presents just one study, the position that it supports may be more broadly applicable across the entire field of Requirements Engineering – that is, that RE research teams could benefit significantly from planning and performing mixed-method studies and simultaneous qualitative and quantitative analyses of their research efforts.

6. Acknowledgements

The authors thank Susan Sim, André van der Hoek, and the anonymous reviewers for SOFTVIS 2006 for their valuable suggestions.

7. References

[Ale04] Alexander, I. Introduction: Scenarios in system development. In *Scenarios, Stories, Use Cases: Through the Systems Development Life-Cycle*. I. Alexander and N. Maiden, eds. John Wiley & Sons, Ltd., pp. 3–24, 2004.

[Als06] Alspaugh, T.A. Relationships Between Scenarios. Institute for Software Research Technical Report UCI-ISR-06-7, University of California, Irvine. May 2006.

[ATB06] Alspaugh, T.A., Tomlinson, B., and Baumer, E. Using Social Agents to Visualize Software Scenarios. To appear: *ACM Symposium on Software Visualization (SOFTVIS '06)*, 2006.

[BFJZ92] Benner, K., Feather, M. S., Johnson, W. L., and Zorman, L. Utilizing scenarios in the software development process. In *IFIP Working Group 8.1 Working Conference on Information Systems Development Processes,* 1992.

[Cre03] Creswell, John W. *Research Design: Qualitative, quantitative, and Mixed Methods Approaches.* Sage Publications, Thousand Oaks, CA, USA, 2003.

[Lat87] Latour, B. *Science in Action*. Harvard University Press, Cambridge, MA, 1987.