High Accuracy Context Recovery using Clustering Mechanisms

Dinh Phung†, Brett Adams†, Kha Tran†, Svetha Venkatesh† and Mohan Kumar‡

†Department of Computing,
Curtin University of Technology,
GPO Box U1987, Perth, WA 6845, Australia
{d.phung,b.adams,k.tran,s.venkatesh}@curtin.edu.au

‡Department of Computer Science and Engineering
The University of Texas at Arlington,
Box 19015, Arlington TX 76019, USA
mkumar@uta.edu

***Abstract*— This paper examines the recovery of user context in indoor environmnents with existing wireless infrastructures to enable assistive systems. We present a novel approach to the extraction of user context, casting the problem of context recovery as an unsupervised, clustering problem. A well known density-based clustering technique, DBSCAN, is adapted to recover user context that includes user motion state, and significant places the user visits from WiFi observations consisting of access point id and signal strength. Furthermore, user rhythms or sequences of places the user visits periodically are derived from the above low level contexts by employing a state-of-the-art probabilistic clustering technique, the Latent Dirichlet Allocation (LDA), to enable a variety of application services. Experimental results with real data are presented to validate the proposed unsupervised learning approach and demonstrate its applicability.**

I. Introduction

The increasing number and capability of mobile devices such as smart phones and ultraportable computers has brought new opportunities and challenges for ubiquitous computing [31]. Mobile computation brings greater variation in a user's situational and interactional context, and two useful constituents of this context are location and level of motion. Location often correlates with certain activities or roles [22], and aspects of the user's physicality, such as motion, are also indicative of activity and the user's affordances (e.g. interruptibility). This information can drive applications at many levels, from automated battery management to assistive systems (e.g., for the visually impaired). Device convergence has made available a number of sensing and communication technologies, including Bluetooth, WiFi and GPS, for extracting these elements of context. In this paper, we design, develop and evaluate novel, high accuracy mechanisms for extracting user contexts in indoor environments. User motion level and significant locations, defined as places where a user spends time, are extracted from raw WiFi signals in a timely, unsupervised, and accurate manner, in existing wireless network infrastructures. The paper provides an example of how this fundamental information can be used to discover such higher-level contexts as the user's daily routines or *rhythms* in an unsupervised setting.

Much work has been done in the use of GPS to derive representations of significant locations in outdoor situations. There has also been significant activity in localization from signatures that penetrate or originate indoors, such as WiFi, GSM, and Bluetooth. Work aimed at characterizing the physical state of a user has tended to make use of sensors that aren't as readily available as ambient radio signatures, such as thermometers, galvanic sensors and accelerometers. A brief review of relevant work is provided in Section IV. In [28] location and orientation estimations based on Bayesian filtering of Received Signal Strength (RSS) justifies the use of WiFi signals for extracting location context. The Locadio positioning system of Krumm et al. [17] uses WiFi signals to infer whether or not a user is moving based on the variance of signal of the strongest access point (AP), with an accuracy of 85%. The noisy, sparse nature of WiFi signatures renders Gaussian assumptions problematic. Another shortcoming of variance-based methods is the requirement for training. Moreover, in the original setting of [17], prediction is made with a latency of 20 seconds, which disqualifies the approach from real-time applications, such as navigation assistance for the visually impaired.

Rather than viewing motion state detection as a supervised classification problem, we cast it as an unsupervised and incremental clustering problem. A window of consecutive WiFi signatures observed from the same location, when the user is still, are likely to be similar, and thus form a cluster as opposed to those when the user is moving. Similarly, if WiFi signals observed during a user's daily life are collated, locations where the user spends time repeatedly, for example at their desk at work, will also emerge from a clustering process. We define a measure of distance between two WiFi observations appropriate to their characteristics, notably allowing for missing data from the vectors of AP signal strength. We use a density-based technique, DBSCAN [10] and its incremental version [9], to recover user motion level and significant locations. Use of incremental DBSCAN allows for motion level classification with latency under 2s, which is suitable for many real-time applications. We conduct comprehensive experiments to compare variance-based methods with our density-based approach for detecting user state. We achieve up to 95% in accuracy with the clustering technique, proving that our method is more robust with noisy and incomplete WiFi data.

978-1-4244-3304-9/09/$25.00 ©2009 IEEE

We experiment with detection of significant locations, using pre-filtering to remove observations when the user is moving, resulting in an accuracy of above 97%.

To further motivate the extraction of motion state and significant locations, we also present a technique for discovering user behaviour over time, termed *rhythms*. It has been shown that travel episodes often correspond to hidden agendas or 'social projects' [5], and we posit that a similar situation occurs at the finer resolution of, say, the office. Discovery of these rhythms offers potentially rich information about user intent and activity. We adapt a probabilistic graphical model, Latent Dirichlet Allocation (LDA)[4], for this task. LDA is an unsupervised probabilistic clustering technique used to discover latent topics from bags of words in text by finding co-occurrences of words in documents. Here, significant locations and their observed times are extracted and are mapped to words. These are then collated over a day and become analogous to a document. The latent topics discovered by LDA in this way are interpreted as user rhythms. We experiment with the discovery of rhythms for a user over the course of a one month period. It is worth noting from the perspective of assistive systems that the incidence of strict routines is even higher among the visually impaired, presumably due in part to the desire to decrease the number of variables that might induce danger or inconvenience for themselves or others, making rhythms more compelling in this application domain.

Our contributions to user context extraction in indoor situations include: (i) a novel WiFi distance measure and unsupervised algorithms for high accuracy motion classification in real-time; (ii) discovery of significant indoor locations at fine resolutions; and (iii) formulation and extraction of fine-scale user behaviour over time as latent topic discovery.

