Experience Sampling for Building Predictive User Models: A Comparative Study

Ashish Kapoor and Eric Horvitz
Microsoft Research
One Microsoft Way, Redmond, WA 98052 USA
{akapoor, horvitz}@microsoft.com

ABSTRACT

Experience sampling has been employed for decades to collect assessments of subjects' intentions, needs, and affective states. In recent years, investigators have employed automated experience sampling to collect data to build predictive user models. To date, most procedures have relied on random sampling or simple heuristics. We perform a comparative analysis of several automated strategies for guiding experience sampling, spanning a spectrum of sophistication, from a random sampling procedure to increasingly sophisticated active learning. The more sophisticated methods take a decision-theoretic approach, centering on the computation of the expected value of information of a probe, weighing the cost of the short-term disruptiveness of probes with their benefits in enhancing the long-term performance of predictive models. We test the different approaches in a field study, focused on the task of learning predictive models of the cost of interruption.

Author Keywords

Interruption, Decision Theory, Experience Sampling

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces; I.5.4 [Pattern Recognition]: Applications

INTRODUCTION

Interest has continued to grow on the use of machine reasoning and learning to enhance human-computer interaction. Several systems have relied on the use of statistical user models to predict states of the user or context, *e.g.*, see [1, 6, 8, 9, 10, 12]. One of the key problems in constructing such predictive user models is the collection of labeled data, where labels contain information about hidden states of computer users such as their interruptability, intentions, needs, and affect. In some cases, labels need not be assessed through the explicit engagement of users; rather, states can be associated in an implicit manner with other sensed data via a process of *in-stream supervision*. As an example, in-stream supervision has been used to construct predictive models for user goals and the timing of actions in the mixed-initiative Lookout system [6]. The system seeks to identify users' intentions by observing their interactions with the Microsoft Outlook application, making the assumption that a calendaring action, occurring within a period of time after an email message is viewed, associates the content of the message with the goal of reviewing the calendar.

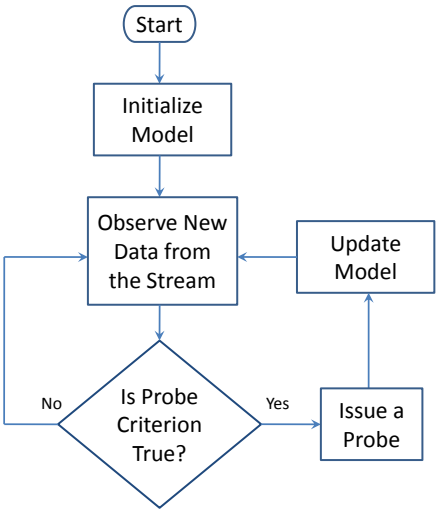


Figure 1. Overall flow of automated experience sampling methods. Key differences among methodologies center on the probe criterion.

Unfortunately, in-stream supervision is often infeasible; users more typically must be engaged to provide information about hidden states in a manual manner. Experience sampling methodologies (ESM) have been used for acquiring real-time labels of situations. With ESM, people are asked, in the course of their normal activities, to reveal hidden states required to build a case library for machine learning. While providing an effective way to label cases, ESM can be disruptive and frustrating for users.

Figure 1 shows the overall flow of automated ESM. The ESM systems are provided with access to a stream of data about the user and/or context and couple the sensed information with user feedback accessed via the probes to provide insights to researchers or to construct predictive models. Automated ESM systems continue to make decisions as to if and when to issue a probe to users, based on some *probing criterion*. The probing criterion lays at the heart of

CHI 2008, April 5 - 10, 2008, Florence, Italy.

the method: changes in the criterion influence the behavior of the system and the quality of the predictive models constructed from the collected data over time.

In the past, relatively simple policies have been used to guide experience sampling. Most experience sampling to date has been based on random sampling. With random sampling, the times for probing for feedback from subjects is randomized. Variants of random sampling have included methods that allow users to modulate the density and distribution of probes via controls that can be used to adjust parameters of the random sampling [8]. Landmark-based heuristics have also been investigated where predefined contextual cues are used to herald an appropriate time for a probe. For example, specific sensed events or locations have been used as triggers for probes [5, 11]. Landmark-based ESM require that triggering events are known in advance and that the apparatus has a reliable means for detecting those events. More sophisticated ESM systems can use predictive models constructed from the data collected in earlier probes to guide decision making about if and when to probe subjects for new assessments. These systems have machinery for deliberating in an explicit manner about when to probe, promising to reduce interactions while maximizing the value of information being acquired with each probe.

We have pursued a comparative analysis of four different approaches to automated experience sampling, spanning a spectrum of sophistication. In addition to a randomized policy, we consider three methods that employ inferential mechanisms which seek to identify the best times for real-time engagement of subjects for their assessments of state. In distinction to the randomized policy, these methods consider an economics of information and disruption; they are designed to be selective about collecting information from users, and seek to maximize the value of probes. We have been particularly interested in understanding the potential value of employing decision-theoretic methods to enhance ESM by balancing the costs and benefits of probes, via reasoning about the current state of subjects and the world, and performing computations of the expected value of the information gleaned from a real-time probe.

In the next section, we review the four ESM policies that we shall study. Then, we discuss as a testbed domain the challenge of building systems that can learn personalized models with the ability to predict the cost of interruption for a user based on desktop activity, calendar information, and such contextual information as ambient acoustical signals and wifi signals.

We extend prior work on BusyBody [8], a system that was first introduced with an ESM method that performs random probing, optionally modulated by parameters provided by users. Then, we describe the details of a two-week field study exploring the use of the different ESM methods by people with a variety of roles at our organization. We perform a comparative analysis including summaries of the attributes of the four different ESM methods, such as the quantity and dynamics of probing over time, the quality of the models constructed, and the overall experience with the methods. We also seek in a post-study survey to understand how recent experience with using the different systems influences general feelings subjects hold about systems that learn from them. Finally, we summarize the results, discuss the implications of the findings, and present future research directions.

