Crossing Qualitative and Quantitative Evaluation in the domain of Ubiquitous Computing

Jennifer Mankoff
HCII
CMU, Pittsburgh, PA
Plan to participate

Scott Carter
EECS
UC Berkeley, Berkeley, CA
Plan to participate

ABSTRACT

Ubiquitous computing (Ubicomp) applications are applications that are embedded seamlessly and ubiquitously into our everyday lives. The study of Ubicomp is concerned with enabling a future in which the most useful Ubicomp applications are feasible to build and pleasing to use. But what is useful? What is usable? What do people actually need? These questions are only beginning to be answered, partly because Ubicomp systems more difficult to evaluate, particularly at the early stages of design, than desktop applications. We argue that an effective evaluation of deployed Ubicomp technology should combine qualitative and quantitative methods, and should make use of as many unobtrusive methods for gathering data as possible.

INTRODUCTION

Ubiquitous computing (Ubicomp) applications are applications that are embedded seamlessly and ubiquitously into our everyday lives. The study of Ubicomp is concerned with enabling a future in which the most useful Ubicomp applications are feasible to build and pleasing to use. But what is useful? What is usable? What do people actually need? These questions are only beginning to be answered, partly because Ubicomp systems more difficult to evaluate, particularly at the early stages of design, than desktop applications. This difficulty is due to issues like scale and a tendency to apply Ubicomp in ongoing, daily life settings unlike task and work oriented desktop systems. We argue that an effective evaluation of deployed Ubicomp technology should combine qualitative and quantitative methods, and should make use of as many unobtrusive methods for gathering data as possible.

Our work to date has involved a study of diary methods (to be published at CHI 2005) investigating the value of different kinds of data gathered by participants (including audio, tangible information, and images) at supporting qualitative interviews. We believe that *critical incidents* can be generated based on a log of sensed data about the application participants are using and the environment in which they are using it. These critical incidents could serve multiple functions in enhancing diary methods. First, they could be used during an interview, just like data gathered by participants, to elicit responses. Additionally, if any critical incidents were ongoing when other data (such as photos) were gathered, this information could be of great use to experimenters. Second, they could be used during the data gathering phase, as the basis for an ESM method.

We are proposing to create a tool that can either respond to or learn *critical incidents* based on a log of sensed data, and identify ongoing critical incidents at times when qualitative data is gathered.

BACKGROUND

One of the hardest problems that application developers face today is evaluating ubiquitous computing applications, such as those that adjust output based on complicated contextual cues or use natural input technologies such as gesture and speech. We define evaluation to include not only lab and field experiments that assess the usability of built systems, but also field studies that drive invention, early-stage requirements gathering, and prototype iteration.

Because of the difficulty of each of these steps, the evaluation of Ubicomp systems is an active area of discussion, as evidenced by two recent workshops on the topic, and an active area of research. Scholtz's 2001 workshop laid the groundwork for evaluation as a way of moving the Ubicomp community forward and bringing it together. Two workshops held on similar topics the following year had the goal of collecting "best known methods" and supporting case studies for dissemination in the community (Scholtz *et al.*, 2002) and discussing evaluation techniques for a subset of ubiquitous computing, notification systems (including ambient and peripheral displays) (Bartram and Czerwinski, 2002). Meanwhile, researchers have begun to study evaluation in the context of its application to different Ubicomp systems.

For example, Consolvo *et al.* (2002) recently published an evaluation of an enhanced biology lab which they evaluated using Lag Sequential Analysis (the first time that technique had been applied to a Ubicomp system). Mankoff *et al.* (2003) have developed modifications to Heuristic Evaluation that enhance its applicability to a sub-domain of Ubicomp, ambient displays. Intille, et al., have developed tools and techniques for experience sampling in Ubicomp settings (2002).

Evaluation is crucial at all stages of design, and the best designs include evaluations that involve users in the design process repeatedly throughout a series of design iterations.

Evaluation research can help us develop a suite of techniques applicable to finished systems (such as Consolvo *et al.*'s work) and early stage design (such as Mankoff *et al.*'s work).

DIARY STUDY

Researchers have a handful of tools and techniques available for understanding everyday human behavior. But many of these techniques either require significant time and resource investment by researchers, such as contextual inquiries, or are divorced from empirical evidence, such as surveys. The diary study is a method of understanding participant behavior and intent that attempts to manage this gap by having participants record events as they happen. This recording usually occurs in one of two ways: participants answer predefined questions about events (feedback studies) or participant capture media that are then used as prompts for discussion in interviews (elicitation studies).

Field studies that require the researchers' persistent presence are difficult to scale. On the other hand, because of their reliance on participants to collect data, feedback studies have the potential to be scalable. However, participants are often reluctant to use them because the act of answering questions is a significant distraction from their main task. Also, because of the lack of an objective observer there is no way to verify to what extent logged information matches actual events.

Media elicitation studies mitigate both of these concerns. In a media elicitation study, participants capture events, usually by taking a photo, and are asked about the event during an interview at a later point in time. Thus for elicitation studies, capture is quick, and while the captured media still represents a subjective point-of-view, it has some empirical value.

Barsalou posited that episodic memory can be improved when a person is presented with cues about an event such as who was involved, where it occurred or what was done just before and after the event (1988). However, while researchers have recently begun using diary studies using photo-elicitation, it is not evident how well media capture these cues and to what extent media facilitate participant reconstruction of events. Also, different media types will likely evoke different reconstructions and attitudes towards an event, but no study has yet shown how.