The ability to infer the context of the mobile user is a vital, foundational component of a broad array of pervasive computing applications. We present work enabling both richer representation and more accurate extraction of aspects of context, and hence the significance of this work is potentially great. It can serve as a basis for both annotation and prediction at a number of levels of the services stack, from context-sensitive device resource and interface management, to semi-automatic calendaring, personal life logs and collaboration tools, personalized push-information such as advertising, and navigation assistance for the visually impaired. In a shared context, this information can aid market research, surveillance and urban planning. Importantly, the absence of a requirement for calibration and use of existing infrastructure make for a low barrier to deployment.

II. A System for Context Detection using Clustering Mechanisms

This section begins with an overview of the system, together with examples of its envisaged setting and uses by way of motivation. Separate sections are then devoted to motion classification, significant location extraction, and rhythm detection, respectively.

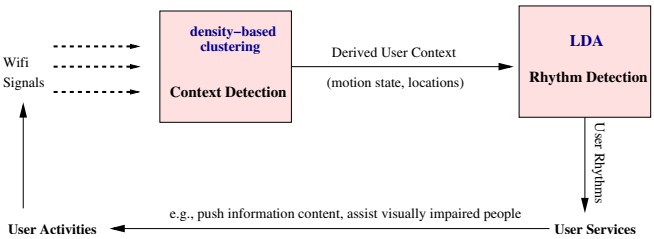


Fig. 1. System overview: context recovery using clustering.

A. System Overview

The envisaged setting of the algorithms detailed below is any mobile device equipped with a WiFi receiver. The system would typically run as a background process, making context information available as it is extracted. The input to the system consists of time-stamped vectors of received signal strength , each identified by the broadcasting AP: {time, AP id, RSS}. This is depicted on the left of Figure 1. Motion classification is performed with the arrival of each new WiFi sample (after an initial, small startup latency period) and is available immediately, for example, for driving agent or interface behaviour. E.g., when navigating, points at which the user becomes stationary are candidates for issuing new information. Referring to Figure 1, motion classification is an output of the *Context Detection* box. If the device also has the ability to store a historical record of WiFi samples, significant location discovery can also begin immediately, resulting in a growing representation of where the user is spending time. E.g., in the course of a normal work day, a handful of locations might be discovered corresponding to the office, cafe, library and a colleague's room. Significant locations are also depicted as outputs of context detection in Figure 1. These locations can be used as annotations to associate activities or media items (e.g., this is the set of applications you run at this location; you took these photos in the same place). If appropriate, labelling these locations meaningfully would be performed as a secondary activity, e.g. via active learning prompts, user-derived sources such as a calendar, or centrally-sourced such as beacon databases or pre-calibrated maps [2]. Finally, at the coarser resolution of days and weeks, rhythm detection becomes appropriate. As depicted on the right side of Figure 1, the rhythm detector accepts the user's history of time-stamped landmarks and yields patterns of behaviour in the user's whereabouts. E.g., discovered rhythms might correspond to: an average work day, involving the office and home; a work day that includes collaboration or shopping; and weekend routines that have little overlap. Rhythms, in addition to constituting a higher-order object for annotation (e.g., these photos were taken at work, but not an average work day), provide the basis for prediction. E.g., the user typically doesn't appear at these locations over the weekend.

B. WiFi observation distance

As we desire to cluster WiFi observations, we require a measure of distance between two such observations. In

theory, the relationship between RSS and distance for a given AP is inverse squared, and at first glance, modelling these points according to a Gaussian distribution and then performing hypothesis testing on the concentration of these points may offer a straight solution. However, there are a number of factors that complicate this model in practice: RSS is attenuated by physical structures and other environmental factors, which result in relatively high signal variability. Moreover, measuring the distance of a pair of *sets* of APs is complicated by missing values from one observation to the next. This also renders clustering algorithms that utilize Gaussian properties (e.g., GMM) unsuitable. One advantage of density-based clustering approaches is a degree of freedom in the formulation of a suitable function of distance between two observations.

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ be the set of all access points available. For a WiFi observation p, let P be a subset of $\{1, 2, \ldots, N\}$ denoting the set of AP indices observed and X_P be the actual set of APs. E.g., if $P = \{2, 5\}$ then $X_P = \{x_2, x_5\}$. Furthermore, we denote by $y_i^{(p)}$ the corresponding RSS reading in observation p from the source x_i. Given two WiFi observations p and q, denoting the common APs set by $C = P \cap Q$, the distance between them is defined as:

$$\text{dist}\,(p,q) = \sqrt{\frac{1}{|C|}\sum_{j\in C}\left[y_j^{(p)} - y_j^{(q)}\right]^2}$$

where $|C|$denotes the cardinality of set C. By this distance measure, only shared signal strength from common APs between two observations is taken into account, any missing ones donot affect the distance.

In practice, the spatial proximity of two WiFi observations affects the difference between X_p and X_q . Let $\eta = |C| / \max(|P|, |Q|)$ and using a threshold $\eta_0 \in [0, 1]$, the distance between p and q is adjusted to:

$$\text{dist}\,(p,q) = \begin{cases} \sqrt{\frac{1}{|C|}\sum_{j\in C}\left[y_j^{(p)} - y_j^{(q)}\right]^2} & \text{if } \eta > \eta_0 \\ +\infty & \text{otherwise} \end{cases}$$

Intuitively, the introduction of η_0 is to account for the case when the difference of observed APs in two observations is too large. For example, with $\eta_0 = 0.5$, any pair of observations that share less than half their APs in common will be set to be totally different ($+\infty$).

C. Motion classification

As mentioned earlier, the key observation used to infer about a user's motion state is the level of 'denseness' or 'connectedness' of WiFi observations accumulated within a short time interval acquired incrementally in real-time. A good fit for this task is DBSCAN [10], a density-based clustering algorithm, with the additional advantage of being non-parametric in number of clusters.

