A Feasibility Study: Mining Daily Traces for Home Heating Control

Dezhi Hong and Kamin Whitehouse
Department of Computer Science
University of Virginia
{hong, whitehouse}@virginia.edu

ABSTRACT

HVAC systems take up the largest portion of utility bills in a home and they are also large electricity consumers nationwide. Recent work has been focused on automated control based on occupancy prediction, where some look into historical patterns while others leverage real-time position information of the occupant. We believe combining these two techniques could help predict when someone will come home. In this paper, we propose to look at an occupant's "landmarks" (time of leaving home and work) in history and make predictions of arrival times. Our approach requires the minimum efforts for heating controls from users. We evaluate the model on the data from 4 users and show the potential 8.3%–27.9% energy savings as well as 14.9%–59.2% reduction in miss time.

Keywords

Energy, home heating, daily traces, prediction

1. INTRODUCTION

Heating, ventilation and cooling (HVAC) contributes most to a home's energy bills, accounting for 48% of residential energy consumption in the U.S. and 61% in the U.K., 64% in Canada where climates are colder [4, 11, 5]. This has incentivized efforts from both the industry and academia for years to increase the efficiency of home heating as well as reduce GHG emissions. Programmable thermostats were found to be not as effective as expected [6, 12] after being used for a while. An alternative solution is to turn on/off the HVAC system in the home automatically based on occupancy prediction, which is a mainstream research direction.

State-of-the-art research on occupancy prediction is primarily focused on two kinds. The first kind predicts occupancy in home setting [10, 13] or commercial buildings [1] with statistical models by deploying numbers of motion sensors in the monitored environment. The second kind, on the other hand, exploits real-time GPS data to estimate travel home time and dynamically controls the HVAC system [8]. In general, automated home heating control is a trade-off between energy use and comfort in the sense that too aggressive a prediction can always guarantee the temperature while incurring unnecessary waste of energy, and vice versa. Our approach concentrates on improving the prediction of occupant's arrival times thus making better preheating decisions.

In this paper, we propose a model to predict when a person will come home each day by matching the times that they leave the house and leave the office with his historical commuting timelines. We exploits the intuition that commuting patterns of a person can change from day to day but will tend to repeat themselves in a long run. We also believe that the position information of a person outside the home throughout a day gives a hint on when he might be heading home. We present a preliminary analysis of 4 users' daily trajectory data in a period ranging from 120 days to 180 days [3], and evaluate our model in respect of arrival time prediction error, energy savings as well as average miss time (total time when an occupied house is below setpoint temperature). The main contributions of this paper are:

- We describe a model for predicting an occupant's arrival times by matching each day with historical days that have similar landmarks and evaluate the prediction error with real data from 4 users.

- Compared to the approach used in [10] as a baseline, our model improves on both energy consumption about 8.3%–27.9% and miss time about 14.9%–

3rd International Workshop on Mobile Sensing, April 8–11, 2013, Philadelphia, Pennsylvania, USA.

59.2% without requiring homeowners to do any programming for occupancy schedule prediction.

2. BACKGROUND AND RELATED WORK

There have been remarkable efforts dedicated to the research on home heating savings. Prior work investigated to improve programmable thermostats [7] and some alternatives employed sophisticated controls based on real-time GPS information [8] or occupancy prediction [10, 13, 1]. Programmable thermostats control the HVAC equipment based on a setback schedule and waste energy in several ways such as when occupants leave home earlier than the scheduled turning off time or an improper setback temperature is used. Gao and Whitehouse [7] design and evaluate a self-programming thermostat to fix the problems of programmable thermostats by automatically choosing the optimal setback schedule based on historical occupancy data. However, that system still produces a static schedule and such a static schedule must sacrifice either energy or comfort considering changes in occupancy patterns.

Gupta et al. [8] propose to control the heating system leveraging real-time GPS data. They make online estimations of travel-to-home time with GPS coordinates sent to a web service and dynamically adjust the HVAC system in the house. They assume the home can always be brought to the setpoint temperature in the time that an occupant travels home from a certain location. Their achieved energy savings are not significant compared with a simple always-on manual thermostat (the system resumes heating whenever the temperature drops below the setpoint) and a pre-programmed thermostat. The limitation of this work is that they only use the status in a current day but doesn't leverage historical information.