Method	Considers Error Cost	Considers Probe Cost	Adapts To Data Dynamics
Random	×	×	×
Uncertainty	✓	×	×
DT	✓	✓	×
DT-dyna	✓	✓	✓

Table 1. Selective Experience Sampling Methodologies.

METHODS FOR GUIDING EXPERIENCE SAMPLING

We shall explore four experience sampling methods, each harnessing a different policy for probing subjects as follows:

- **Random probe:** Probes appear at random times.
- **Uncertainty probe:** A predictive model, constructed with data collected so far, is harnessed to generate probes for situations associated with the most uncertainty.
- **Decision-theoretic probe (DT):** Based on the probabilities inferred with the current predictive model about the user's internal state, the expected value of information is computed, weighing the costs and benefits of the probe. When the value is positive, a probe is recommended.
- **Decision-theoretic dynamic probe (DT-dyna):** The decision-theoretic probe approach is extended with a method for addressing the potential unmodeled dynamics of the context. With this extension, the system continues to deliberate about removing and caching assessed states for later reconsideration, allowing for the implicit construction of multiple models across time with changes in context that are not represented explicitly.

The *uncertainty*, *DT*, and *DT-dyna* probes are examples of *selective* ESM probing strategies in that they attempt to use inferential methods to determine the best time to assess information from users. We used the degree of uncertainty as the criterion in one of the policies because it is the simplest active-learning policy we know and it has been employed in numerous studies of active learning within the machine-learning community [15, 20, 21]. At the core of the two more sophisticated ESM methods, DT and DT-dyna, are *decision-theoretic* probing policies. These methods endow the ESM system with the ability to balance the cost of probing users for labels of states with the benefits of the increased accuracy of models. The decision-theoretic methods consider uncertainties and preferences, and compute a formal quantity called the *expected value of information* by considering the costs associated with the accurate versus erroneous functioning of the predictive model in downstream uses of the predictive system. DT and DT-dyna employ estimations about how much better the predictive model will perform with information that is expected to come from the additional probe. This value of information is based on a weighing of the expected costs and benefits of the informa-

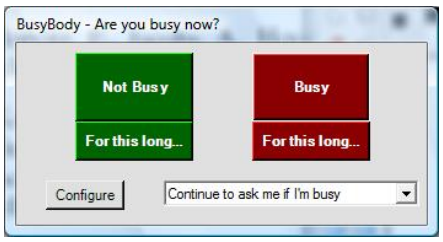


Figure 2. BusyBody probe for user feedback. When running in a binary modality, the probe inquires about whether the user is highly non-interruptible versus in another state.

tion gleaned from the probe, given the current inferred uncertainties about a subject's response to the probe. The computation considers the cost of interrupting a user to inquire about a situation or state, and how the new data point will enhance the downstream performance of a predictive model.

The DT-dyna approach introduces additional machinery that enables the system to be flexible should aspects of the context change in ways that are not necessarily recognized by the model. For example, a laptop user may be working in very different settings, such as home, office, and conference room, and these different venues may not be encoded explicitly with predefined variables sensed and used by the system. With DT-dyna, the system has the ability to continue to deliberate about whether to include in its construction of predictive models each of the cases it has learned so far from users. In the approach, cases, created by joining probes with contextual data may be removed, cached, and reconsidered at a later time. The decision-theoretic dynamic probe ESM methodology continues to reason about *forgetting* and caching, versus *remembering* data acquired from probes. This dynamic approach to admitting data into the learning of models allows the system to respond locally in a flexible manner to multiple contexts. Technical details of this approach are described in [13]. In brief, the DT-dyna method employs a variant of value of information called the *value of forgetting*. When integrated into the cycle of reflection about the best points to gather, the value of forgetting allows a system to learn in an efficient manner in potentially dynamic environments—contexts where unmodeled exogenous variables lead to changes in the relationships among observations and states of interest.

In terms of computational requirements, random probing simply requires the generation of a random number which is compared against a preset threshold. Using the uncertainty criterion requires that we use the available predictive model to compute the posterior probabilities of the current state of the target inference—the cost of interruption in the case of BusyBody's domain. The complexity of inference depends on the learning and reasoning methodology. For the inferential methodology (expectation propagation [18]) we have used in this study, the complexity is $O(d^2)$, where d is the dimensionality of the data. The DT and DT-dyna approaches require repeated application of the predictive model on a buffer of previously encountered data points. In this work d is 48 and the computation with models of this dimensionality did not provide significant overhead.

DOMAIN: LEARNING THE COST OF INTERRUPTION

The problem domain of learning the cost of interruption is an especially interesting task for experience sampling; the predictive model—being constructed for use in applications that balance the context-sensitive cost of alerts with the value of increased awareness of messages—provides an ESM policy with inferences about the cost of interrupting users with probes aimed at enhancing the predictive model. As the predictive model is learned incrementally, the adaptive ESM methods have access to estimates of the expected cost of probing a user at different times and can make use of this estimate in deliberation about if and when to probe for additional data.

The work on BusyBody was an early effort to explore the use of automated ESM in a closed-loop manner for learning models for predicting the cost of interrupting people, based on observations of user activity and context [8]. Other work on learning about interruptability includes [10, 7, 3]. BusyBody can operate in different training modalities. In its binary assessment mode, users are probed with a binary decision task; they are asked to indicate if they are highly non-interruptable versus in states of lower cost of interruption. In use, the BusyBody system continues to generate an expected cost of interruption by computing the probability that the user is highly non-interruptable. At run-time, the system provides other applications with a current expected cost of interruption or with information on when the expected cost of interruption exceeds a user-set threshold.