DBSCAN develops three concepts that are naturally relevant to our problem: *directly density reachable*, *density reachable* and *density-connected*. It requires a pair of parameters (ϵ, D) (which may be inferred automatically from the data) where ϵ is a radius around point p to form its neighboring set $N(p|\epsilon)$ and D serves as a threshold to test if two points p and q are *directly density reachable*. Two points p and q are then called *density reachable* if there is a sequence of points $p = p_1, p_2 ..., p_l = q$ that links them, where p_i is directly density reachable to p_{i+1}; and finally p is called density connected to q if a point v that is density reachable from both p and q can be found. These concepts are depicted in Figure 2. DBSCAN then seeks to form clusters that are maximal in density-connectedness. Incremental DBSCAN [9] also uses the above fundamental density concepts but operates in an online manner. Assuming that all points prior to the arrival of a new point have been clustered, the essential step is the INSERT operator, which updates the points effected by the newly arrived point. A DELETE operation can also be similarly performed to remove stale points.

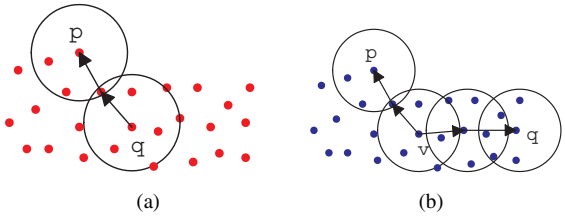


Fig. 2. (a) directly density reachable; (b) density reachable [10].

To determine whether the user is static or moving, we examine the similarity of WiFi observations within a window. If the user is static, a regular 'cluster' will be returned, and if the user is moving, no cluster should be formed because of the variability in WiFi signal strength or visible access points.

Algorithm 1 Motion state detection.

Input: current window of WiFi observations
For each WiFi observation p in current window
  If no cluster found then
    If p is unclassified then
      If number of neighbors $|N(p|\epsilon)| \geq D$ then
        p and its neighboring points form a cluster.
    Else p is noise point
Output: motion state: 'moving' if no cluster found or cluster size is smaller a threshold δ (discussed more in the texts); or 'static' otherwise.

The batch approach is restricted by window size (Alg. 1). E.g., if the window size is 20 seconds, a result can be obtained only after each 20 second window is processed, leading to clear real time limitations. This can be overcome by an overlapping window approach that employs Incremental DBSCAN. As each new overlapping window is introduced, new WiFi observations are added, and outdated WiFi observations are removed. As each WiFi observation p is introduced, there are three possibilities: (1) Noise: p is

Algorithm 2 Incremental INSERT and DELETE operations.

Input: new point p (to INSERT) and existing data points
If number of neighbors $|N(p|\epsilon)| \geq D$ then
  p belongs to the existing cluster or form a new one.
For each neighboring point q of p
  Update neighboring set of q by adding p
  If $|N(q|\epsilon)| \geq D$ then
    q belongs to the existing cluster or form a new one.
Output: updated clustering result and motion state.

Input: existing point p (to DELETE) and other data points
For each neighbor point q of p
  Update neighboring set of q by subtracting p
  If $|N(q|\epsilon)| < D$ then
    Update cluster status
Output: updated clustering result and motion state.

considered noise, (2) Creation: p and some previous noise points form a new cluster, or (3) Absorption: p is absorbed into the existing cluster (Alg. 2). There are two cases when a point p is to be removed: (1) Removal: p's neighbors are decreased and the existing cluster may disappear, or (2) Reduction: p's neighbors are decreased but cluster status is unchanged (see Alg. 2). To take advantage of the existence of at most one cluster, the algorithm terminates as soon as a cluster is found.

Recall that DBSCAN requires two parameters: the neighborhood radius ϵ and number of neighbors D. Ester *et. al.* [10] propose a simple but efficient heuristic to determine ϵ and D in terms of the "thinnest" cluster in the database. Let k-dist be the distance from each point p to k-th nearest neighbor q of p. A sorted k-distance graph is produced by sorting all the points in descending value of k-dist. A good threshold is suggested to be empirically chosen as the point where there is a rapid change in the sorted k-distance graph. In Figure 3 we empirically show a threshold that intuitively gives a good balance on this criteria. If the number of observations assigned to a particular cluster within a window is greater than a threshold δ, the state is deemed to be static. For a window with N observations, D is calculated as $D \approx \delta \times N$.

D. Significant location discovery

The notion of significant or *meaningful* locations has been discussed in the literature from a range of perspectives. For our purposes it is sufficient to define a significant location as a place where a user spends time. This in turn requires that the user be repeatedly *stationary* at a location.

Hence, in our setting, this involves a two step process. Firstly, the user's motion level is classified over a window of time, and a cluster is formed if they are inferred to be still. If so, the observations that belong to the static cluster are extracted. For each AP id in these observations, the average signal strength is computed. This is termed the "average observation" for the given window. In the second

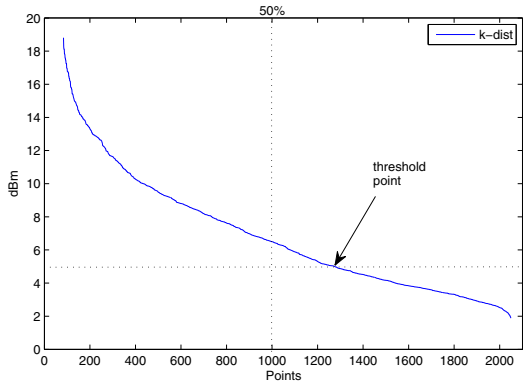


Fig. 3. The sorted k-distance graph , $\delta = 90\%$, window size = $20s$.

step, we use the original version of Incremental DBSCAN presented in [9] to cluster this new average observation. As each WiFi observation p is introduced, there are four possibilities: (1) Noise: p is a noise point, (2) Creation: p and some previous noise points form a new cluster, (3) Absorption: p is absorbed in the existing cluster, or (4) Merge: several clusters and noise neighbors are merged to form a cluster. Parameters are determined using the k-distance criteria referred to above, barring the parameter D, which is directly input from the system.

