

Activity Rhythm Detection and Modeling

Rosco Hill, James “Bo” Begole

Sun Microsystems Laboratories

2600 Casey Ave

Mountain View, CA 94043 USA

rhill@engmail.uwaterloo.ca, James.Begole@Sun.COM

ABSTRACT

We present an algorithm for detecting and modeling rhythmic temporal patterns from the record of an individual’s computer activity, or online “presence.” The model is both predictive and descriptive of temporal features and is constructed with minimal *a priori* knowledge.

Keywords

Awareness, rhythms, CSCW, user modeling, statistics

INTRODUCTION

People exhibit temporal patterns, or rhythms, in their daily behavior such as when they typically arrive at the office, take breaks, attend recurring meetings, and commute to and from work sites. Zerubavel [5] demonstrated that collocated coworkers share a sense of these patterns and use this to form expectations of each others’ availability. As more work is distributed, however, it becomes difficult to maintain awareness of coworkers’ rhythmic patterns. The goal of this research is to help restore some of the awareness of rhythmic patterns among distributed coworkers.

In past work, Begole *et al.* [1] described human-observable patterns seen in the record of individuals’ computer, email, instant messaging and phone activity collected using a prototype awareness system called Awarenex [4]. In this paper, we present an algorithm for *computer* discovery and modeling of such rhythmic patterns.

Previous researchers [2, 3] have used Bayesian networks that include time along with other factors to construct predictive models of a person’s likely location or availability. Our model differs in that it is both predictive and *descriptive* of the temporal patterns. While a Bayesian network can show the factors that contribute to a prediction, it does not describe the changes in a factor’s influence over time. Both types of model can answer a query about a specific point in time, but our model additionally describes a forecast of activity and online presence over all points in time.

This descriptive model has the ability to support applications beyond a solely predictive model. Users can visualize their temporal patterns, possibly making inferences of their own. Also, the descriptive model can infer a reason for an absence, such as “at lunch.” We have integrated this

capability into Awarenex so that people do not need to explicitly set their “away” status.

NON-PARAMETRIC MODEL

Many rhythmic patterns did not conform to common parametric statistical distributions (e.g., Gaussian, Poisson, etc.). For example, the distribution in Figure 1 has peaks at approximately 20 minute intervals as a result of being constrained to a mass transit schedule [1]. Therefore, our model does not assume parametric distributions.

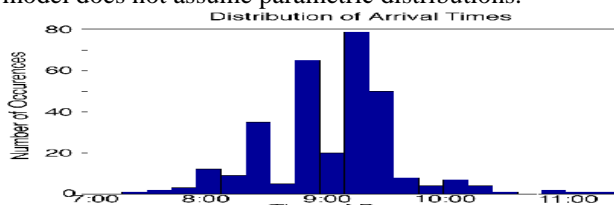


Figure 1. An example distribution that does not conform to a parametric distribution.

Another goal of our modeling technique is to minimize *a priori* knowledge of the structure of a person’s day. Beyond the coarsest generality that people tend to start activity in the mornings and end in the evenings, peoples’ patterns vary widely. Even a common feature like “lunch” does not show up in all people’s data. Because it is impossible to enumerate all of the rhythmic patterns we should look for, we discover features dynamically.

Feature Discovery

The first step in building the model is to detect points in time that potentially designate a significant rhythmic feature. We are interested in two types of pattern: (1) recurring periods of inactivity and (2) recurring transitions between locations. Recurring inactivity occurs due to recurring meetings, lunch and regular breaks. Examples of location transitions are commuting between home and office and moving from one’s office to a laboratory or other site.

To discover such patterns, we first filter the historical data according to factors that significantly influence rhythm, such as day of week, and location [1]. To detect recurring inactivity, the algorithm examines the varying average level of activity at each minute of the day. Figure 2 shows an individual’s average level of activity throughout the day on his office computer on Mondays. When the average activity crosses threshold values, it is identified as a potential feature. Upper and lower bounds are computed for threshold values to prevent small changes in activity due to natural variance (noise) from generating spurious features. We compute the upper and lower bounds by initially setting both to the 50th percentile of the activity levels and moving

each up and down iteratively to maximize the number of significant features detected while minimizing the number of small, spurious threshold crossings.

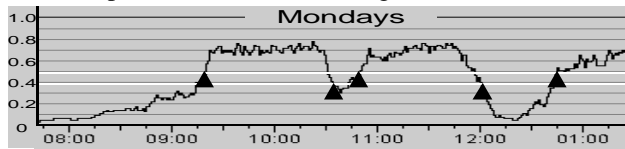


Figure 2. Thresholds identify potential features.

When activity is declining, it must cross the lower threshold boundary and when activity is rising, it must cross the upper threshold boundary. The black arrows in Figure 2 represent transition points detected by this algorithm.

The other type of feature of interest, location transitions, are found by noting instances where an inactivity period begins and ends in different locations.