When BusyBody is in training mode, the system intermittently probes users with a pop-up query, requesting an assessment of the computer user's current or recent interruptability. The initial version of the system probed users at random times, constrained to an overall maximum rate and inter-probe interval as set by users via a set of controls available on the BusyBody probe pop up. Figure 2 shows a request by BusyBody for input that is used when the system is running in a binary-hypothesis modality.

Details about the original BusyBody are described in [8]. BusyBody employs an event infrastructure that logs desktop activities including such activities as typing, mouse movements, windows in focus, recent sequences of applications and window titles, and high-level statistics about the rates of switching among applications and windows. The system also considers several kinds of contextual variables, including the time of day and day of week, the name of the computer being used, the presence and properties of meetings drawn from an electronic calendar, and wireless signals. The system employs a conversation-detection system, using a module that detects signals in the human-voice range of the audio spectrum. Responses to probes about the current cost of interruption are stored, along with the sensed evidence. The system takes such assessments as labels, and builds a rich case library by joining the labels with a large vector of evidence about computer activity. This case library is used to construct models that can predict the expected cost of interrupting the user at different times, based on observed activity at the computer. Bayesian structure search was used for learning inferential models in the original BusyBody work.

Later versions of BusyBody [13] employed Gaussian Process classification because they allow for efficient computations of the value of information, an analysis that requires multiple steps of learning and inference.

For our field study, we make use of a version of BusyBody using the default randomized probe developed initially for the first version of the system. In addition, we created three new versions of Busybody, each using a different ESM as described earlier.

RESEARCH IN ACTIVE LEARNING

Before moving on to a field study, we shall briefly review relevant research in active learning. Active learning is an area of research in machine learning. Active learning algorithms are employed to make decisions about the next unlabeled case within a library of unlabeled and labeled cases, that should receive a label (*e.g.*, via an active probe), so as to maximize the learner's ability to classify the data.

Variants of Active Learning

Most work in active learning assumes a *pool-based* setting where the set of labeled and unlabeled data are provided and an algorithm selects the points from the pool to query. Various heuristics have been employed as criteria for active learning. Within the Gaussian Process framework, the expected informativeness of an unlabeled data point has been used [14, 16]. For SVMs, Tong and Koller [21] explored the criterion of minimizing the version space to select the unlabeled points to query. Other pool-based methods have been based on a combination of active learning with semi-supervised classification [17, 19, 22]. Shen and Dietterich [20] have used active learning with an entropy-based measure to learn predictive models from noisy data.

Stream-based learning has been less explored than pool-based scenarios. In stream-based settings, the learner views a continuing series of unlabeled points and can make a decision at each step to determine whether to query for the label of a newly presented unlabeled data point. Many of the existing approaches to stream-based active learning can be viewed as adaptations of pool-based active learning strategies to the stream-based scenario. As an example, an approach to stream-based active learning relies on the selection of unlabeled points that the existing classification is most uncertain about [15]. In another approach, researchers have adapted methods that consider the disagreement of a committee of classifiers to the stream-based scenario [4]. Similarly, stream-based active learning for linear classifiers has been proposed [2].

Note that all of these approaches seek to address the challenge of adding points to the active set; none of the methods tackles the issue of eliminating irrelevant, outdated data, and recalling older data that might become relevant. Further, none of the prior methods are targeted at optimizing the criterion upon which the system is ultimately evaluated.

DECISION-THEORETIC EXPERIENCE SAMPLING

We harness recent work on decision-theoretic active learning [13] for adaptive ESM. The method identifies the most valuable unlabeled instances to label by considering the costs of labeling cases and the cost of misclassification. The method also addresses the challenge of learning in a dynamic environment and identifies when the current predictive model conflicts with the current situation. Specifically, the method has the ability to forget and cache older labeled data points in an automated manner via the computation of the expected value of forgetting (VOF) previously labeled instances. Similarly, the method also considers the possibility that cases cached earlier might again become relevant in the current context. Cases are reconsidered by computing the expected value of recall (VOR). The cycle of forgetting and recalling cases can be valuable for learning predictive models within domains that have poorly characterized non-stationarities. Non-stationarity may be founded in "evidential incompleteness"—the absence of consideration in the system of important evidential distinctions that could capture important indicators of change over time. Examples of critical exogenous variables that may be absent in a system for predicting a user's interruptability include the user's appointment status and the day of week. Busybody considers these variable. However, if the system had no knowledge of meetings or of the distinction between weekdays and weekends, the system might find models learned for some settings sometimes perform poorly in other settings for unknown reasons. Moving beyond salient examples of incompleteness, it is safe to say that any learning and reasoning system will likely rely on a representation of the world with "blindspots"—a representation that is incomplete in important ways.

FIELD STUDY

We shall now describe the field study that we conducted to compare the different automated experience sampling methods. Our aim was to study how different policies for probing would affect the behavior and performance of Busybody in a real-life working environment.

Details of the Probing Policies

We sought to compare the four different policies described earlier. Specifically, we built the four different versions of BusyBody to probe the user for labels and to update the predictive model of the user's cost of interruption. The implementation details are as follows:

- **Policy 1: Random probe.** This policy enables BusyBody to probe the user for their cost of interruption in a random manner. The default rate was set to four probes an hour and users had the ability to change this rate.
- **Policy 2: Uncertainty probe.** According to this policy, Busybody issues a probe whenever the probability of classifying the state as busy is between 0.25 and 0.75.
- **Policy 3: Decision-theoretic probe (DT).** For this policy, we assume that the cost of classifying the user state as not-busy when she is busy (R_{12}) 2 USD and cost of classifying the state as busy when she is not (R_{21}) is 1 USD. Similarly, the cost of a probe when the user is busy (C_1) is 2 USD and the cost when not busy (C_2) is 1 USD. Furthermore, the buffer and the optimization horizon consider the past 30 minutes of data and the buffer is updated in 1 minute.