E. User rhythm detection

While the notion of significant locations is useful for many applications, location alone is an insufficient index for others; Often diverse activities may be folded into the same location, and are only differentiable when the context of behaviour over time is considered. [**?**] have demonstrated, at a coarse scale using logs from 100000 mobile phone users over a 6 month period, that human trajectories show "a high degree of temporal and spatial regularity." [15] note that, in addition to location, other social facets including routine are equally important. Below we detail an approach to detecting periodicities in time and location that arise from daily routines, termed rhythms.

We propose to apply the LDA model [4] to the problem of rhythm extraction. Figure 4 is a graphical representation of LDA in plate-notation, where WiFi observations are mapped to words, these observations collated for a day form a document, all days form a corpus, and the latent topics discovered from this corpus are the sought-after rhythms. In this figure, T is the number of topics, N_d is the number of words contained in document d. For each document d, a mixing topic proportion $\theta \sim \text{Dir}(\alpha)$ is sampled from a Dirichlet distribution parameterised by the hyperparameter α, each word in the document is generated by first sampling a topic z from a multinomial distribution $z \sim \text{Mult}(\theta)$ and then sampling $w \sim \text{Mult}(\phi_z)$ where, again, each ϕ_z is simplex distributed according to $\text{Dir}(\beta)$. LDA thus models each document as a mixture of topics, similar to probabilistic Latent Semantic Indexing (PLSI),

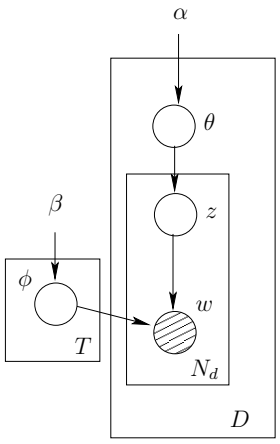


Fig. 4. Graphical representation for latent Dirichlet allocation (LDA) [4].

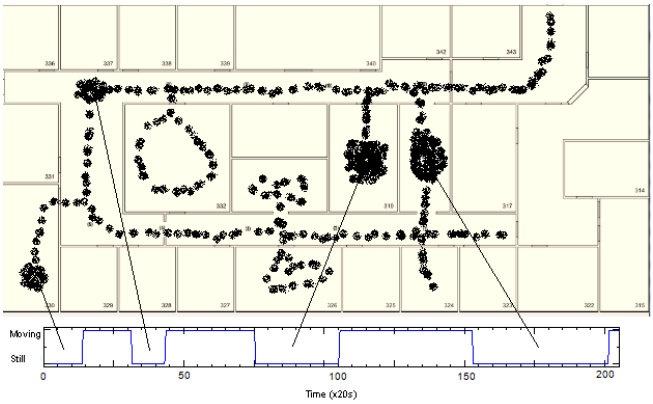


Fig. 5. Example a user movement trace for motion classification task (groundtruth is plotted underneath).

but places a smooth distribution over the topic distribution. In our case, a rhythm is defined as repetitive visitation of significant locations across days. For example, typical office rhythms may include 'having lunch in the staff room around noon' or 'seminar every Monday at 1pm in the boardroom.' These hidden timetables or agendas, which drive the user's trajectory, are the latent topics we seek to model with LDA and interpret as rhythms.

Exact inference in LDA is known to be intractable. Options include the variational approach [4], expectation propagation (EP) [20] or collapsed Gibbs sampling [11]. Despite being deterministic with an analytical bound, the variational method is known to be biased and may wrongly estimate the parameter. EP requires memory storage in the order of number of topics $\times$ total words in the corpus and quickly becomes infeasible with a large corpus. Besides, EP is known to have problems with sparse data. In this work, we use collapsed Gibbs sampling proposed in [11], which iteratively draws samples from the conditional distribution for each topic z_i after marginalizing out the parameters[1]:

$$\Pr\left(z_i^d = z \mid \mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta\right) \propto \left(\alpha_z + n_z^{d,-i}\right) \frac{\beta_w + n_{zw}^{-i}}{\sum_v \left(\beta + n_{z,v}^{-i}\right)}$$

where $\mathbf{z}_{-i}$ denotes the sequence of topic assignments excluding position i (in document d), $\mathbf{w}$ denotes the entire observed sequence of words, $n_z^{d,-i}$ denotes the number of topic z being assigned to document d excluding position i, $n_{z,w}^{-i}$ denotes the number of the current word $w = w_i^d$ being assigned to topic z , and $n_{z,v}^{-i}$ denotes the number of a vocabulary v being assigned to topic z, again excluding position i. The first term is proportional to the number of the current topic z within document d and the second term is proportional to the count of the current word w in document d to the topic z. Intuitively the effect of co-occurrence is achieved by assigning higher probability to two words in the same document being assigned to the same topic.

[1]Please see [12] for full derivation.

The collapsed Gibbs sampling presented here is a form of Markov Chain Monte Carlo method and is guaranteed to converge to the target distribution with sufficient iterations. This method has been found to work very efficiently in practice when applied to different types of corpora. In this work, we employ a version of symmetric Dirichlet for both α and β which implies the entire Dirichlet is controlled by only one parameter. As β gets smaller and $\beta \to 0$, the model favors more discriminative topics. I.e., the samples generated from the Dirichlet distribution are concentrated at the corners of the simplex. When $\beta \to 1$ the Dirichlet behaves more like a uniform distribution, and when β is large, its samples are concentrated in the center of the simplex, favouring topics which are more similar (by Kullback-Leibler divergence).

