

Inferring Player Engagement in a Pervasive Experience

Joel E. Fischer, Steve Benford

The Mixed Reality Laboratory

University of Nottingham

Nottingham, NG8 1BB, UK

{jef, sdb}@cs.nott.ac.uk

ABSTRACT

We investigate the prediction of player engagement to address temporal issues arising from the long-term character of pervasive experiences such as interruptibility, mutual player state awareness, disengagement and synchronization on re-engagement. We introduce a model that operationalizes engagement in terms of the *elapsed* and *response time* in game messages. We designed and conducted an experiment based on the experience-sampling method to evaluate our model on the basis of a long-term SMS-based game called *Day of the Figurines*. Statistical analysis supports the hypothesis that player engagement can be predicted by the continuous data properties *elapsed time* and *response time*. Our findings point towards further research towards the adaptation of pervasive experiences to the player's temporal context.

Author Keywords

Pervasive experience, engagement, context-awareness, experience-sampling method

ACM Classification Keywords

H5.1. Multimedia Information Systems: Evaluation/methodology.

INTRODUCTION

The work presented here draws on the lessons learned from the experience of staging a long-term pervasive game called *Day of the Figurines* (DoF) that was played by over 1,000 members of the public in Berlin, Singapore and the UK. DoF is a deliberately slow-paced game that requires players to send and receive only a few messages per day [7]. Each player controls a character by sending and receiving SMS to explore a virtual town as a story unfolds through a series of scheduled events. The player interacts to visit destinations, chat to and help other players, and receive dilemmas and missions.

As players increasingly adapt pervasive experiences to fit their everyday lives [1], the challenge becomes how to

support their temporal patterns of interaction. The majority of the player's interaction in DoF turned out to be episodic, as they dipped in and out of the game, frequently disengaging and reengaging [3]. In mobile experiences, every message to the player potentially interrupts their real world activity – and their level of engagement determines if the interruption is welcome or not. Disengaged players often felt 'flooded by messages' they continued to receive. When re-engaging, players reported that outdated messages confused them into taking actions that were no longer relevant. Some were frustrated by being 'ignored' by others as the system did not provide awareness and availability information about the potential recipient of the message.

The experience would benefit if the system could adapt to each player's level of engagement – backing off when they are disengaged and increasing activity at other times. In this paper, we investigate if it is possible to infer levels of engagement from system logs of game messages. We introduce a model that uses *elapsed time* and *response time* derived from system logs to infer levels of engagement. We describe an experiment that uses the experience-sampling method to assess the veracity of the model. Our experiment confirms that *elapsed time* and *response time* can predict levels of engagement from data in system logs, enabling unobtrusive monitoring and thus avoiding the overhead of users providing explicit feedback on availability [11]. For example, [8] points out that the users of a telephone system rarely remembered to change their availability state explicitly.

PLAYER ENGAGEMENT IN DOF

The players of DoF actively engage with the game by sending text messages. Their interaction is often organized in sequences of turns as a player action is usually a response to a game message which is then responded to, and so on. The discipline that studies turn-taking in depth is conversation analysis, which has been adopted for previous studies of technology-mediated communication [13, 14, 12]. It was especially the description of how different *conversational styles* are characterized by the lapses between turns or sequences of turns that occur when participants disengage and re-engage [13], which inspired the idea to utilize the lengths of the lapses between turns to estimate player engagement. For instance, the lapses between any two

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2009, April 4–9, 2009, Boston, MA, USA.

Copyright 2009 ACM 978-1-60558-246-7/09/04...\$5.00.

turns usually remain small in a *focused conversation*, for example in the temporally clearly-framed encounters on the telephone [13]. A *bursty conversation style* is characterized by lapses that occur between sequences of turns. An *intermittent conversation style* was observed in instant messaging [12], for which long lapses between the turns occur that would normally form a single sequence of turns at talk [13]. All three forms of conversational style were seen in DoF. The interaction throughout the entire experience can best be described as *bursty*, as lapses occurred during sequences of activity. Within this, however, there were many examples of *focused* interactions. The *intermittent* style became apparent when players sent messages with long lapses between each message. Building on these ideas, the core issue for this paper is whether the lengths of the lapses between players' actions can predict their shifting levels of engagement.

