Evaluation Methodologies for Context-Aware Computing

Jean Scholtz, Defense Advanced Research Projects Agency
jscholtz@darpa.mil, +1 (703) 696 4469

INTRODUCTION

In this position paper I present a case for evaluation of context-aware computing systems and discuss some possibilities for evaluation of context-aware computing systems. Evaluation should serve two purposes: first it should enable the researchers to measure progress in a field; secondly, evaluation can help researchers better understand research issues and produce a more focused effort to overcome particular hard research problems.

We currently lack a systematic user-centered evaluation methodology for interactive systems. Research makes progress by evaluating work and addressing the current limitations. For context-aware computing, I envision an evaluation of how well systems can reason about context. But the most desirable evaluation is user-centered. Does context-aware computing deliver useful information to the user at a time and in a form that it can be comprehended? And can the user's cognitive load be lessened by context-aware systems? If it's an entertainment task, does a context-aware system increase the user's enjoyment? Will it be possible to define a user-centered evaluation methodology and a set of benchmark tasks to measure progress in context-aware computing? I propose that we need evaluation methodologies in order to facilitate progress in the area. HCI is about evaluation and many evaluation methodologies have been developed and used within our community. However, we lack higher-level evaluations that can increase our overall understanding and help to advance HCI research.

TEXT RETRIEVAL CONFERECE

The information retrieval community has been extremely successful as a community because of a common evaluation effort. The Text Retrieval Conferences (TREC) co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA) allow researchers to use the same set of data to compare results, substantially advancing the state of the art in information retrieval (http://trec.nist.gov/). TREC was started in 1992 as part of the federal TIPSTER Text program (http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/) . Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, the TREC workshop series has the following goals:

- encourage research in information retrieval based on large test collections;
- increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

In accordance with the last goal, TReC expanded to include a number of different comparisons including multi-lingual retrieval tracks, interactive tracks, and question and answering tracks to focus more on the "user-experience." A common corpus of data and a set of retrieval tasks are provided to researchers. At TREC, the results of the various algorithms are examined for the different types of queries. This helps researchers gain insights into advantages and disadvantages of the various approaches and has led to refinements and sometimes combinations of approaches. As the field has progressed, so has the evaluation. TREC is now in it's 8th year. This year's conference had the largest number of participants yet. New tracks have been added to include different languages and interactive queries. A new track of question answering on the web has recently been added. Similar evaluation efforts at NIST have been used for speech recognition and mixed initiative dialogues.

One important aspect of these evaluations is that the research community that uses the evaluations also designs them. The researchers must believe that the evaluations will be helpful to them in assessing research progress. Additionally for evaluations to be useful to the researchers, the metrics must also be developed by the community.

The interactive track in TREC has proven more difficult than other tracks as it involves user studies. Training and expertise of users is an issue. As many systems are involved in the track, users cannot use and compare all systems. The approach taken here has been to use a control system and to have users complete tasks on both the control system and the comparison system. Thus all TREC participants compare their system to the standard control.

USER DISTRACTIONS

Let's consider how one might evaluate a context-aware system or ubiquitous computing. One of the first issues is to identify meaningful metrics. These metrics must useful to the researchers to identify progress. The goal of ubiquitous computing is to minimize "user distractions" given that the user's attention is the limiting quantity in human-computer interaction. The term "user distraction" or "user overhead" refers to tasks that a user currently performs that should become the responsibility of a ubiquitous, distributed system. If we assume that the main purpose of ubiquitous computing is to provide a user with information whether the purpose is for work, entertainment, or education, then there are several problems the user currently has in accessing this information: distribution of the information, the volume of information, the level of information, and access in a mobile domain. Distribution concerns the manner in which the information is requested and disseminated to a user. Currently the user has to realize a need for information and make this request to the computer using whatever interactions are available in the user interface. Volume is certainly an issue; there is an increasing amount of information available and it is growing at a rapid rate. It is extremely difficult for the user to sort through all and ascertain the information that is relevant to the situation. The level of information is also a problem. Much information is in tables, charts, etc. while some exists as video data or image data with the bulk of information text-based. The user is faced with having to assimilate much information and interpret it to answer a higher level need. The last barrier to information access is support for mobility. As users move between environments, computing resources change. Data has to be move so that it is accessible; applications have to be restarted in a different computing environment. Much of this responsibility falls on the end user currently and is another distraction that the user must deal with in order to access information.

