NextPlace: A Spatio-Temporal Prediction Framework for Pervasive Systems

Salvatore Scellato[1] Mirco Musolesi[2] Cecilia Mascolo[1]
Vito Latora[3] Andrew T. Campbell[4]

[1] Computer Laboratory, University of Cambridge, UK
[2] School of Computer Science, University of St. Andrews, UK
[3] Dipartimento di Fisica, University of Catania, Italy
[4] Department of Computer Science, Dartmouth College, USA

Abstract. Accurate and fine-grained prediction of future user location and geographical profile has interesting and promising applications including targeted content service, advertisement dissemination for mobile users, and recreational social networking tools for smart-phones. Existing techniques based on linear and probabilistic models are not able to provide accurate prediction of the location patterns from a spatio-temporal perspective, especially for long-term estimation. More specifically, they are able to only forecast the next location of a user, but not his/her *arrival time* and *residence time*, i.e., the interval of time spent in that location. Moreover, these techniques are often based on prediction models that are not able to extend predictions further in the future.

In this paper we present NextPlace, a novel approach to location prediction based on nonlinear time series analysis of the arrival and residence times of users in relevant places. NextPlace focuses on the predictability of single users when they visit their most important places, rather than on the transitions between different locations. We report about our evaluation using four different datasets and we compare our forecasting results to those obtained by means of the prediction techniques proposed in the literature. We show how we achieve higher performance compared to other predictors and also more stability over time, with an overall prediction precision of up to 90% and a performance increment of at least 50% with respect to the state of the art.

1 Introduction

The ability to predict future locations of people allows for a rich set of novel pervasive applications and systems: accurate content dissemination of location related information such as advertisement, leisure events reports and notifications [1, 20] could be implemented in a more effective way, avoiding the delivery of information to uninterested users, and, therefore providing, a better user experience. For example, by exploiting the availability of future location information, Web search engines such as Google, Bing or Yahoo! and location-based social network services such as Facebook Places and Foursquare may provide "location-aware" sponsored advertisements together with search results that are relevant to the predicted user movement patterns.

The increasing popularity of smart-phones equipped with GPS sensors makes location-aware computing a reality. Even in the case of devices where this information is not currently available, location can be roughly estimated by means of triangulation and cell estimation techniques or by profiling places through the analysis of the MAC addresses advertised by nearby devices and 802.11 access points [17]. In addition, these devices are increasingly always connected to the Internet, at least in areas where GPRS/EDGE or WiFi connectivity is present. Therefore, information about the current positions of users can be transmitted to a back-end server, where analysis of the data can be performed at run-time in order to predict future location patterns.

In this paper we propose NextPlace, a new prediction framework based on *nonlinear* time series analysis [12] for forecasting user behavior in different locations from a *spatio-temporal* point of view. NextPlace focuses on the temporal predictability of users presence when they visit their most important places. We do not focus on the transitions between different locations: instead, we focus on the estimation of the duration of a visit to a certain location and of the intervals between two subsequent visits. The existing techniques are able to forecast the next location of a user, but *not* his/her *arrival* and *residence time*, i.e., the interval of time spent in that location. Moreover, these techniques are often based on prediction models that are not able to extend predictions further in the future, since they mainly focus on the next movement of a user [2, 14, 16, 19, 23, 26].

We focus instead on patterns of residence in the set of locations that are more frequently visited by users. We show that, at least in the datasets under analysis, human presence in important places is characterized by a behavior that, even if at first glance seems apparently random, can be effectively captured by nonlinear models. Predictions are based on the collection of movement data that can be of different types: sets of GPS coordinates, registration patterns to access points or also information about presence in locations by means of passive and active transponders (such as badges). In addition, check-ins performed in location-based social networking services can be exploited to acquire movement data.

The proposed prediction technique consists of two steps. Firstly, we need to identify significant locations among which users move more frequently. Secondly, we apply a model able to predict user presence within these locations and relative residence time by means of techniques drawn from nonlinear time series analysis [12]. More specifically, the contribution of this paper can be summarized as follows:

- We describe NextPlace, a novel approach to user location prediction based on nonlinear time series analysis of visits that users pay to their most significant locations. NextPlace estimates the time of the future visits and expected residence time in those locations.
- We analyze four datasets of human movements: two GPS-based (representing respectively the positions of the users involved in the deployment of the CenceMe application at Dartmouth College [21] and the locations of cabs in San Francisco [24]) and two containing registration patterns of WiFi access points (at Dartmouth College [15] and within the Ile Sans Fils wireless network in Montreal, Canada [18]). We identify regularity and, more specifically, some previously uncaptured degree of determinism in patterns of user visits to their significant places by means of nonlinear analysis.