Operationalizing engagement

Elapsed time (et) is the time interval between two player activities; in the case of DoF, the time period between two messages. *Et* is a property of a message a player sends to the game, it represents the time period since they last sent a message. As the pattern of player interaction varies between different styles as described above, the engagement is to be expected to vary along with the temporal patterns of interaction. Lapses between activities vary in length, as players alternate between different levels of engagement and disengagement. Low values of *et* mean that the lapses between turns are small. In turn, high values arise when the player is disengaged, i.e. does not send a message to the game.

Response time (rt) is defined as the time interval since the player last *received* a message, as opposed to *et*, which is the time interval since the player last *sent* a message. Figure 1 outlines the concept of *et* and *rt*.

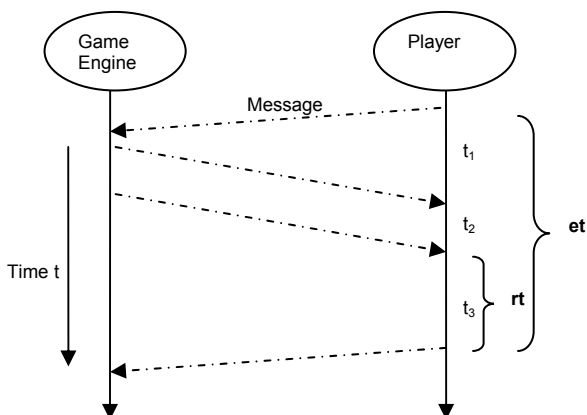


Figure 1. The concept of elapsed (*et*) and response time (*rt*).

In the example pattern of interaction depicted in figure 1, the player receives two messages before he sends one.

The *rt* in this case is the time t_3 . The *et* is the period since the player last sent a message, or the sum of $t_1 + t_2 + t_3$.

Et and *rt* potentially capture different aspects of engagement. We propose that *et* might provide an overview of the general pattern of player activity throughout, capturing the overall intensity with which they are interacting. *Rt*, on the other hand, might potentially capture responsiveness to the characteristics of particular messages received from the game (e.g., do players respond differently to dilemmas from the game compared to chat message from other players), although further exploration of this question falls outside the scope of this paper, and remains a topic for future research.

EXPERIMENTAL DESIGN OF AN SMS-BASED ESM

The experience sampling method (ESM) has been designed to accommodate for the shortcomings of post-hoc evaluation techniques such as diaries and questionnaires that rely on the retrospective assessment of an experience, which may lead to distorted assessments of the actual experience, e.g. through social or cultural connotations that are applied post-hoc [6]. The ESM introduces an *in situ* approach to measuring the quality of experience by prompting participants to fill out questionnaires during their *current* experience with a signaling device, i.e. a ‘beeper’, over longer periods of time [5]. ESM has previously been applied to the evaluation of ubicomp applications [4], to an examination of attitudes towards availability and interruption [8], and to evaluate a model that aims at estimating interruptibility to enhance human interaction and computer-mediated communication [9].

Adapting ESM for evaluating engagement in DoF

Our hypothesis is that short *elapsed* and *response time* intervals between two player actions indicate high engagement. We adapted the ESM to test this hypothesis. Since players interact with the game using their mobile phone we employ the same platform for the experiment. This approach has an advantage over PDA-based [4] or pager-based [8] ESM in that the participants would not have to carry expensive extra devices. The goal of the experiment is to assess if the proposed relation between data properties and engagement exists and to get a more accurate picture of which values of *elapsed* and *response time* express which levels of engagement. The participants in the study, who are playing DoF at the same time, are prompted via SMS to fill out a questionnaire on their mobile's WML browser about their current game experience three times a day – a frequency previously found to be appropriate for experiences of this duration [10]. The questionnaire asked the participant to “Please select which description best matches your current engagement in DoF.” They chose the answer from four categories “Disengaged - not playing; Passively following the game; Responding to messages; Proactively looking for action in the game.” The categories were derived

through introspection informed by the experience of playing the game. We chose nominal instead of ordinal categories to guarantee a high comparability among the answers in order not to introduce implicit assumptions about the participants' understanding of the concept as it would have been when asking them to rate 'on a scale'.