Lu et al. [10] exploit an HMM on motion sensor data to determine when to turn off the system and use historical arrival times to decide when to preheat. They choose a reactive approach (the leaves and arrivals of occupants are detected both with motion sensors) as a baseline and evaluate their approach's miss time and energy savings for each day in a week using the EnergyPlus simulator. Their usage of sensor data is limited inside the home but doesn't know where the person currently is while being outside home, and they generally use all the historical days to model the occupancy patterns in order to make predictions.

Scott et al. [13] formulate an online algorithm for occupancy prediction by matching with historical days. Standing at any time point in a day, they compute the occupancy probability in the coming hours based on similar historical days and set a threshold probability to determine when to preheat the house. They choose a simple pre-programmed thermostat control as a baseline and evaluate their system in energy savings and miss time. Again, they don't make use of the position information of a person outside the home. Krumm and Brush [9] also present an occupancy prediction algorithm that gives occupancy probabilities at different times of day. However, they compute a representative day (i.e., Monday) for each day of the week, being unable to accustom to changing occupancy patterns.

Our approach combines the techniques in aforementioned work by leveraging historical GPS data to model an occupant's commuting patterns in order to predict occupancy. We look at typical time points (i.e., when they leave the house and the workplace) of similar days in history and make predictions of arrival times based on a statistical distribution. We choose the approach used in [10] as a baseline to evaluate energy savings and miss time of our model.

3. APPROACH

As a whole, there are two parts in home heating controls: determining when to turn on and when to turn off the HVAC system. This paper targets at better predicting when an occupant will come home (i.e., provides some insight into the preheating decision), so we will not discuss about when to turn off the heating system.

The daily commuting timeline of a person can vary from day to day but will show some regularity, so looking into the historical data to search for days with similar landmarks could provide some confidence in prediction. By looking for conditionally similar days in the past, we compute the distribution of arrival times in those similar days and, based on the distribution, predict when an occupant will arrive home in the future.

3.1 Conditional Distribution Model

Basically, each particular day is represented by a two-element tuple,

$$d_i = (t_h^i, t_w^i)$$

where, t_h^i is the time of leaving the house of a particular day i and t_w^i is the counterpart for leaving the workplace. These two time points are the features we leverage for prediction and extracted from daily traces. A day, where the occupant both leaves and comes home once which is the most common case, is represented with one such tuple. For a day where the occupant comes and leaves home several times throughout the day, we check the timestamps of two adjacent events: if the last leave and a new arrival happen within 30 minutes, which means the house in unoccupied for less than 30 minutes, we deem this period as occupied; likewise, if the last arrival and a new leave happen within 30 minutes, we consider this period as unoccupied. Otherwise, we use extra tuple(s) for such a day.

With every day represented as a tuple, for a given day on which we want to predict when the occupant will come home (we call it a "query day"), we look into the historical data and search for days with similar times of leaving home and leaving work. Here we define "similar" as such that the times of leaving home and leaving work on that day satisfy both,

$$\left|t_h^i - t_h^q\right| \le \varepsilon, \quad \left|t_w^i - t_w^q\right| \le \varepsilon$$

where, t_h^i is the time of leaving home in a particular historical day i, t_h^q is for the query day, ε is the allowed error in time difference between a historical day and a query day, say, 5 minutes. And t_w^i, t_w^q are for the time of leaving work.

After finishing the search, we compute the arrival times distribution for all similar days and then compute the arrival time error between the query day and each similar historical day. Each day is partitioned into slots of 5 minutes and the arrival time of that day is expressed as the ith slot it falls into. Therefore, error is calculated at a granularity of 5 minutes.

4. EVALUATION

We evaluate our model along with the baseline on a data set comprising 4 users' daily trajectory, and each user's data collection period spans from 120 days to 180 days. First, we compare the prediction error in two models. The error distribution for each user, which is a normalized frequency count, is obtained with leave-one-out cross validation on all his historical days. Second, we calculate the back-of-the-envelope energy savings based on the error distribution. Third, we evaluate the average miss time in both models.