Refining the properties of a feature

The next step is to refine the estimates of the start, end and durations of the potential features detected in the previous step. We search the data set for periods of inactivity that are instances of the detected potential features.

Equation 1 compares two time periods, represented as points in three dimensional space with start, end and duration as the axes. A lower value indicates greater similarity.

$$f = \sqrt{\left(\frac{\Delta start}{\hat{\sigma}_{start}}\right)^2 + \left(\frac{\Delta end}{\hat{\sigma}_{end}}\right)^2 + \left(\frac{\Delta dur}{\text{Min}(\widehat{dur}, \overline{dur})}\right)^2}$$

Equation 1. Similarity metric.

The numerators, $\Delta start$, Δend and $\Delta duration$, are the absolute value of the difference between the estimated values of the feature and a specific instance of an inactivity period.

Each dimension in the space is normalized by weighting the parameter by how far it can vary before significantly affecting similarity. Where $\hat{\sigma}_{start}$ and $\hat{\sigma}_{end}$ are the estimated standard deviations of the start and end times of the rhythm feature. Although we do not assume a normal distribution, standard deviation provides a useful measure of variance. Initial estimates are bootstrapped manually and the refinement pass is iterated until the proportional difference between the estimate and the refined value converges. We weight the duration term by the minimum of the estimated duration, \widehat{dur} , and the observed duration, \overline{dur} , to account for the skew toward longer durations.

We use a threshold distance of three to determine whether a time period is an instance of the feature. Three allows all of a period's properties to lie within one normalized unit from the rhythm feature, or any two to lie between one and two normalized units.

From this set of instances, we calculate statistics and determine the probability distribution for each property of the rhythm feature: start, end and duration. The lower portion of Figure 3 shows the rhythm features extracted for Mondays in the office. The distributions for the start, duration and end are drawn on different rows to avoid overlap.

Note that "start of day" and "end of day" are special case, *one-sided*, features that do not have a duration distribution.

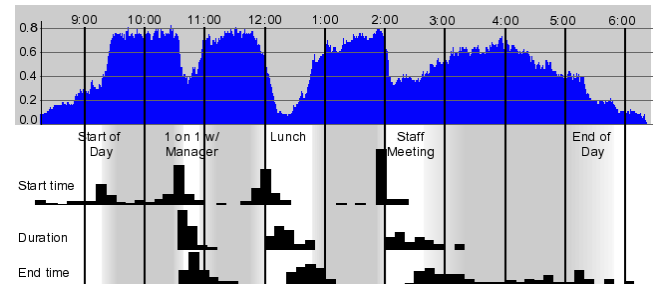


Figure 3. Average activity levels (upper) result in rhythm model (lower) with probability distributions for start, duration and end of each feature.

In this example, the start of day rises gradually until a sharp climb around 9:20. The meeting at 10:30 generally takes less than 30 minutes. Lunch regularly begins around noon and lasts between 15 and 45 minutes. This person attends a staff meeting starting promptly at 2:00 with an irregular duration, often ending before 3:00. The end of day is diffuse with short peaks around 4:50, 5:20 and 5:50.

RHYTHM NAMING

We use a simple naming algorithm. The feature closest to noon is named "Lunch." A feature that corresponds to a recurring appointment is named after that appointment. A location-transition feature is named after the locations in which it begins and ends, such as "Office to Home." Users may specify different names.

CONCLUSIONS

We have presented an algorithm to build a model of the temporal patterns in a person's computer activity and on-line presence. To ensure privacy, full details of the model are not available to end users but are used to support applications such as predicting when someone will be "present" and automatically setting their "away" status in an instant messaging system, such as "at lunch" or "commuting from home to office." Such applications help restore rhythm awareness among members of a distributed team.

REFERENCES

1. J. Begole, J. Tang, R. Smith and N. Yankelovich, "Work rhythms: Analyzing visualizations of awareness histories of distributed groups", *Proceedings of CSCW 2002*, ACM Press, pp.334-343.
2. E. Horvitz, P. Koch, C. Kadie, and A. Jacobs, "Coordinate: Probabilistic Forecasting of Presence and Availability", *Proceedings of the 2002 Conference on Uncertainty and Artificial Intelligence*, July 2002, AAAI Press, pp. 224-233.
3. S. Hudson, J. Fogarty, C. Atkeson, J. Forlizzi, S. Kiesler, J. Lee and J. Yang, "Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study", *Proceedings of CHI 2003*, ACM Press, to appear.
4. J. Tang, N. Yankelovich, J. Begole, M. Van Kleek, F. Li and J. Bhalodia, "ConNexus to Awarenex: Extending awareness to mobile users", *Proceedings of CHI 2001*, ACM Press, pp. 221- 228.
5. E. Zerubavel, *Hidden Rhythms: Schedules and Calendars in Social Life*, Chicago: The University of Chicago Press, 1981.