Based on these concerns, one contribution of the paper we present at CHI 2005 is a set of suggested improvements to the diary study technique, derived from three studies of the technique itself in action. For two of these studies we played the role of a participant observer by involving ourselves in an ongoing study. Specifically, we observed the process of using the method, analyzed results from the study and interviewed the researchers involved about their experiences. The other study we ran ourselves, to gain first-hand insight into the issues involved in running a diary study and to compare and contrast the use of different capture media: photos, audio clips and tangible (physical) objects. While photo diary studies are gaining in popularity, use of the other two media is limited.

Our studies revealed a need for situated annotation of captured events in elicitation studies. We found that the best approach to feedback studies may be to combine media capture with structured, question-and-answer based annotations. Our studies also revealed the usefulness of different media in different situations. Specifically, we found that images lead to more specific recall than any other medium, but that audio, in addition to making it easier for participants to capture information that does not have a visual representation, can be used clandestinely in situations in which participants do not feel comfortable using a photo to capture an event. We found that information about location does not significantly impact recall, and that tangible objects are more likely than other media to prompt discussion of broad attitudes and beliefs.

We also noticed unforeseen issues in elicitation interviews. For example, while media capture lent itself to a sequential review of data, interview discussion tended to follow themes, causing problems for participants and researchers when they referenced captured data out-of-sequence.

Our experience with media-based diary studies as well as reports in the literature, indicate that it is important to mitigate the impact of a study on participant's everyday interactions and encourage participant recall of ambiguous data. We also found it important to provide support for interview preparation. To address these issues we proposed a diary study pipeline that borrows from both feedback and elicitation methods to maximize participant recall and interview preparation while minimizing situated logging. We then built and tested a lightweight tool, *Reporter*, to support this pipeline. Results showed that participants were able to learn the tool rapidly.

TOOL

While *Reporter* aids participant recall as well as researcher preparation, critical incident recognition and data capture for diary studies can still be improved. The tool we propose should be able to automatically annotate streams of data with out-of-the-ordinary events. Also, the tool should support both *implicit* and *cued* data capture for diary studies. *Implicit* capture is tied to an infrastructure (*e.g.*, a camera attached to a door) whereas *cued* capture occurs on a personal device (*e.g.*, a PDA). Critical incidents recognized from log files allow researchers to quickly navigate to data of interest during elicitation interviews, whereas critical incidents recognized *in situ* can expand the possibilities for the types of events researchers can discuss with participants.

The tool should support both researcher-defined critical incidents as well as a set of universal critical incidents that researchers can quickly parameterize. In many diary studies, researchers are interested in what captures participant's attention, factors that arouse emotional responses in participants, and what prompts participants to shift between activities. The tool should be able to support standard types of attention recognition for implicit capture. For example, given an image source the tool should be able to detect whether someone is looking or gesturing towards something in which the researcher is interested (*e.g.*, a peripheral display) to queue data capture. Also, the tool should be able to support galvanic response for cued capture of affective events. Researchers should be able rapidly to parameterize this to cue participant-driven capture for particular emotional states. Finally, the tool should support cued capture of changes in physical activity as well as implicit capture of changes in computational activity. Cued capture might be accomplished using a few rudimentary wearable sensors as input. For implicit capture, the tool could use log data from other tools being developed in applied Activity Theory to detect computational activity changes.

CONCLUSIONS

In conclusion, when technology use is embedded in the daily life and environments of consumers, answering questions about how technology is used becomes more difficult. The diary study is a scalable method for gathering data in everyday environments. However, while it has seen significant use, only recently have researchers begun to enhance it with technology such as digital photos, mobile phones, and so on (Carter and Mankoff, 2005).

Based an study of how media is used in diary studies, and the development of a tool to support media-driven diary studies, we are proposing to create a sensor-based tool that logs critical incidents ranging from changes in physiological arousal relating to affect to low level activity changes. In addition to logging this data implicitly for later perusal by researchers, our tool could support cued, *in situ* capture.

We believe that the combination of the qualitative diary study technique with the logging of sensed events is a powerful and important contribution to the experimental techniques available to Ubicomp researchers.

REFERENCES

Barsalou, L. W. Remembering reconsidered: Ecological and traditional approaches to the study of memory. In *The content and organization of autobiographical memories*. eds. U. Neisser & E. Winograd. Cambridge University Press, 1988 193—243.

Bartram, L. and Czerwinski, M. (2002). Design and evaluation of notification interfaces for ubiquitous computing. Ubicomp 2002 Workshop 9.

Carter, S.and Mankoff, J. (2005), When Participants Do the Capturing: The Role of Media in Diary Studies, CHI 2005, To Appear .

Consolvo, S., Arnstein, L., and Franza, B. R. (2002). User study techniques in the design and evaluation of a ubicomp environment. In Ubicomp 2002, volume 2498 of Lecture Notes in Computer Science, pages 73–90.

Intille, et al., Eliciting user preferences using image based experience sampling and reflection, In *Extended Abstracts of CHI 2002*, pp. 738-739.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Ames, M., and Lederer, S. (2003). Heuristic evaluation of ambient displays. Proceedings of the ACM CHI'03 Conference on Human Factors in Computing Systems, CHI Letters, 5(1):169–176.

Scholtz, J. (2001). Evaluation methods for ubiquitous computing. Ubicomp 2001 Workshop.

Scholtz, J., Tabassi, E., Consolvo, S., and Schilit, B. (2002). User centered evaluations for ubiquitous computing systems: Best known methods. Ubicomp 2002 Workshop 2.