4.1 Data Set and Pre-processing

The data set (obtained from [2]) was collected with a service running in the background to automatically collect the user's mobility and to trace sensor usage time [3]. To get the ground truth, participants were asked to explicitly label the places and kept a record of places they had visited with entrance and departure times. Therefore we need to perform some basic pre-processing on the data set to filter out the times of leaving home, leaving work and arriving home in each day.

4.2 Baseline

The baseline model in this paper is the one exploited in [10]: instead of searching for similar days in the historical data, it simply uses every single day in the past to compute the distribution for arrival times and then obtain the prediction error between a query day and each historical day.

4.3 Prediction Error

The prediction error of arrival times are computed on all historical days of each user with leave-one-out cross validation, and different values of ε, from 5 minutes to 30 minutes, are used to compare the error distributions as shown in column one of Figure 1. We see that, the conditional models have more "vertical" distributions than baseline, meaning that the error distribute more aside zero. The prediction error of our model and the baseline are illustrated in Figure 2a. The bars for conditional models are the average of errors with different ε values.

4.4 Energy Savings

To obtain the statistics of energy savings, we need to first differentiate between two types of energy penalty for early and late predictions ("early" means the predicted time is ahead of the real arrival time and, similar for "late"). According to an empirical measure of different heating stages of a conventional three-stage HVAC system [10]: on average, stage-1 (preheat the house to the setpoint) needs around 24 minutes to raise the temperature by 1°F while consuming about 1.1 kWh power. Stage-2 (maintain the temperature before an occupant comes home), which is the most energy efficient stage, takes 18 minutes and 0.9 kWh. Stage-3 (react to quickly heat up the house) has the fastest response time (6 minutes) but a higher energy cost (1.6 kWh).

Generally, the cost of an early prediction is $E_{S1}+E_{S2}$, where E_{S1} is the energy consumed in stage-1, E_{S2} is the energy in stage-2 which is the actual penalty. Assuming after reaching the setpoint, the temperature drops by 1 degree in 18 minutes and recall that it takes stage-2 18 minutes to heat up the house by 1 degree with 0.9 kWh, thus the system is duty-cycled in half of each hour – the penalty would be $P_E = (60\times0.5)/18\times0.9 =$ 1.5 kWh/hour.