- **Policy 4: Decision-theoretic dynamic probe (DT-dyna).** The parameters for this policy were the same as the ones for Policy 3, except that the framework incorporates the caching and recalling functionality to adapt to the potential non-stationarity of the domain.

Participants

We recruited 44 participants from our organization. The subjects had roles that can be classified as: 1) software developer, 2) researcher, 3) program manager and 4) group manager. A group manager differs from a program manager in that the group manager manages a team as part of her job, whereas program managers are individual contributors. The breakdown of the participants is described in Table 2.

The subjects were randomly assigned to one of the four different conditions. At the beginning of the study, each version of BusyBody was assigned to 11 subjects. Out of the 44 participants, three subjects failed to install BusyBody as directed. Thus, we started with 41 subjects. Out of these 41 participants, one subject had a computer hardware failure and dropped out of the study. Three other subjects did not provide the data log files because of their work commitments. Thus, the results we report are based on our observations on the remaining 37 subjects.

Duration of the Study

The study spanned two weeks and the participants were given a small gratuity for their time as well as a promise to view a report of the analysis of when they were the most busy and free based on the analysis of the data collected during the study.

Setup and Equipment

BusyBody was installed on the computers of subjects. All of the computers were Intel architecture machines running either Windows XP or Vista. The computers were primarily used by subjects for their work throughout the course of the study. The participants also had the ability to access the desktop remotely from other machines and such access did not negatively influence the ability to collect probes, as well as data about activity and context, per the design of BusyBody.

Procedure and Design

The participants were provided with written instructions on how to install Busybody on their system. Following the installation, we did a quick check to insure that the system was properly installed.

The subjects were told that the BusyBody probes would appear occasionally with an associated audio chime, and that they could provide feedback about their sense for their current cost of interruption at that time. Subjects were asked to continue their regular office work on the desktop with the installed BusyBody for a period of two weeks.

In an attempt to standardize the definition of "busy" per the question that the BusyBody probe was asking them to answer when in binary mode, we instructed the subjects as follows:

Condition	Developer	Researcher	Program Manager	Group Manager	Total
Random	4	2	1	2	9
Uncertainty	1	4	3	1	9
DT	2	2	4	1	9
DT-dyna	3	2	3	2	10

Table 2. Details of the subjects who participated in the study.

> "Please only click *Busy* if the cost of stopping immediately is such that you wouldn't stop what you are doing right away and review the alert, even if you knew that the incoming message contained urgent information. Otherwise, click *Not Busy*."

Participants were not informed about the nature of the different experience sampling methods, nor was information provided that multiple versions of the system were being tested. Beyond the probes being generated by the ESM policy, BusyBody randomly probed at a rate of two probes per hour to collect validation data. The goal of this random probing was to gather data to test the performance of the system.

The Busybody prototype on each machine collected the data and probed the participants over two weeks. At the end of the study, the data was transmitted in the form of a set of text files. We analyzed the data to compare the performance among the four methods. Finally, at the end of the study, the participants were asked to complete a survey where we asked questions related to their experience with the BusyBody system and more general assessments of reflections about working with tools that learn from users via experience sampling. We describe the observations, data analysis, and the results below.

RESULTS

The goal of the automated ESM approach is to build a predictive model that has maximal classification performance and to do this with minimum disruption. Hence, to characterize the performance of the ESM probing policies, we analyze both the probing behavior over the two-week study, as well the classification performance at different times—often referred to as learning curves.

Analysis of Probing Behavior

We first compare the number of probes issued by the different ESM methods. The number of probes captures how much supervision the learner requested from subjects. The amount of disruption caused by the ESM system scales with the number of probes issued by the learner; thus, minimizing the number of probes is a desirable goal of an automated ESM system.

Figure 3 shows the statistics for the number of probes issued per day for the different policies. The numbers are averaged across all subjects for each ESM system. The highest number of probes was issued by the random policy (21.35 probes per day), followed by the policy based on uncertainty (19.73) and decision-theoretic probing (16.17). The lowest number of probes were issued by the decision-theoretic dynamic probing (4.65). A one-way Analysis of Variance (ANOVA)

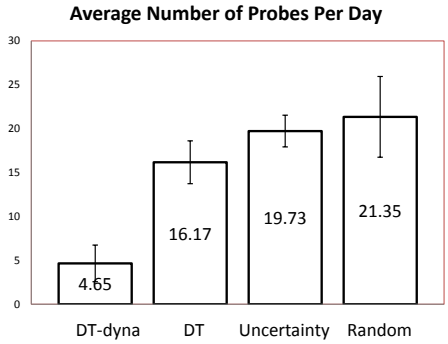


Figure 3. The average number of probes issued per day for different conditions. The decision-theoretic dynamic probing methodology generated significantly fewer probes than the other approaches. The error bars show the standard error.

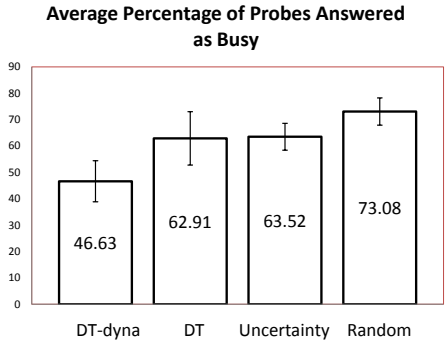


Figure 4. The average proportion of probes that were tagged as *busy* for different conditions. Compared to random probing, the DT-dyna probing methodology issued significantly fewer probes while the users were in a state they labeled as busy. The error bars show the standard error.