– We evaluate NextPlace comparing it with a probabilistic technique based on spatio-temporal Markov predictors [26] and with a linear model [6]. We report an overall prediction precision over the four datasets of up to 90%, with precision of up to 65% even after a number of hours, and a performance increment of at least 50% over Markov-based predictors. We show how the adoption of a nonlinear prediction framework can improve forecasting precision with respect to other techniques even for long-term predictions.

The rest of this paper is organized as follows: Section 2 describes NextPlace and its novel approach to prediction based on nonlinear time series analysis as well as illustrates the techniques we use for the extraction of significant places. Section 3 presents the implementation issues and the validation of our approach using real-world measurements, also reporting the results of the evaluation of our method against other predictors. Section 4 discusses related work and Section 5 concludes the paper illustrating potential future work.

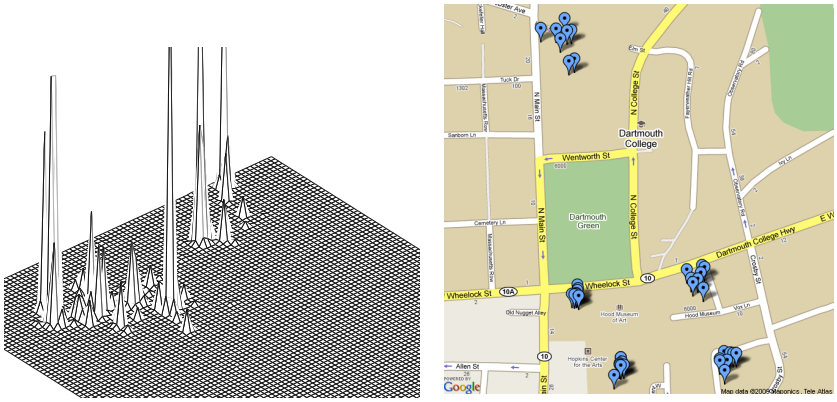
2 Predicting Spatio-temporal Properties of Mobile Users

Any prediction of future user behavior is based on the assumption of determinism. From a practical point of view, determinism simply means that future events are determined by past events, so that every time a particular configuration or situation is observed, the same (or a similar) outcome will follow. Since in human societies daily and weekly routines are well-established, human activities are characterized by a certain degree of regularity and predictability [8].

The intuition behind NextPlace is that the sequence of important locations that an individual visits every day is more or less fixed, with only minor variations that are also usually deterministically defined. As an example, if a woman periodically goes to the gym on Mondays and Thursdays, she may change her routine for those days, but the changed routine will be more or less the same over different weeks. Therefore, the sequence of events may still be predictable.

From a formal point of view, let us consider a certain number of mobile users, where user i freely moves among different locations. For the moment, we do not explicitly focus on how these locations can be identified, and only assume that the start time and the duration of each visit of a user to a given location can be determined. A visit of a user is simply defined by the tuple (u, l, t, d), where t and d are respectively the time of arrival and the residence time of user u in location l. It is worth noting that this approach does not model movements but, rather, residence time in some locations, hence, it can also be adopted in systems without any spatial or geographical information about locations, i.e., access points in 802.11 WLANs.

We now introduce the two steps of NextPlace and the basic theory behind them. We first describe how we isolate the user's significant places, exploiting the technique proposed by Kim et al. in [14]. Then, we describe our novel method for the estimation of future times of arrival and residence times in the different significant places and how we exploit this prediction to compute accurate estimation of where the user will be after a given time interval. Finally, we describe the mathematical details of the prediction techniques behind our approach.



(a) (b)

Fig. 1: Example of frequency map using GPS traces. Higher peaks in (a) reveal places where user spent most of their time and which represent its significant locations: in (b) we show some visits to these significant places reported on a geographical map.

