Simulating Events to Generate Synthetic Data for Pervasive Spaces

Andres Mendez-Vazquez
Mobile and Pervasive Computing Lab
University of Florida
E301 CSE Building, PO Box 116120, Gainesville, FL 32611
amendez-@cise.ufl.edu

Abdelsalam (Sumi) Helal
Mobile and Pervasive Computing Lab
University of Florida
E301 CSE Building, PO Box 116120, Gainesville, FL 32611
helal@cise.ufl.edu

Diane J. Cook
School of Electrical Engineering and Computer Science
EME 121 Spokane Street
Box 642752
Washington State University
Pullman, WA 99164-2752

Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research, April 4, 2008, MIT, Massachusetts, USA.

Abstract
Actual data collected from real deployments is ultimately the best data that can be had and used in evaluating systems and new concepts. However, available data may not be specific enough to drive certain evaluation goals. Therefore, it is necessary to propose, as an alternative, a method to generate synthetic data for pervasive spaces. A first step in this direction is trying to simulate a chain of events in the sensor output space by the use of pattern recognition models. Specifically, we propose the use of Markov Chains to generate patterns of events due to the fact that they have been used successfully to classify chains of events in machine learning and computer vision. We believe this strategy will make for a more realistic simulation of sensor activity in a pervasive space.

Keywords
Pervasive spaces, synthetic data generation, simulation, Markov chains, state machines, spike Poisson generator.

ACM Classification Keywords
I.6.1 Simulation Theory: Types of simulation (continuous and discrete).

Introduction
In recent years, it has become obvious that the increasing cost of building pervasive spaces and the lack of applicable data makes research in this area a bit more than difficult [1]. Not everybody has a large budget to build a pervasive space to test new algorithms and ideas. And even if budget is not an issue, it is usually very time consuming to generate enough data for a meaningful collection of patterns or events. For instance some of the available data sets are useful to some researchers but not all, depending on the events captured and the specific sensors available in the space where the data was collected. Another

difficulty is recruiting participants to perform all of the activities under all possible conditions or contexts that we wish to consider. We believe it is necessary to look for alternative ways to create focused simulations [8][9] of events to fine-tune the analysis and understanding of the specific pervasive space and its associated algorithms and applications. This early stage simulation can help researchers evaluate their ideas quickly and with reasonable accuracy.

In this paper, we propose to use of a combination of Markov chains [2,11], Poisson processes [3,12] and probability distributions [4] for the simulation and generation of synthetic data for pervasive spaces. Initially, The Markov Chain will be used to generate pattern of activities based on *apriori* knowledge of the target user behavior. Markov chains have been successfully used in the detection and recognition of patterns of activities [5,6,10,13,14,15,16]. Here, we utilize Markov Chains in the inverse way – in generating the patterns of activities. Now, we need a way to generate random time stamps in a given time interval for each event or sensor output to be simulated. A popular way to generate time stamps in a given time interval for events is by the use of the Poisson distribution. Finally, the sensor output values will be generated using probability distributions based in *apriori* assumptions like the output range for each sensor.

An immediate benefit of synthetic data generation is the degree of control that can be attained. For example, we can control what patterns of activity are going to be generated, what kind of noise can be added to the patterns of activity, etc. This is something that can only be done in a limited way in a real pervasive space. For this and other reasons, we believe that the proposed idea to simulate to events or sensor outputs in smart spaces is worthwhile.

PROPOSED ALGORITHM

Here, we will describe the basics of the proposed data generation algorithm. More details can be found in the Technical Report describing this work (see [19]).

Synthetic Data Algorithm

1. Set Time Stamp = 0. This is done to have a point of reference for each element of the Markov chain.
2. Set NODE = Initial Node. Normally, the initial node represents the beginning of a specified activity.
3. Set N. This is the number of sensor outputs to be generated out of restricted set of possible sensor given a series of probabilities
4. *Chain* a storage for the Markov Chain
5. For i = 1 to N
    a. Sample a multinomial distribution with state NODE and probability $p_1,....,p_k$.
    b. Change NODE = Sample.
    c. Then:
        i. Generate a time stamp T by using the spike Poisson generator.
        j. Generate a value O using the distribution for *NODE*.
        k. $Chain = Chain \bigcup \langle NODE\_ID, T, O \rangle$
6. End For
7. Return *Chain*.

Validation

In this section, we present a brief evaluation of the proposed algorithm by looking at an actual data set, simulating the events, and then comparing actual and

syntheic data. We use actual data collected through an actigraph [17] device to gather data from a subject activity. Each entry in the data set contains four fields,

1. **Label -** sleep, walk, sit, read, exercise, house working.
2. **Time** – in hours.
3. **Outdoor temperature** – in Fahrenheit.
4. **Energy expenditure** – in calories.

It is natural to use the labels as our nodes in the state machine for the Markov chain. In this particular case we did not use the spike Poisson spike generator because the discrete times are measured in hours in the data set. The finite state machine for the simulation of a 50 hour period in the daily living activities of a person is given in Figure 1.

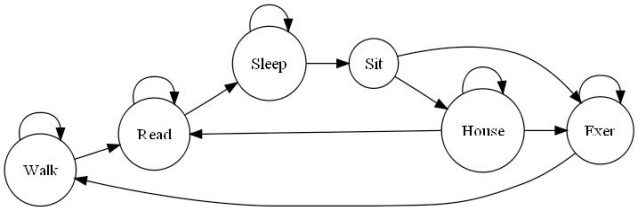

Figure 1 State machine for the simulation of activities

The edge probabilities and sensor range activities were assigned to try to resemble in the most faithful way the original activity graph (Figure 3). In this activity graphic the X axis is the time axis, and the Y axis corresponds to the calories consumed.