If we recast these barriers to information access (distribution, volume, level, and mobility) into research areas for removing distractions, we can see that there is research needed at the human-computer interaction level, at the information management level, and at the middleware level. At the human-computer interaction level, we can address distribution distractions. Distribution currently requires the user to request all the information that is needed. By having the system be aware of the user's situation and inferring information needs, some of the distribution can occur in a push or automatic fashion.

At the information management level, we can eliminate distractions having to do with the volume and level of information. The volume can be reduced by using context as a filter in an information push mode. The user will always want to request information as well. If these requests can be made at a higher level of interaction that is more related to the task the user is doing, the user is relieved of having to break down a high level task into lower level information needs. For example, if the user wants to locate nearby entertainment, a question might be "what is there to do this afternoon within 10 or 12 blocks?" The user doesn't want to have to query a database to see what museums are open, what galleries are having shows, what plays are on, what movies are shown, etc. At the information management level we are interested in formulating those queries automatically and consolidating the resulting information into a summary that answers the user's high level question. The terminology "task" is not meant to suggest that the user is always engaged in work. Tasks apply to entertainment as well. In the case of entertainment, we might consider assessing whether context-aware systems increase the pleasure of the user. For educational uses, we might measure the amount and retention of learning with a context-aware system as compared to a benchmark system.

At the middleware level, we are interested in making the user's interaction with a ubiquitous computing environment seamless. This means that the task the user is doing needs to flow from one set of computing devices to another with no interaction on the part of the user. This is another type of context that the system must be aware of and respond to: the set of computing devices and connectivity that the user currently has access to.

CONTEXT-AWARE EVALUATIONS

User distractions could be used as an evaluation mechanism for context-aware computing. One methodology for an overall evaluation would be to select a set of representative scenarios. Whittaker, Terveen, and Nardi [1] argue for benchmark tasks for HCI as well. These could range from work to entertainment to domestic tasks. Baseline measurements could be taken to determine the steps a user currently has to undertake to accomplish each of these tasks. As we develop context-aware systems, we should demonstrate them using these baseline tasks and measure the steps now needed to determine if we have eliminated user distractions. As we make inferences about what a user needs, the evaluation must take into account the correctness of our reasoning. If we draw incorrect conclusions, the user will most likely have to make an explicit request – correcting a system mistake could cost more steps (user distractions) than the original baseline task.

Another possibility would be to do evaluations at the inference level. Suppose that we knew the basic task that a user was doing and had some knowledge of the basic information needs for that task. Then one possibility would be to record situational information as the user goes about doing this particular task and noting what actions the user does at a particular time. This could be tagged as ground truth and stored in a database. It would then be possible to compare the inferences made by looking at the situational data to the actual user request. This type of evaluation, of course, assumes that we have identified the appropriate situational data (both hard and soft sensor data, user profiles, user history, etc.) to collect and are able to store it in a time-stamped database. This type of evaluation could be a first step to help compare various context extraction and inferencing algorithms.

SUMMARY

There are many research issues that would have to be addressed in order to identify a reasonable evaluation methodology including, how to identify, capture, store, and annotate situational data. Is the situational information needed different from one user to another? What types of tasks are representative and should be selected for evaluation? What is the minimum set of contextual information that is needed to make reasonably accurate inferences? What metrics are suitable for educational and entertainment activities? These questions will also have to be considered as we develop tools and environments for context-aware computing. Research in evaluation needs to be iterative. As the research progresses, the tasks used for evaluation must change as well to continue to challenge the research efforts. Evaluation for context-aware systems will not be trivial but should be considered a necessary part of the actual research. Evaluation should be used to measure progress and to help researchers identify outstanding investigation areas. The methods suggested here are very preliminary thoughts but as we move to ubiquitous and aware computing environments we need to devote some research efforts to evaluation methodologies.

REFERENCES

1. Whittaker, Steve, Loren Terveen, and Bonnie Nardi. Let's stop pushing the envelope and start addressing it: A reference task agenda for HCI. *Human Computer Interaction*, 15, 75-106.