III. Experiments

A. Data and groundtruth

All WiFi data was collected using a handheld HP iPAQ HW6569 and custom logger written with .NET Compact Framework 2.0 and the free OpenNetCF package[2]. For motion classification task, groundtruth was manually marked down using a GUI interface as the user walked around the designated area. Data was collected over a 60 minute period at a sampling rate of $0.5Hz$ for 5 days. Figure 5 shows an example user trajectory, including static periods when the user was immobile at the same place for a few minutes. Data for the significant locations experiment was drawn using the same apparatus, but over a 28 day period of one user's normal daily routines, during the hours of 8:00AM to 17:30PM. Groundtruth was labelled from among 6 landmarks indicated by the user to be significant, and, as with the motion groundtruth, was indicated using the logging software. For the rhythm experiment, the 28 days of data collected for the significant locations experiment was collated and used.

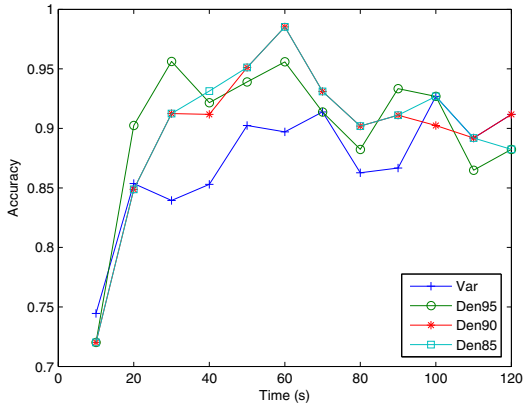
[2]www.opennetcf.com

B. Motion classification results

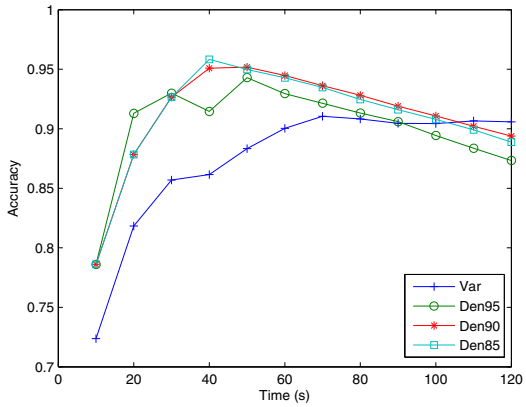
A number of different parameterizations were used to experiment on motion classification, including both overlapping and non-overlapping windows, and window sizes ranging from 10 to 120 seconds. The cluster quality threshold parameter δ was tested with $85\%, 90\%$ and 95%; and $\eta_0 = 0.75$. We also compared our method with the variance-based approach of Krumm et al. [17], which requires supervised training to learn the probabilities $P(\sigma^2 \mid \text{still})$ and $P(\sigma^2 \mid \text{moving})$.

To evaluate the algorithm, we compute the accuracy for each class defined as the ratio of the number of observations detected for that class to the total number of observations in testing.

The accuracy of the variance-based and density-based approaches are shown in Figures 6(a) (non-overlapping windows) and 6(b) (overlapping windows). Performance is poor for short windows ($10s$), probably due to the limnited number of WiFi observations within the window. Accuracy gradually improves as the window is lengthened, levelling out subsequently. In general, the density-based approach shows superior performance compared to the variance-based approach in all cases. Moreover, overlapping windows lead to more consistent performance. It is to be noted that while overlapping and non-overlapping windows yield more or less similar results, overlapping windows have shorter latency and thus can be used in real time.



(a) Non-overlapping windows



(b) Overlapping windows

Fig. 6. The accuracy when window size ranges from $10s$ to $120s$ in non-overlapping and overlapping cases. Legend consists of a variance-based approach (Var) and varying thresholds for the density-based approach.

C. Significant location extraction results

We first perform the motion classification step at the quality threshold of $\delta = 90\%$ to determine if the motion state within a non-overlap $60s$-Window is static or not. Recall that a $60s$-Window shows the best performance in non-overlap situation (Figure 6a). Two parameters in the batch algorithm are automatically derived, $\epsilon = 5$dBm and $D = 60s \times 0.5Hz \times 90\% = 10$. We call this the "average observation" for a *minute interval*.

In the second step, wee clusters this new average observation incrementally as described earlier. While the neighbor distance remains at $\epsilon = 5$dBm, the number of neighbors is configured to reflect a duration of 5 minutes, appropriate to discovery of significant locations, $D = 5$. Note that the meaning of D in each step should be distinguished because of the different time intervals integrated in their observations.

For performance evaluation, we use the *cluster purity* measured as the percentage of match between the extracted clusters $\{c_1, ..., c_L\}$ and the groundtruth clusters $\{\psi_1, ..., \psi_K\}$. The purity of cluster i is defined as

$$P_i = \frac{\max_j |c_j \cap \psi_i|}{N_i}$$

where N_i is the number of points in the groundtruth of cluster i. The average purity across K clusters is computed as:

$$\bar{P} = \sum_{i=1}^{K} \frac{N_i P_i}{N}$$

where N is total number of points $N = \sum_{i=1}^{K} N_i$.

The clusters are compared with the groundtruth and the purity of the clusters are shown in Table I. Overall cluster purity is 98.77%.

To test the model's predictive power, one third of the collected data not used for the above clustering is used for testing. Again, incremental DBSCAN ($\epsilon = 5$dBm and $D = 5$) is used to assign clusters, and thus significant locations. Results are shown in Table II. It can be seen that when sufficient data is used for detecting significant locations (Landmarks 1–5), accuracy is very high.