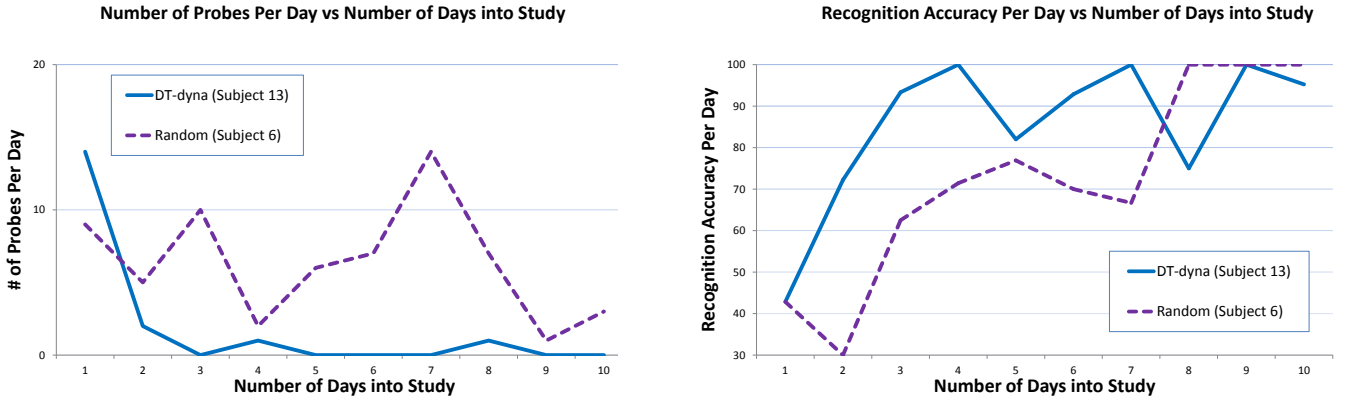
was performed on the data. A main overall effect of probing policy was observed, $F(3, 33) = 7.01, p = 0.001$. Paired comparisons using the Tukey procedure showed that the number of probes issued per day by the DT-dyna policy was significantly less than those issued by the DT ($p = 0.037$), Uncertainty ($p = 0.004$) and Random ($p = 0.001$) policies. The significantly lower number of probes for DT-dyna is likely based in the power of the caching and recalling capabilities of this method, enabling the predictive model to be flexible and adaptable to the subtle dynamics of the domain; the model appears to achieve good performance by adjusting itself to the current context without requiring additional feedback from the user. The uncertainty-based probing and decision-theoretic probing rely on the current state of the model to issue new probes; they consequently seemed to issue smaller numbers of probes than the random policy that is oblivious to the state of the world and user.

Next, we plot the percentage of responses labeled as *busy*. The cost of probing a user varies over time. In general, a decision-theoretic probing policy should be generally expected to issue fewer probes when probing is expensive. Figure 4 shows the percentage of busy responses. The DT-dyna probing generates the lowest number of probes (46.63%) when the user was in the busy state. Similarly, on average the Uncertainty (63.52%) and DT methods (62.91%) issue lower numbers of probes in the busy state when compared to Random (73.08%). A one-way Analysis of Variance (ANOVA) on the data showed a main overall effect of probing policy, $F(3, 33) = 2.89$, $p = 0.05$. Further, paired comparisons using the Tukey procedure showed a significant difference between DT-dyna and the random probing policy ($p = 0.03$). No other effects were observed in other pairwise comparisons. Both decision-theoretic probing policies consider the cost of probing in deciding when to probe; their behavior with disrupting subjects less was in line with the intent of their machinery for doing cost-benefit analysis of the costs and benefits of probing at different times.

We also investigated the change in probing behavior over time for the different methods. Figure 5(a) shows how the number of probes issued per day change as data accrues over time for two subjects that were being probed by the Random and DT-dyna probing methodologies. The random-probing policy does not take into account any information about the state of the world and the model; thus, the probing behavior shows minimal changes. In contrast, for the DT-dyna probing methodology, the number of probes issued decreases significantly over time. The probing policy is aware of the state of the world as well as the model; as the system learns over the time, the number of probes issued per day decreases.

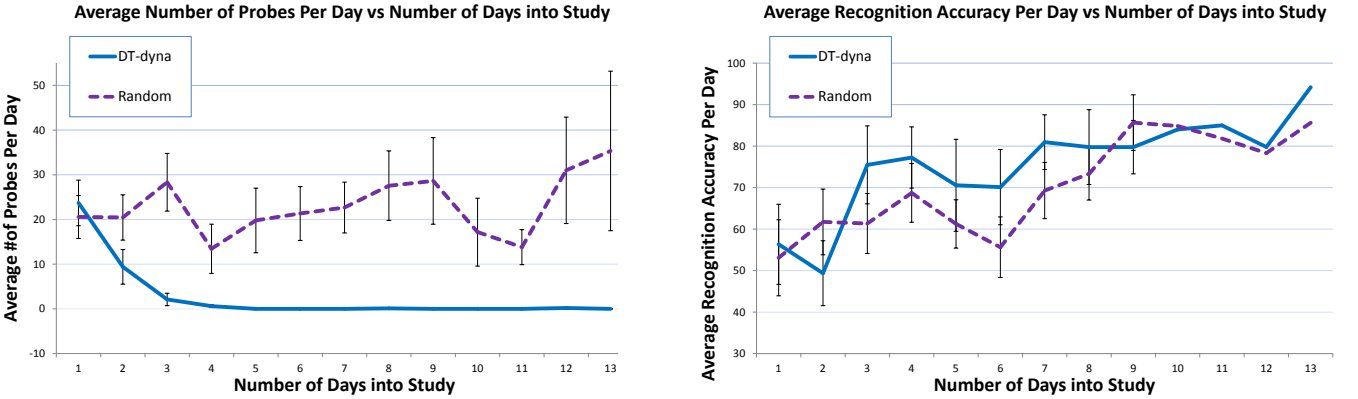
Figure 6(a) displays plots of statistics averaged over all the subjects within each condition. A similar trend is expressed over the averaged values for the groups of users using the policies. These results highlight the potential benefits of employing decision-theoretic dynamic probing. By considering the current state of the user, the current predictive model, and the information acquired in the past, the DT-dyna policy refrains from issuing a new probe unless the benefit in performance gain is greater than the cost of the disruption.