The participants are prompted dependent on their in-game behaviour, so that the probability of being prompted n minutes after engaging with the game decreases exponentially. One of the three daily prompts is sent at a random time. This implementation reflects the idea to probe engagement right after players sent a message to test the hypothesis that engagement is high at that time, while trying to avoid a systematic bias towards *always* probing participants after they just sent a message and thereby biasing towards a highly engaged sample. For each time-stamped ESM questionnaire answer, parallel in-game properties were computed to analyze their correlation with the participants' self-reports. This augments the questionnaire response with data about in-game behaviour: When did this player last send a message (*elapsed time*)? When did they last receive a message (*response time*)? Does this relate to their self-reported state of engagement?

The instance of the game that was studied for this experiment was staged publicly at the Southbank centre in London from June 12th until July 5th 2008, with 331 players playing the game in total. We recruited 16 participants to play DoF and participate in the survey; none of them were involved in the development of DoF. Several players' characters died in the game before it was over, and one left the game. Overall, the players played half of the time of the game that lasts for 24 days.

In total, the 16 participants played 201 days of the game and completed 101 self-reports, while they were prompted to fill out 564 self-reports in total. This is an approximate return rate of 1/6 – on average every sixth prompt resulted in a successful completion of the questionnaire. No systematic differences were found for the participant sample and the rest of the 331 players in terms of the average number of messages sent and average *elapsed* and *response time*; indicating that that the prompts – which may be regarded as “reminders” of the game – did not bias our participants towards more engagement than the rest of the players.

RESULTS

Here, we will focus on self-reported states of engagement and its interrelation with *elapsed time* and *response time*. The distribution of the self-reported levels of engagement is as follows: 10 times participants reported they were “disengaged – not playing”, 38 times they reported to be “passively following the game”, 39 times “responding to messages” and 11 times they reported to be “proactively searching for action in the game”. 3 missing values lead to

Levels of engagement	N	Mean	Std. Deviation	Minimum	Maximum
disengaged	10	3457.3	2372.2	71.0	7144.0
passive	33	1228.9	1408.5	12.0	6871.0
responding	29	1356.9	1824.1	5.0	6988.0
proactive	10	271.4	478.7	0.0	1338.0
missing value	3	841.0	597.2	162.0	1285.0
Total	85	1408.4	1783.1	0.0	7144.0

Table 1: Elapsed time by levels of engagement.

excluding those three cases from the subsequent inferential analysis.

We tested statistically if the observed *et* and *rt* differ with respect to the participants' self-reported levels of engagement. Note that, for each participant, the variables are not computed for their first reply to the questionnaire since the variables describe continuous properties *since they last answered* a questionnaire. This explains the $N=85$ (101 responses minus 16 first replies to the questionnaire, one per participant). An analysis of variance (ANOVA) for the means of *et* for the four different states of engagement (see table 1) showed F to be significant beyond the .01 level: $F = 5.53$; $p < .01$. The size of the effect proves to be large with an *Eta squared* value of .21 according to Cohen's classification of effect size. The trend of the distribution is as predicted: with increasing *et* the levels of engagement decrease. An exception is that the mean of *et* for the level “responding” is larger than the mean for the level “passive”. However, considering that the variable is not normally distributed a look the median, which is less sensitive to outliers, changes the picture. The median for “responding” is smaller than that for “passive”, thanks to the lessening of the outliers' weights. Inspecting the differences closely, a Mann-Whitney test showed that whereas the differences between “disengaged” and the three other categories and “proactive” and the three others are all significant at the .01 level, the categories “responding” and “passive” are not significantly different from each other.

Similar results are obtained for the interrelation of self-reported levels of engagement and *rt*. Again, an ANOVA yielded significant differences for the means of *rt* for the different categories of engagement. It showed F to be significant at the .01 level: $F = 3.47$; $p = .012$. Again, the mean for the *rt* was larger for “responding” than for “passive”. And yet again, the median changes this trend. A Mann-Whitney test confirms the assumption that the differences in *rt* for participants that report to be “passive” or “responding” do not differ significantly as well.