D. Rhythm detection results

Significant locations are discovered using the previous technique, and are mapped to words by discretizing each day into 30-minute intervals from 8am to 5:30pm to obtain a sequence of data consisting of pairs of (time, significant

TABLE I

PERFORMANCE OF CLUSTERING SIGNIFICANT LOCATIONS.

Landmark	#Groundtruth	#Clustered	Purity (%)
1	2574	2572	99.92
2	296	291	98.31
3	40	39	97.50
4	41	41	100.00
5	193	187	96.89
6	19	19	100.00

TABLE II

PREDICTION OF SIGNIFICANT LOCATIONS.

Landmark	#Prediction	#Correct	Accuracy (%)
1	857	855	99.76
2	100	94	94.00
3	13	13	100.00
4	14	13	92.85
5	35	31	88.57
6	7	5	71.42

place label) tuples. Gibbs sampling is used for inference where symmetric Dirichlet hyper-parameters are set to $\alpha = 0.01$ and $\beta = 0.01$. The number of Gibbs iterations is 2000 in which the first 200 iterations are discarded (burn-in stage) and samples are collected in every 10 iterations (lag) [11]. As a measure of the goodness of fit of the model, perplexity is used to determine the suitable number of topics, and is computed as:

$$\text{Perp}\,(\mathbf{w}) = \exp\left\{-\frac{\log \Pr(\mathbf{w})}{N}\right\}$$

where N is number of words. Adapting [29] the probability of the corpus $\Pr(\mathbf{w})$ is computed as follows. After each sampling step, let $n_{d,k}$ be the number of times that topic k appears in document d, $n_{k,v}$ be the number of times a word v is assigned to topic k, parameters θ and ϕ are first estimated:

$$\hat{\theta^s}_{d,k} = \frac{n_{d,k} + \alpha}{\sum_k n_{d,k} + T\alpha} \quad ; \quad \hat{\phi^s}_{k,v} = \frac{n_{k,v} + \beta}{\sum_v n_{k,v} + W\beta}$$

and then $\Pr(\mathbf{w})$ is computed as:

$$\Pr(\mathbf{w}) = \prod_{d=1}^{D}\prod_{v=1}^{W}\sum_{k=1}^{T}\frac{1}{S}\sum_{s=1}^{S}\hat{\theta^s}_{d,k}\hat{\phi^s}_{k,v}$$

where S is the number of collected samples.

Figure 7 shows the perplexity in log form when the number of topics is changing. The lower the perplexity of the model, the better it fits. In addition, the greater the number of topics, the more computation Gibbs sampling requires. It is therefore desirable to choose the 'simpler' model with good degree of fitness. In our case (cf. Figure 7) we choose the number of topics $T = 5$. For this choice the perplexity has decreased exponentially and is stabilizing. It thus offers both goodness of fit and model simplicity.

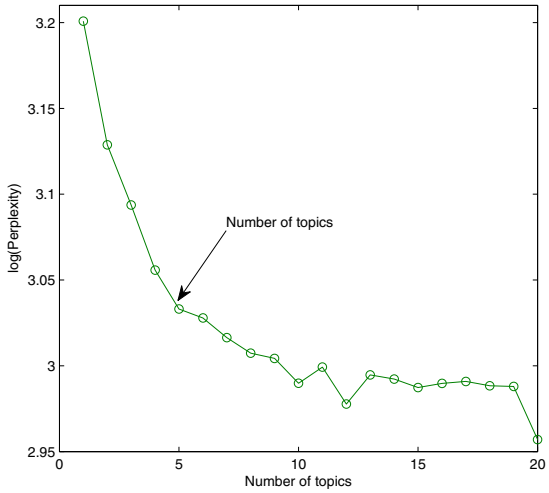


Fig. 7. Choosing the number of topics using perplexity.

Five significant rhythms (R1-R5) are shown in Figure 8 in which each row corresponds to a specific landmark and the height of each bar depicts the accumulated time spent. Figure 9 provides a different perspective on the same detected rhythms. In general, the simple rhythms illustrate the repetitive nature of a research student. For example, most of the time was spent in his office or at lunch (Rhythm 1). This person participates in the institute's academic activities, such as meeting with advisors (Rhythm 3) and attending seminars (Rhythm 5). In addition, his frequent visits to two advisors' rooms for discussions are also extracted (Rhythm 2 and Rhythm 4). Rhythm 1 reflects the daily routine of this research student, while Rhythms 2-5 are weekly routines.

IV. RELATED WORK

Using GPS signals, context such as location-based activity and significant places are more or less solved in outdoor environments [7], [34], [1], [19].

We focus primarily on work that recovers context with an indoor component in terms of location, proximity of others, and/or some definition of activity via a range of sensors. Table III provides a brief breakdown of related work by signal type, method and context discovered.

The work of [6] develops a wearable system including video and audio sensors for extracting the events and scenes. HMMs are used to infer the events, however recent recognition techniques in video and audio are still unreliable. However, using HMMs required supervised training to learn the parameters and thus user-dependent. Aiming for an unsupervised approach we deliberately avoid this class of models. [8] uses GSM positioning and Bluetooth proximity to extract the repeated activities of individuals and community patterns by extracting the principal eigenbehaviors from the eigenvalues of the day (row) vs time

TABLE III
REVIEW

Environment	Signal	Context	Work	Method
Indoor/outdoor	Video	Event/Scene	[6]	short/long time-scale HMM
	Audio			
	GSM	Eigenbehavior	[8]	SVD
	Bluetooth	Proximity		
	Accelerometer	Activity	[14]	Naive Bayes
		Routine		LDA
Indoor	WIFI	Location	[3]	KNN
	WIFI	Location	[18], [26], [33], [32]	Naive Bayes
	WIFI	State	[17]	Variance of signal strength
		Location		Naive Bayes
	RFID	Activity	[24]	Dynamic Bayes model
	Powerline	Location	[23]	Naive Bayes