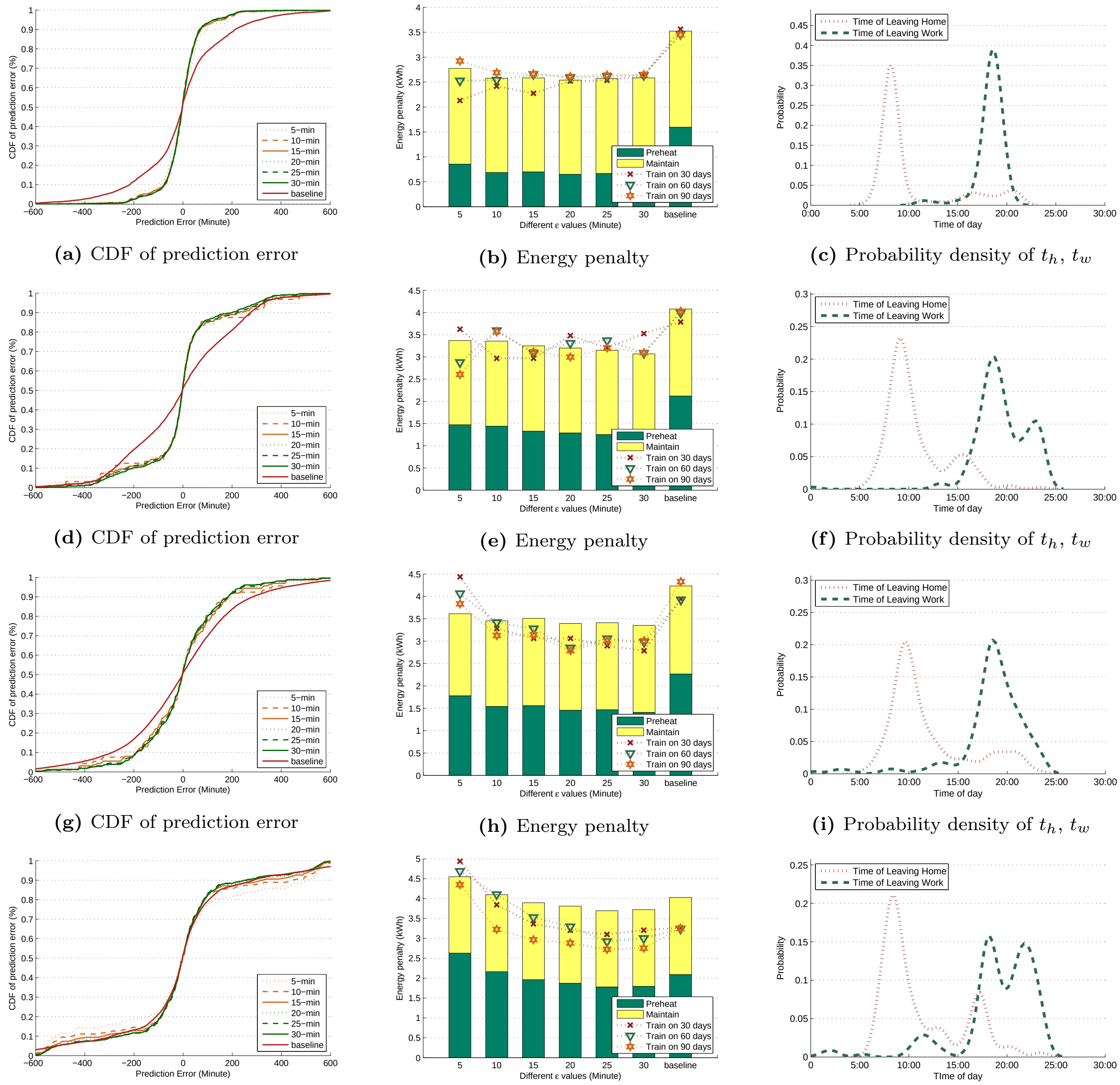
For a late prediction, the cost is simply E_{S3}, which is the energy consumed with stage-3 heating. The waste is the extra energy consumed by using stage-3 rather than stage-1, because if we predict earlier, we could preheat using less energy with stage-1. Recall the statistics earlier and assume the house will be heated up by 8 degrees – the penalty would be $P_L = (1.6–1.1)\times8 = 4$ kWh.

With the unit penalty for different prediction conditions, we compute the gross energy penalty based on the prediction error distribution using,

$$P = \begin{cases} \sum Pr(t_{err}^i) \cdot t_{err}^i \cdot P_E, \text{ if } t_{err}^i < 0 \\ \sum Pr(t_{err}^i) \cdot P_L, \text{ if } t_{err}^i > 0 \end{cases}$$

where, $Pr(\cdot)$ is the probability of a case where the prediction error is t_{err}^i, i denotes which 5-minute-long time slot the t_{err} is in. The energy penalty of different conditional distributions (i.e., different ε) for different users are shown in column two of Figure 1.

Discussion: We see that, different users have different



(a) CDF of prediction error (b) Energy penalty (c) Probability density of t_h, t_w

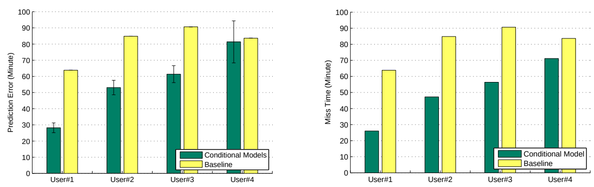
(d) CDF of prediction error (e) Energy penalty (f) Probability density of t_h, t_w

(g) CDF of prediction error (h) Energy penalty (i) Probability density of t_h, t_w

(j) CDF of prediction error (k) Energy penalty (l) Probability density of t_h, t_w

Figure 1: Each row contains the results for one user: the first column is the CDF of prediction error distribution for arrival times; the second column is the energy penalty, the optimal case of each user has energy savings ranging from 8.3% to 27.9% than baseline; the last column is the probability distribution of each user's leaving home times and leaving work times.