We explored the evolution of the system's behaviors over time. Returning to the two subjects above, Table 3 shows the predictive accuracy on the test points seen so far, as well as the total number of probes issued at the end of weeks 1 and 2 of the study. As demonstrated by the data, DT-dyna issues a very small number of probes in the second week as it already has acquired most of the information that it infers (per the decision-theoretic machinery) it needs in the first week. Table 4 shows these statistics averaged over all the subjects and for all of the four probe conditions. Note that the table also shows that, on average, the two decision-theoretic methods (DT and DT-dyna) issue far fewer probes during the second week. The Uncertainty policy also issues on average fewer probes in the second week when compared to Random. To judge the significance of these numbers (Table 4), we performed a one-way ANOVA analysis on the number of probes. A main effect was observed for both weeks: $F(3, 33) = 4.85$, $p = 0.007$ (week 1) and $F(3, 33) = 5.80$, $p = 0.003$ (week 2). Follow-up tests were conducted to evaluate pairwise differences among the means. As the variances were not homogenous, we conducted the post-hoc tests us-



(a) (b)

Figure 5. Variation in (a) number of probes issued per day and (b) recognition accuracy achieved per day with time for two subjects in the study. The DT-dyna probing methodology reduces the number of probes issued as time progresses while improving the recognition accuracy. Conversely, the random probing methodology continues to probe the user at unchanging rates.



(a) (b)

Figure 6. Variation in (a) number of probes issued per day and (b) recognition accuracy achieved per day with time. The averages are computed over all the subjects in a condition. The DT-dyna probing methodology showed fewer probes as time progressed while yielding a higher recognition accuracy. The random-probing methodology continues to probe the user at relatively high rates.

	AT WEEK 1		AT WEEK 2	
Method	**Accuracy**	**Total Probes**	**Accuracy**	**Total Probes**
DT-dyna (Sub 13)	80.20	17	87.13	18
Random (Sub 6)	55.26	60	64.21	98

Table 3. Summary of the cumulative performance on the validation data after one and two weeks of usage. The averages are computed across the subjects in each ESM probing condition.

ing the Dunnett's C test. For both weeks, the tests showed that there is a significant difference between the mean of the total number of probes for DT-dyna and that of the other three policies. No other significant effects were observed. In summary, when compared to the other methods, the DT-dyna approach uses the least number of probes to attain a similar classification accuracy.

Analysis of Performance

The ultimate goal of automated ESM probes is to learn an effective predictive model. We now explore the classification performance of the models built with the data collected

	AT WEEK 1		AT WEEK 2	
Method	**Average Accuracy**	**Mean Total Probes**	**Average Accuracy**	**Mean Total Probes**
DT-dyna	67.35	35.90	69.80	36.00
DT	70.94	145.78	70.98	169.11
Uncertainty	60.70	172.44	62.83	207.56
Random	63.09	174.11	65.76	256.22

Table 4. Summary of the cumulative performance on the validation data after one and two weeks of usage. The averages are computed across subjects.

by the different ESM probing policies. We examine learning curves that capture how the system's classification performance evolves over the time. Figure 5(b) shows the accuracy achieved per day over the time for the two individual subjects we have been focusing on. The DT-dyna policy shows better recognition performance even during the early phase of the study. As the probes issued by the DT-dyna policy are targeted to gather information inferred to be the most valuable for enhancing predictions, the system achieve a good performance level with fewer probes, and thus interruptions, for the user.

Again, we plot these statistics averaged over all the subjects in each condition (see Figure 6(b)). We can see that a similar trend is expressed over the averaged values. These results highlight the benefits of the decision-theoretic dynamic probing. We can combine the observations from Figures 6(a) and (b) and Figures 5(a) and (b) to conclude that using decision-theoretic experience sampling can help achieve a good performance level at lower probe costs for this domain. The methods will likely have similar attributes for other domains given the general principles underlying the cost-benefit analysis that lies at the heart of the ESM probing methodology.

By looking at the evolution of the system from one week to another for the two subjects (Table 3), we can see a gain in total recognition accuracy for all of the policies. Similarly, examining the statistics averaged over the multiple subjects in different conditions (Table 4), shows that there is an improvement in recognition accuracy over time for all of the approaches. We conducted a one-way Repeated-Measures ANOVA and observed that the difference in accuracy for week 1 versus week 2 was significant for Random probing (Wilk's $\Lambda = 0.576$, $F(1, 9) = 6.62$, $p = 0.03$). The difference in accuracy for week 1 versus week 2 was not significant for rest of the policies. These results suggest that besides random probing, the ESM methods do not gain much performance from probing in the second week. This finding is likely explained by the fact that the probing decisions in these policies are guided by data seen in the past; leveraging data collected earlier helps the methods to better use downstream probes more efficiently than the way that collected data is used by the random strategy.

SURVEY

Following the study, we asked all the users to take a survey. The survey was designed to acquire information about the participants experience with the BusyBody system and to see if the experience with the different ESM methods might have had influence on their general feelings about employing experience sampling within adaptive applications. We note that there was no way for the subjects to distinguish between an authentic probe triggered to collect the label or a validation probe issued to collect the test points; consequently, the responses to the survey considers all the probes experienced by the users. Out of the 37 subjects, only 1 subject failed to submit the survey.