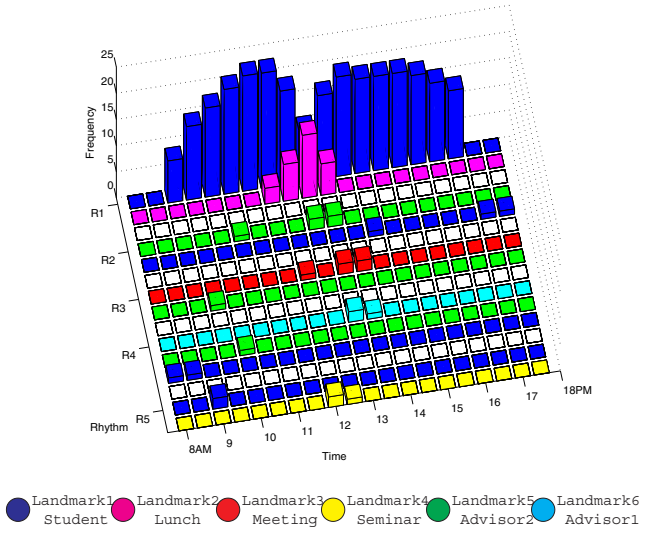


Fig. 8. Detected user rhythms – histograms by landmark (rhythms are delineated by white boxes; best viewed in colour).

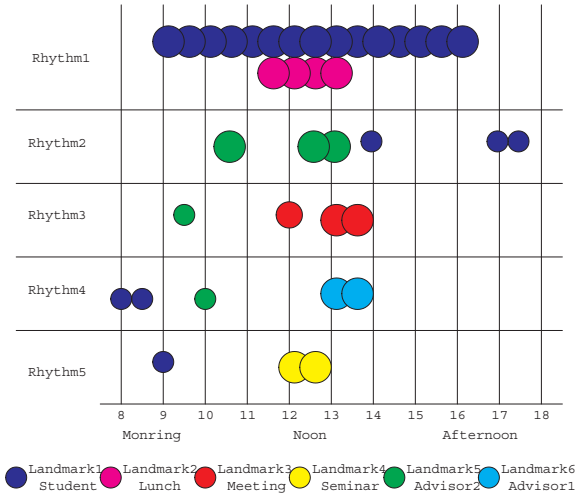


Fig. 9. Detected user rhythms – mixtures of landmarks (best viewed in colour).

(column) matrix, in which an entry indicates if the user was at that spatial temporal location. However, GSM is too coarse for use in indoor activities. Further they do not extract rhythms.

By integrating the RFID tag for each equipment with meaningful description [24], the context is extracted directly from the description and then higher-order activities are inferred using the same model in [7].

In [27], a Nash H-learning mechanism is used to predict user mobilities for efficient resource management. To complement location-based activity recognition, [14] propose a new approach to extract activity patterns using accelerometers. Naive Bayes classifier is used for recognizing the normal activities such as sitting, walking, driving etc.

The work of [25] proposes a novel model LSDA, an extended version of N-gram model [30], to extract the socially hidden rhythms of a user using GPS traces. The corpus of words and documents is generated from GPS data [1] in which word is mapped to <time,significant place> and a document consists of all words in each day. The advantage of LSDA is its ability to map consecutive locations as a N-gram of words and thus in its ability to extract meaningful social themes.

Applying the original LDA model with variational inference [4], [14] constructs a dictionary of word and a corpus of documents in which each word is an activity and a document consists of all words during particular day and leverage the daily patterns from normal activities collected in 16-day experiment using accelerometer.

Understanding context and providing context-aware application services are critical to dynamic pervasive environments. Context continues to be a topic of research focus as context data and their associated sources exhibit dynamism. Henrickson and Induslka [13] discuss shortfalls of context modeling and reasoning with ontologies for understanding context. Nicklas et al. [21] investigate the use of hybrid reasoning to augment the NEXUS framework. Naive Bayesian classifiers are used in [16] to derive high level contexts. To summarize, dynamic Bayesian models have been primarily

used on either WiFi or accelerometer data to derive context in indoor environments. These models are supervised. Limited use has been made to extract context with topic models on accelerometer and GPS data, but none of these have been applied to WiFi data. Thus there is a requirement to produce techniques that work in an unsupervised manner on noisy WiFi data to extract context.

V. Conclusion

Motivated by the need to build assistive systems in indoor environments, we have presented a novel, clustering-based method for the extraction of user context from ambient WiFi. This includes state of motion, significant locations, and rhythms. Experiments validate the accuracy of our techniques. The advantages of our approach lie in its use of existing wireless infrastructure, without requirements for calibration, and its ability to support real-time services. Our future work includes investigating the applicability of this approach in a much larger setting and seeking better modeling of temporal information for the task of rhythms extraction such as n-gram models.

References

[1] B. Adams, D. Phung, and S. Venkatesh. Extraction of social context and application to personal multimedia exploration. *Proceedings of The 14th Annual ACM International Conference on Multimedia*, pages 987–996, 2006.

[2] B. Adams, D. Phung, and S Venkatesh. Sensing and Using Social Context. *ACM Transaction on Multimedia Computing, Communications and Applications (TOMCAP)*, 2008.

[3] P. Bahl and VN Padmanabhan. RADAR: an in-building RF-based user location and tracking system. *Proceedings of The 19th Annual Joint Conference of The IEEE Computer and Communications Societies (INFOCOM)*, 2, 2000.

[4] David M. Blei, Andrew Y. Ng., and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] J. Carrasco and E. Miller. Exploring the propensity to perform social activities: a social network approach. *Transportation*, 33(5):463–480, September 2006. available at http://ideas.repec.org/a/kap/transp/v33y2006i5p463-480.html.