DISCUSSION

The study empirically affirmed our hypothesis. The fact that significant differences for *elapsed time* and *response time* in correlation to self-reported levels of engagement

were found support the empirical veracity of our model of engagement. The model allows for the inference of player engagement from the pattern of ongoing activity similar to [15], operationalized by the lapses that occur between the player's actions. We even predicted the trend correctly: With increasing level of engagement from "disengaged" to "proactive", the values for *et* and *rt* decrease. This also supports the choice of categories of engagement we presented to the participants: The categories worked to express different *ordinal levels* of engagement. Yet, the fact that *et* and *rt* did not differ significantly for the two middle categories "responding" and "passive" suggest to either test with a larger sample or perhaps merge the categories into one. The presented means of *et* and *rt* provide a first impression of the allocation of periods of time to levels of engagement. However, the substantial standard deviations imply some uncertainty in the reliability of the allocation of certain values of *et* to levels of engagement. The findings remind us that, with respect to the prediction of individual context, it is difficult to interpolate across individuals. Sensitivity is needed when building a system that adapts the experience to the temporal context of the player.

CONCLUSIONS

In general, the results show that game inherent continuous properties can be utilized as predictors of engagement. The finding that *elapsed time* and *response time* varied significantly for different *levels of engagement* suggests that the concept should be maintained, refined and developed further to become part of an inference mechanism in a context-adaptive system for long-term pervasive experiences. If the system is able to detect the player's level of engagement, it could hold back certain messages when assuming disengagement and inform other players of their unavailability and send summaries on re-engagement. However, as uncertainty in the allocation of time slots to levels of engagement remains, a context-aware system would need to offer explicit mechanisms for the user to correct an adaptation [2]. The findings may also inspire designers of other long-term games to find out which properties of their players' interactions indicate engagement to 1) acknowledge that different specific levels of engagement exist in their games and 2) design the game to adapt to the players' different levels of engagement so as to motivate or facilitate re-engagement when sensing disengagement or by rewarding high engagement.

The experience-sampling method is an interesting tool to collect quantitative data in a setting where the participants are locally dispersed and temporally independent. Thus, it is ideal to collect subjective self-reports in pervasive experiences in a fashion that they become immediately available to sound statistical analysis and to meaningful interlacing with data that is gathered throughout the experience such as log files.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the EPSRC through the *Challenge of Widespread Ubiquitous Computing* project (EP/F03038X/1).

REFERENCES

1. Bell, M., Chalmers, M., et al., Interweaving mobile games with everyday life, *Proc. CHI 2006*, ACM.
2. Bellotti, V. and Edwards, K., Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human-Computer Interaction*, 16, 2 (2001), 193 - 212.
3. Benford, S. and Giannachi, G., Temporal Trajectories in Shared Interactive Narratives, *Proc. CHI 2008*, ACM, 73-82.
4. Consolvo, S. and Walker M., Using the experience sampling method to evaluate ubicomp applications. *Pervasive Computing*, IEEE, 2, 2 (2003), 24-31.
5. Csikszentmihalyi, M., Larson R., and Prescott, S., The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6, 3 (1977), 281-294.
6. Csikszentmihalyi, M. and LeFevre, J., Optimal experience in work and leisure. *Journal of Personality and Social Psychology*, 56, 5 (1989), 815-822.
7. Flintham, M., Smith, K., Benford, S., Capra M., Green, J., Greenhalgh, C., et al. A slow narrative-driven game for mobile phones using text messaging, *Proc PerGames 2007*, Salzburg, June 11-12 2007.
8. Hudson, J.M., Kristensen, J., Kellogg, W., Erikson, T., "I'd be overwhelmed, but it's just one more thing to do": availability and interruption in research management, *Proc. CHI. 2002*, ACM.
9. Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., et al., Predicting human interruptibility with sensors: a Wizard of Oz feasibility study, *CHI 2003*, ACM.
10. Kubey, R. and M. Csikszentmihalyi, Experience Sampling Method Applications to Communication Research Questions. *Journal of Communication*, 46, 2 (1996), 99-119.
11. Middleton, S.E., Roure, D.C.D., and Shadbolt, N.R., Capturing knowledge of user preferences: ontologies in recommender systems, *Proc. K-CAP. 2001*, ACM.
12. Nardi, B.A., Whittaker, S., and Bradner E., Interaction and outeraction: instant messaging in action, *Proc. CSCW 2000*, ACM.
13. Woodruff, A. and Aoki, P.M., How push-to-talk makes talk less pushy, *Proc. GROUP 2003*, ACM.
14. Woodruff, A. and Aoki, P.M., Conversation Analysis and the User Experience, *Proc. CHI 2004*, ACM.
15. Yu, C., Aoki, P.M., and Woodruff, A., Detecting User Engagement in Everyday Conversations, in *Proc. INTERSPEECH 2004*, ISCA.