Specifically, we asked the users to rate how annoying the system was using a 10 point scale (1-not at all and 10-highly annoying). Figure 7 summarizes the responses to this question. An independent t-test (DT-dyna vs. rest) was conducted to evaluate the hypothesis that the BusyBody system with DT-dyna policy was assessed as less annoying than the BusyBody that employed any of the other three policies. The test was significant, $t(34) = 2.04, p = 0.049$, suggesting that there is a considerable difference between the annoyance for the DT-dyna versus the other probing policies.

We also asked the participants to estimate how often they recalled the probe to have appeared. Figure 8 summarizes the responses. Again, DT-dyna was perceived to have issued

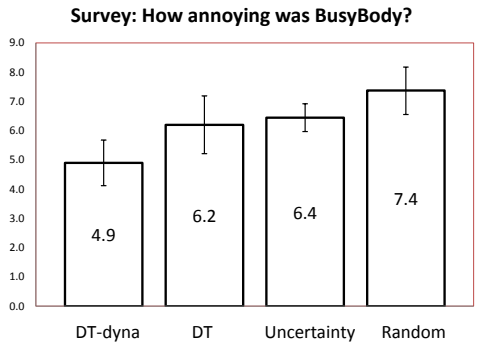


Figure 7. Average of the scores reported by the users on how *annoying* BusyBody probe was (1-not at all, 10-highly annoying). The average score attained by DT-dyna is significantly less than the score for the other probing policies. The error bars show the standard error.

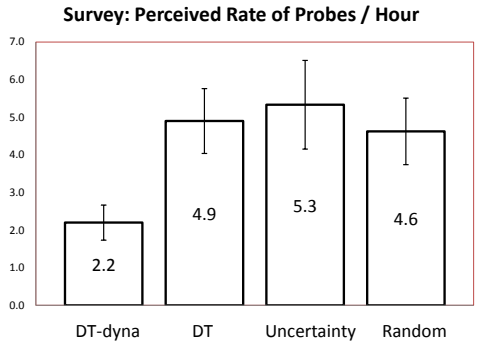


Figure 8. Average of number of probes per hour *perceived* by the users of the different BusyBody probing policies. The version of BusyBody that used DT-dyna probing was recalled as having significantly fewer numbers of probes than recollections about the other probing policies. The error bars show the standard error.

a smaller number of probes than the other policies. This effect was identified as significant using the independent t-test ($t(34) = 3.98, p = 0.008$).

Long term vs short term deployment

One of the key challenges in building systems that interact with humans is that the users and the context may change constantly; such dynamism is a challenge for systems that have only been allowed to learn during a pre-assigned training phase. However, users may be averse to a system that learns continuously by probing similar to that used in BusyBody. We sought to explore users preferences about using systems that employed ongoing selective probing. Specifically, we asked the participants if they thought that the system would be better if the probing was limited to a well-defined training phase.

Most of the participants (24 out of 36) in the survey specified a preference that the probing be limited to a short-term period. However, we found that majority of the subjects who

had leaned towards allowing a longer-term deployment of a probing system (5 out of 11, see Table 5) had experienced the BusyBody probe guided by the DT-dyna method. This finding suggests that for long-term usage, automated ESM methods that take into account the costs and benefits associated with probes might be more acceptable than the ones that are random. Having access to probing over extended periods of time could allow systems to deliver higher performance than those with a limited training period.

We found it interesting that two subjects who were not in DT-dyna group explicitly requested a smarter probing methodology. One stated: *"I would want something that did more inference, and perhaps asked me for verification at select times only."* The other participant said: *"For this approach to be effective, the probes have to be smart and not disruptive or annoying. Whether a probe comes up or not should be influenced by what you are doing on the machine rather than a preset frequency."*

Is the training effort worth it?

We also explored in the survey how worthwhile the subjects believed the training to be for the goal of building systems that could predict their cost of interruption. Specifically, were they willing to put up with the probing to train a system that could eventually become good at inferring their interruptability?

Most of the subjects answered favorably (22 out of 36). Again a large number of subjects who said yes (see Table 5) were probed using the DT-dyna policy. We were impressed how much influence the subtleties of a training procedure could have on the longer-term perceptions of the overall value of building and using adaptive systems.

Note that for this specific question we received more positive responses (8 yes, 1 no) for the Uncertainty policy than for DT-dyna (7 yes, 2 no). The difference here is insignificant and does not imply that users believe that the Uncertainty policy is better appreciated than DT-dyna; they were answering a question more generally about the potential value of investing in the training of a system that could provide long-term payoffs.

DISCUSSION

Experience sampling methods can be disruptive and, thus, frustrating, especially when the users are interrupted to provide training input over a long period of time. One of our aims was to develop and test methods that could minimize the disruption and annoyance associated with such systems by using information available to selectively query about the most valuable probes and to consider users' states of interruptability in an ongoing manner as data is collected.

Beyond the quantitative results, the results of the qualitative evaluation, suggests that probing procedures that take into account the data dynamics, user state, and the cost of probing can provide a way automate ESM in a manner that will be acceptable to users. When compared to the commonly used random probing policies, the annoyance levels of the users were lower when the ESM system probed in a manner

Question: *Would the system be better if the probing was simply limited only to a training phase of, say, a few days?*

Method	Yes	No	No Response
DT-dyna	5	5	0
DT	6	3	0
Uncertainty	9	0	0
Random	4	3	1

Question: *Would the training effort required to learn how to predict your interruptability be worth it to build a system that, when running, would work on your behalf to minimize interruptions?*

Method	Yes	No	No Response
DT-dyna	7	2	1
DT	3	5	1
Uncertainty	8	1	0
Random	4	3	1

Table 5. Responses to survey questions.

that balances the performance gain and the cost of the interruption. Furthermore, we believe that the ability of the DT-dyna approach to adapt to the changing data dynamics and context switches is a valuable feature; the method enables an ESM probing system to harness all prior data before seeking out new information from the user. This in turns results in significantly fewer probes—a behavior that we found to be associated with a higher tolerance by the users.