optimal values of ε for best energy savings. An occupant with more centralized distribution of t_h and t_w (indicating a more regular commuting pattern) has smaller difference in savings between different ε (#3 vs. #4) and tends to need a smaller ε for best energy saving (e.g., #1 vs. #2 or #1 vs. #4). Column two of Figure 1 also presents the result of different numbers of training days for each occupant (the bars are obtained using all historical days). As the number of training days varies, the optimal ε deviates in some cases (#1 and #2), but the saving of a deviated ε has only a slight difference from that of the "real" optimal value (obtained with all historical days) and the optimal ε almost converges to the "real" value when 60 days or more is used. Therefore,



(a) Prediction error　　**(b)** Miss time

Figure 2: (a) our model has 2.7%–55.8% lower prediction error than baseline; (b) there is a reduction of 14.9%–59.2% in miss time observed in our model than baseline.

the number of days used has little impact on selecting the optimal value of ε for energy savings. As observed in Figure 1, our conditional model has energy savings of 8.3%–27.9% than baseline.

4.5 Miss Time

Miss time is the time when the house is occupied but the temperature is below setpoint. Here it is calculated as the expectation of prediction error in cases that have late predictions,

$$MissTime = \sum Pr(t^i_{err}) \cdot t^i_{err}, \text{ where } t^i_{err} > 0$$

We choose the case with best energy saving out of each user (e.g., user #1 has the best energy saving when ε is 20 minutes) to compare with the baseline in terms of miss time. The result is shown in Figure 2b. Overall, the conditional models have 14.9%–59.2% less miss time than baseline.

5. CONCLUSION AND FUTURE WORK

This paper pursues a hypothesis that historical commuting timelines of an occupant extracted from GPS data can provide some insight into occupancy prediction. Pivoted on this hypothesis, we explore the historical GPS data of an occupant and propose a model for predicting arrival times by matching each day with similar days in the past. We evaluate the model on a data set comprising 4 users in terms of gross energy savings and miss time to verify the effectiveness of such a model, and show that our model is 8.3%–27.9% and 14.9%–59.2% better in energy savings and miss time respectively than baseline approach.

The next step of work includes: a) to incorporate the influence of seasonal weather changes into our model. In Section 4.4 we assume the temperature in the house drops by 1 degree in 18 minutes. Given the insulation condition of a house, the heat loss rate is associated with weather conditions, for example, the house loses heat more quickly in winters than in falls. Therefore, the unit penalty cost is dependent on weather; b) in this work, we only looked into typical time points in a day such as leaving the house home and the office, there is more information in the data set which could be helpful, such as the daily travel trajectory.

6. ACKNOWLEDGMENTS

We want to thank Yohan Chon, for providing the data set to facilitate our analysis. This work was partly funded by the NSF grant CPS-1035771.

7. REFERENCES

[1] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng. Occupancy-driven energy management for smart building automation. In *BuildSys'10*.

[2] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *UbiComp'12*.

[3] Y. Chon, E. Talipov, H. Shin, and H. Cha. Mobility prediction-based smartphone energy optimization for everyday location monitoring. In *SenSys'11*.

[4] Department of Energy. 2011 Buildings Energy Data Book. `http://buildingsdatabook.eren.doe.gov/`.

[5] Department of Energy & Climate Change. Energy consumption in the UK, 2011. `https://www.gov.uk/government/publications/energy-consumption-in-the-uk`.

[6] Environmental Protection Agency. Summary of Research Findings From the Programmable Thermostat Market. *Technical Report RPDH 15, Building Research Establishment*, 2004.

[7] G. Gao and K. Whitehouse. The self-programming thermostat: optimizing setback schedules based on home occupancy patterns. In *BuildSys'09*.

[8] M. Gupta, S. S. Intille, and K. Larson. Adding gps-control to traditional thermostats: An exploration of potential energy savings and design challenges. In *Pervasive'09*.

[9] J. Krumm and A. J. B. Brush. Learning time-based presence probabilities. In *Pervasive'11*.

[10] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. Whitehouse. The smart thermostat: using occupancy sensors to save energy in homes. In *SenSys'10*.

[11] Natural Resources Canada. Energy Use Data Handbook, 1990 -2008. `http://oee.nrcan.gc.ca/publications/statistics/handbook10/pdf/handbook10.pdf`.

[12] H. Sachs. Programmable Thermostats. *ACEEE'04*.

[13] J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar. Preheat: controlling home heating using occupancy prediction. In *UbiComp'11*.