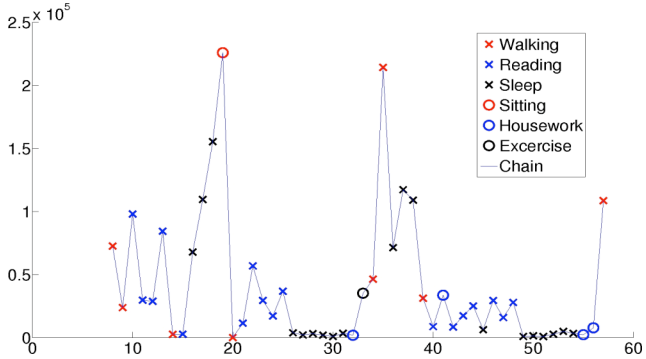

Figure 2 The activity graph (actual data).

We can see that the simulated activity graph (Figure 3) resembles the real activity graph in the succession of patterns of activity. For example, it is possible to see from hour 28 to 45 in (Figure 3) that the person transitions from walk to read activities, and then finally goes to sleep. This is similar to the patterns of activities in (Figure 2) from hour 20 to hour 32. In order to quantifiy these similarities in the patterns, we use dynamic time warping [18].

Conclusion

We believe that generatig smart space data synthetically is very useful and needed. We believe that actual data collected in real deployments is best data to use, but practically, for new ideas and concepts, it is much more realistic and feasible to rely on synthesized data. We briefly presented our data generation algorithm and provided an assessment of its accuracy by comparing its output with an actual data set.

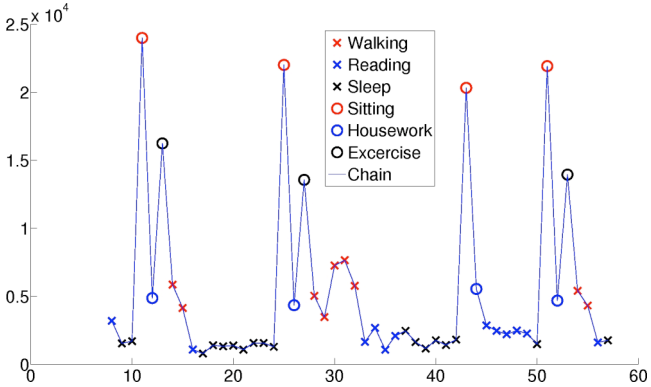

Figure 3 The simulated activity graph.

Citations

[1] I. Armac and D. Retkowitz, "Simulation of Smart Environments," *IEEE International Conference on Pervasive Services*, pp.257-266, 15-20 July 2007.

[2] P. Bremaud, Markov chains. Gibbs fields, Monte Carlo simulation and queues, 3ed ed., New York:Springer, 2008.

[3] D.R. Cox and V.I. Isham, *Point Processes,* Chapman & Hall, 1980.

[4] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove: Duxbury Press, 2002.

[5] S. K. Das, N. Roy and A. Roy, "Context-aware resource management in multi-inhabitant smart homes: A framework based on Nash H-learning," Pervasive and Mobile Computing, Volume 2, Issue 4, Special Issue on PerCom 2006, November 2006, Pages 372-404, ISSN 1574-1192.

[6] A. Dupuy, J. Schwartz, Y. Yemini, and D. Bacon, "NEST: a network simulation and prototyping testbed," *Communications of ACM* 33, 10 (Oct. 1990), 63-74.

[7] D. Heeger, "Poisson Model of Spike Generation," handout, Sept. 2000. http://www.cns.nyu.edu/~david/handouts/poisson.pdf

[8] T.G. Kim, *Theory of Modeling and Simulation*, 2nd ed. Academic Press, 2000.

[9] A. Law and W. D. Kelton, *Simulation Modeling and Analysis*, McGraw-Hill, 1999.

[10] G. Mazeroff, "Markov models for application behavior analysis," in Proceedings of the 4th Annual Workshop on Cyber Security and information intelligence Research: Developing Strategies To Meet the Cyber Security and information intelligence Challenges Ahead, May 12 - 14, vol. 288.

[11] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[12] D.L. Snyder and M.I. Miller, *Random Point Processes in Time and Space,* Springer-Verlag. 1991.

[13] H. Thimbleby, P. Cairns, and M. Jones, "Usability analysis with Markov models," in *ACM Transactions on Computer-Human Interaction*, 8(2), 99-132, 2001.

[14] J. Yin, Q. Yang, D. Shen, and Z. Li, "Activity recognition via user-trace segmentation," *ACM Transactions on Sensor Networks*, 4(4), 2008.

[15] G. M. Youngblood, Diane J. Cook, Lawrence B. Holder, "Managing Adaptive Versatile environments," Pervasive and Mobile Computing, Volume 1, Issue 4, Special Issue on PerCom 2005.

[16] J.W. Yoon, D.H. Jwa, J.H. Kim, H. Park and Y.S. Moon, "Gaussian Distribution for NPC Character in Real-Life Simulation," *International Conference on Intelligent,* pp.132-135, 11-13 Oct. 2007.

[17] H. Pigot, B. Lefebvre, J.G. Meunier, B. Kerhervé, A. Mayers and S. Giroux, "The role of intelligent habitats in upholding elders in residence," *5th international conference on Simulations in Biomedicine*, April 2003, Slovenia.

[18] L. R. Rabiner and B. Juang *Fundamentals of speech recognition*, Prentice-Hall, Inc., 1993

[19] A. Mendez, A. Helal, D. Cook, "Simulating Events to Generate Synthetic Data for Pervasive Spaces," University of Florida Technical Report, available from www.icta.ufl.edu/projects/publications/chiws09.pdf