[6] B.P. Clarkson. *Life patterns: structure from wearable sensors*. PhD thesis, Massachusetts Institute of Technology, 2003.

[7] J. Donald, L. Lin, and F. Dieter. Inferring high-level behavior from low-level sensors. *Proceedings of The 5th International Conference of Ubiquitous Computing. New York, USA*, 139, 2003.

[8] N.N. Eagle. *Machine Perception and Learning of Complex Social Systems*. PhD thesis, Massachusetts Institute of Technology, 2005.

[9] M. Ester, Hans P. Kriegel, Jorg Sander, Michael Wimmer, and Xiaowei Xu. Incremental Clustering for Mining in a Data Warehousing Environment. *Proceedings of the 24rd International Conference on Very Large Data Bases*, pages 323–333, 1998.

[10] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press*, pages 226–231, 1996.

[11] T.L. Griffiths. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_ 1):5228–5235, 2004.

[12] T.L. Griffiths, M. Steyvers, and J.B. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211–44, 2007.

[13] K. Henricksen, J. Indulska, and A. Rakotonirainy. Modeling Context Information in Pervasive Computing Systems. *Proceedings of the First International Conference on Pervasive Computing (Pervasive)*, pages 167–180, 2002.

[14] T. Huynh, M. Fritz, and B. Schiele. Discovery of Activity Patterns using Topic Models. *Proceedings of the Tenth International Conference on Ubiquitous Computing*, 2008.

[15] Q. Jones, S.A. Grandhi, S. Whittaker, K. Chivakula, and L. Terveen. Putting systems into place: a qualitative study of design requirements for location-aware community systems. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 202–211, New York, NY, USA, 2004. ACM Press.

[16] P. Korpipaa, J. Mantyjarvi, J. Kela, H. Keranen, and E.J. Malm. Managing Context Information in Mobile Devices. *IEEE Pervasive Computing*, pages 42–51, 2003.

[17] J. Krumm and E. Horvitz. Locadio: Inferring Motion and Location from Wi-Fi Signal Strengths. *Proceedings of International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous)*, 2004.

[18] A.M. Ladd, K.E. Bekris, A. Rudys, L.E. Kavraki, and D.S. Wallach. Robotics-Based Location Sensing Using Wireless Ethernet. *Proceesings of The 8th ACM International Conference on Mobile Computing and Networking (MOBICOM)*, 2002.

[19] L. Liao, D. Fox, and H. Kautz. Extracting Places and activities from GPS Traces Using Hierarchical Conditional Random Fields. *The International Journal of Robotis Research*, 26(1):119, 2007.

[20] T.P. Minka. Expectation propagation for approximate Bayesian inference. *Uncertainty in Artificial Intelligence*, 17:362–369, 2001.

[21] D. Nicklas, M. Grossmann, J. Mínguez, and M. Wieland. Adding High-level Reasoning to Efficient Low-level Context Management: A Hybrid Approach. *Proceedings of the 2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications-Volume 00*, pages 447–452, 2008.

[22] P. Nurmi and J. Koolwaaij. Identifying meaningful locations. *Mobile and Ubiquitous Systems: Networking & Services, 2006 Third Annual International Conference on*, pages 1–8, 2006.

[23] S.N. Patel, K.N. Truong, and G.D. Abowd. PowerLine Positioning: A Practical Sub-Room-Level Indoor Location System for Domestic Use. *the proc. of Ubicomp*, pages 441–458, 2006.

[24] M. Philipose, K.P. Fishkin, M. Perkowitz, D.J. Patterson, D. Fox, H. Kautz, and D. Hähnel. Inferring Activities from Interactions with Objects. *IEEE PERVASIVE COMPUTING*, pages 50–57, 2004.

[25] Dinh Phung, Brett Adams, and Svetha Venkatesh. Computable Social Patterns from Sparse Sensor Data. *First Int. Workshop on Location Web, World Wide Web Conference (LocWeb08,WWW08)*, 2008.

[26] T. Roos, P. Myllymäki, H. Tirri, P. Misikangas, and J. Sievänen. A Probabilistic Approach to WLAN User Location Estimation. *International Journal of Wireless Information Networks*, 9(3):155–164, 2002.

[27] N. Roy, A. Roy, and S.K. Das. Context-aware resource management in multi-inhabitant smart homes: A nash h-learning based approach. In *Proceedings of the 4th Annual IEEE International Conference on Pervasive Computing and Communications (PERCOM)*, 2006.

[28] V. Seshadri, G.V. Zaruba, and M. Huber. A Bayesian Sampling Approach to In-door Localization of Wireless Devices Using Received Signal Strength Indication. *Proceedings of the 3rd IEEE international Conference on Pervasive Computing and Communications*, pages 75–84, 2005.

[29] Y.W. Teh, D. Newman, and M. Welling. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, 19:1353, 2007.

[30] X. Wang, A. McCallum, and X. Wei. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 697–702, 2007.

[31] M. Weiser. The computer for the 21 stcentury. *ACM SIGMOBILE Mobile Computing and Communications Review*, 3(3):3–11, 1999.

[32] Z. Xiang, S. Song, J. Chen, H. Wang, J. Huang, and X. Gao. A Wireless LAN-based Indoor Positioning Technology. *IBM Journal of Research and Development*, 48(5/6):617–626, 2004.

[33] MA Youssef, A. Agrawala, and A. Udaya Shankar. WLAN location determination via clustering and probability distributions. *Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications (PerCom 2003)*, pages 143–150, 2003.

[34] C. Zhou, S. Shekhar, and L. Terveen. Discovering Personal Paths from Sparse GPS Traces. *1st International Workshop on Data Mining in conjunction with 8th Joint Conference on Information Sciences July*, 2005.