We note that annoyance of subjects is a subjective assessment which might correlate with other variables that are unmodeled by the system. For instance, one of the participants wrote the following in the survey: *"I did notice a correspondence between how annoyed I got with the probe and my external stress levels. In other words, if I'm less stressed or anxious in general, the probe becomes less annoying. I think this is more a reflection of how my psyche operates than a direct reflection of probe though."* This comment suggests that modeling the subjects affect, and perhaps other aspects of the subject and context, might enrich the accuracy of the models of the cost of the ESM probes. Such extensions would be natural additions to the DT and DT-dyna methods.

CONCLUSION AND FUTURE WORK

We reviewed four different methods for automated experience sampling aimed at the task of assessing data from users for building predictive user models. We performed a field study and found that the capabilities of experience sampling based on active learning using decision-theoretic probing were valuable for minimizing the numbers of probes and maximizing the classification accuracy for the task of building predictive models of the cost of interruption. Key concepts were illustrated in the context of the BusyBody system.

We believe that ESM methods that perform context-sensitive cost-benefit analysis of the value of information for probing decisions will be valuable for the use of machine learning to personalize the performance of computing applications. Beyond promising to enhance the performance of predic-

tive models, the annoyance and disruption associated with probes can be lowered significantly by using probing strategies that pay heed to the user state and the costs associated with disruptions.

On research directions, we are interested in characterizing the value of different experience sampling policies for building predictive models for other target inferences and domains. We are also pursuing several extensions and uses of selective probing, including the application of decision-theoretic strategies for extended-horizon and offline ESM. With *extended-horizon* ESM, we relax the real-time constraint, and allow probes to be delayed until a less costly time for an interruption. With delayed probes, the system inquires about past experiences, using landmarks, recordings, or other means to refer to events at earlier times. Extended-horizon probing policies should consider the cost of the potential loss of fidelity of assessments with increasing delay. With *offline* ESM, selective probing policies are applied to reduce the effort required of subjects to label recordings of prior activities. As an example, subjects in [7] were asked to perform the tedious task of viewing several hours of over-the-shoulder video recordings of their desktop activities and to assess changes in their cost of interruption. Decision-theoretic ESM could be applied in a sequential manner to identify the portions of recordings that would be most valuable to assess, significantly reducing the time and effort required to perform such offline assessments.

ACKNOWLEDGMENTS

We thank all the subjects who participated in the field study and Paul Koch for helping with BusyBody components.

REFERENCES

1. Albrecht, D. W., Zukerman, I. and Nicholson, A. E.: Bayesian Models for Keyhole Plan Recognition in an Adventure Game. User Modeling and User-Adapted Interaction Volume 8 (1998).
2. Cesa-Bianchi, N., Conconi, A. and Gentile, C.: Learning probabilistic linear-threshold classifiers via selective sampling. Conference on Learning Theory (2003).
3. Fogarty, J., Hudson, S. E. and Lai, J.: Examining the Robustness of Sensor-Based Statistical Models of Human Interruptability. Proceedings of CHI (2004).
4. Freund, Y., Seung, H. S., Shamir, E. and Tishby, N.: Selective Sampling Using the Query by Committee Algorithm. Machine Learning Volume 28 (1997).
5. Froehlich J, Chen, M. Y., Smith, I. E., and Potter, F.: Voting with Your Feet: An Investigative Study of the Relationship Between Place Visit Behavior and Preference. UbiComp (2006).
6. Horvitz, E.: Principles of Mixed-Initiative User Interfaces. Proceedings of CHI (1999).
7. Horvitz, E. and Apacible, J.: Learning and Reasoning about Interruption. International Conference on Multimodal Interfaces (2003).
8. Horvitz, E., Apacible, J. and Koch, P. BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption. Conference on Computer Supported Cooperative Work (2004).
9. Horvitz, E., Breese, J., Heckerman, D., Hovel, D. and Rommelse, K.: The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. Uncertainty in Artificial Intelligence (1998).
10. Horvitz, E., Jacobs, A. and Hovel, D.: Attention-Sensitive Alerting. Uncertainty in Artificial Intelligence (1999).
11. Intille, S. S., Rondoni, J., Kukla, C., Anacona, I., and Bao, L.: A Context-aware Experience Sampling Tool. Extended Abstract in Proceedings of CHI (2003).
12. Kapoor, A., Burleson, W. and Picard R. W.: Automatic Prediction of Frustration. International Journal of Human Computer Studies Volume 65 (2007).
13. Kapoor, A. and Horvitz, E.: On Discarding, Caching, and Recalling Samples in Active Learning. Uncertainty in Artificial Intelligence (2007).
14. Lawrence, N., Seeger, M. and Herbrich, R.: Fast Sparse Gaussian Process Method: Informative Vector Machines. Neural Information Processing Systems (2002).
15. Lewis, D. D. and Gale, W. A.: A sequential algorithm for training text classifiers. International Conference on Research and Development in Information Retrieval (1994).
16. MacKay, D.: Information-Based Objective Functions for Active Data Selection. Neural Computation Volume 4(4) (1992).
17. McCallum, A. K. and Nigam, K.: Employing EM in pool-based active learning for text classification. International Conference on Machine Learning (1998).
18. Minka, T. P.: A Family of Algorithms for Approximate Bayesian Inference. PhD Thesis, Massachusetts Institue of Technology (2001).
19. Muslea, I., Minton, S. and Knoblock, C. A.: Active + Semi-supervised Learning = Robust Multi-View Learning. International Conference on Machine Learning (2002).
20. Shen, J. and Dietterich, T. G.: Active EM to Reduce Noise in Activity Recognition. International Conference on Intelligent User Interface (2007).
21. Tong, S. and Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification. International Conference on Machine Learning (2000).
22. Zhu, X., Lafferty, J. and Ghahramani, Z.: Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining at ICML (2003).