A Hybrid Evaluation Approach for Ubiquitous Computing Environments

Mark Burnett and Chris P. Rainsford

Information Technology Division
Defence Science and Technology Organisation
Department of Defence, Fern Hill Park
Canberra ACT 2600
AUSTRALIA
{mark.burnett,chris.rainsford}@defence.gov.au

1 Introduction

In this paper we consider the problem of providing an evaluation framework for ubiquitous computing environments. We distinguish between qualitative and quantitative approaches to evaluation and argue that a hybrid framework may be the most suited to ubiquitous computing. We start by discussing evaluation approaches with reference to other evaluations within computer science and conclude the paper by describing a hybrid approach to evaluation for ubiquitous computing.

2 Evaluation Techniques

Evaluation techniques are used by researchers in many fields to benchmark the performance of proposed approaches to solving research problems. For example, in the realm of information retrieval the NIST-sponsored TREC events have been widely used to compare different information retrieval systems [trec.nist.gov]. Prior to this test collections were small, evaluation was ad-hoc and cross-community acceptance of evaluation results was patchy. TREC has been well organized, independently assessed, and operates to widely agreed quantitative metrics.

Another example of a research evaluation is the RoboCup competition where teams of robots play soccer against each other [robocup.org]. This competition is defined by clear goals and allows easy evaluation of results. Interestingly, RoboCup also has a stated goal for the year of 2050 to have a team of fully autonomous humanoid robots capable of winning against the world soccer champions. This use of an overall objective is a worthwhile approach to defining the target problem and worth considering for ubiquitous computing.

Not all evaluations need be purely quantitative. As an example, the data mining community has held data mining challenges in conjunction with major conferences e.g. PKDD in Europe. Their chosen format of evaluation allows the qualitative comparison of diverse techniques against a fixed set of data and hence encourages the examination of diverse and novel techniques. The broad differences between qualitative and quantitative evaluation are described in Table 1.

Typical Characteristics	Quantitative	Qualitative
Target Problem	Well defined, agreed metrics.	Loosely defined, new, unknown metrics
Pros	Well defined outcomes, easy to compare candidates	Allows exploration, encourages diversity, human subjectivity allowed.
Cons	Narrow focus, does not accommodate human subjectivity.	Not well defined, hard to compare candidates, inconclusive.
Evaluation Outcomes	Improved performance	Discovery of new approaches, identification of strong points

Table 1. Comparison of Quantitative and Qualitative evaluation strategies.

Looking at the characteristics of successful evaluations we see the following elements emerging:

- What's being evaluated is clear;

- Quantitative measures of evaluation are clear, and agreed on by practitioners in the field;
- Independent evaluators are used;
- Barriers to participation for research groups are low in terms of the resources needed;
- The evaluations encourage exploration and diversity in approaches.

3 Ubiquitous Computing And Evaluation

There are a number of different elements that define ubiquitous computing:

- *Ubiquity/pervasiveness* – lots of computing devices;
- *Invisibility* – in many situations the interface between the user and device disappears and the device effectively becomes invisible;
- *Connectedness* – the devices are networked in some way to other devices and information;
- *Context-awareness* – the system is aware of the context of users and provides an intelligent bridge between the computational environment and the real world.

Within this space there are a number of sub-problems amenable to quantitative evaluation such as speech processing and biometrics techniques (for example the Information Technology Laboratory of NIST works to provide benchmarking and evaluation of a number of technologies directly relevant to ubiquitous computing [www.itl.nist.gov]). At the same time, there are also a number of problems for which evaluation cannot be easily quantified. We argue that since ubiquitous computing attempts to contextualise information and services for individual users, subjectiveness is an inherent property of this domain. Rather than evaluating a single quantifiable measure one may also want to measure the effectiveness of the overall information ecology provided by the pervasive computing environment. We now consider *how* and *what* to evaluate.

3.1 How to evaluate a ubiquitous computing environment

Our approach to the *how* question is two-fold: (i) choose a well-defined environment for experimentation; (ii) choose some clearly defined work related tasks within the environment for evaluation. A *smart room* [*Scientific American*, April, 1996] is one such environment where testers in the space would perform a series of tasks, which may or may not be known in advance by participants. Each evaluation could involve a different set of tasks, or maybe involve some unknown task categories. Evaluation metrics then involve the speed and accuracy of the tasks performed. As important, we argue, are qualitative measures of the usefulness of the systems. The essence of ubiquitous computing is its invisibility and context-awareness. User-awareness means being able to intelligently gauge user's needs and preferences in support of the tasks they are engaged in. This implies the need for subjective measures for how well this awareness is manifest in the smart room.

3.2 What to evaluate in a ubiquitous computing environment

There are two types of experimentation possible with this environment. The first is an evaluation of the software infrastructure and services available in support of specified tasks. In this situation the device and networking infrastructure of the room is standardized, and participants provide their own software. Participants could either build their own smart room to the requirements or use someone else's. The latter option points to a low barrier to participation.

The second type of evaluation is a generalization of the first and allows participants to build their own smart room. Device, networks and software are all put into the mix for researchers to experiment with. Drawbacks of this approach are that the question of what is being evaluated is not immediately obvious, and costs for entry – in terms of the diverse expertise necessary to build a smart room as much as the cost involved – may be prohibitive.

4 Conclusion

Overall the approach described here has a number of key strengths. Situating the experiments in a smart room reflects the nature of ubiquitous computing as situated in the real world and assisting people in commonplace tasks. Quantitative evaluation measures for tasks can be relatively easily defined e.g. speed, cost, accuracy. In addition, qualitative measures based on user's experiences can be catered for by allowing evaluators to give feedback on what aspects of their environment they did or did not find helpful. It must be noted however, that the tasks would need to be carefully designed so that domain knowledge and uneven or time-dependent tester competence in a task does not skew the evaluation.