2.1 Significant Places Extraction

In this section we present two methods we use to extract significant locations from both GPS information and WiFi association logs, the two most commonly available sources of data about user movements.

Extracting Places from GPS Data Many solutions for the extraction of significant places from GPS measurements have been presented in the literature [2, 11, 28]. We choose one that is based on the residence time of a user to quantify the importance of a place for him/her: the intuition being that permanence at a place is directly proportional to the importance that is attributed to it by the user.

As proposed in [14], we apply a 2-D Gaussian distribution weighted by the residence time at each GPS point. This means that at each point the Gaussian distribution uniformly contributes also to nearby points, smoothing out values that are close together. The value of the variance for the Gaussian distributions that we choose is $\sigma = 10$ meters, which is related to the average GPS accuracy[5]. The resulting *frequency map* contains peaks which give information about the position of popular locations: we consider regions that are above a certain threshold T as significant places. The threshold T can be chosen as a fraction of the maximum value of the frequency map. We will show the application of this technique and how the value of the threshold T can be selected using two GPS-based datasets in Section 3.

In Figure 1(a) a close-up of a frequency map is shown: when a threshold is applied to the map, only higher peaks are selected and each peak generates an area defined

[5] `http://www.gps.gov/`

by a continuous boundary. All GPS points within that area result in visits to the same significant place. As an example, if we choose a threshold equal to 15% of the highest peak of the map, we obtain the visits to significant places shown over the area map in Figure 1(b).

Extracting Places from WiFi Logs Alternatively, we can derive significant places from user registrations to 802.11 access points. Since these access points are fixed and easily identifiable from their globally unique MAC address, this information can be exploited to extract visit patterns to a set of locations in a straightforward manner. From this point of view, the most frequently seen access points are natural candidates to represent significant places. Hence, we can define as popular places for a user the access points he/she connects to more often, providing that a sufficient number of visits has been recorded to a given access point. More specifically, we define an access point as a significant place for a certain user if this user has a sequence of at least n visits to the access point, in order to filter out all the access points that are seldom visited and to have a sufficient number of observations from a statistical point of view. For the analysis presented in this paper, we select n equal to 20.

2.2 Predicting User Behavior

We now describe NextPlace's location prediction algorithm: in order to obtain an estimation of the future behavior, the history of visits of a user to each of its significant locations is considered. Then, for each location we try to predict when the next visits will take place and for how long they will last. After this estimation, the predictions obtained for different locations is analyzed, in order to produce a unique prediction of where the user will be at a given future instant of time. A theoretical foundation of this technique is described in Section 2.3.

For each user we keep track of all previous visits to a set of locations, that is, for each visit we consider the instant when it started and how long it lasted. The algorithm predicts the next visits to a given location by means of the previous history of visits $((t_1, d_1), (t_2, d_2), \ldots, (t_n, d_n))$:

1. two time series are created from the sequence of previous visits: the time series of the visit daily start times C and the time series of the visit durations D defined as follows:
$$C = (c_1, c_2, \ldots, c_n)$$
$$D = (d_1, d_2, \ldots, d_n)$$
where c_i is the time of the day in seconds corresponding to the time instant t_i (i.e., c_i is in the range $[0, 86400]$);
2. we search in the time series C sequences of m consecutive values $(c_{i-m+1}, \ldots, c_i)$ that are closely similar to the last m values $(c_{n-m+1}, \ldots, c_n)$[6];
3. the next value of time series C is estimated by averaging all the values c_{i+1} that follow each found sequence;

[6] We will discuss the choice of parameter m in the next section.

4. at the same time, in time series D the corresponding sequences $(d_{i-m+1}, \ldots, d_i)$ are selected; the sequences have to be located exactly at the same indexes as those in C;
5. the next value of time series D is then estimated by averaging all the values d_{i+1} that follow these sequences.

As an example, if the last three visits of a certain user to a location are Monday at 6:30pm, Monday at 10:00pm and Tuesday at 8:15am, we analyze the history of visits in order to find sequences that are numerically close to (6:30pm, 10:00pm, 8:15am), i.e. (6:10pm, 9:50pm, 8:35am) and (6:35pm, 10:10pm, 8:00am): then, assuming that the next visits that follow these subsequences start at 1:10pm and 12:40pm and last for 40 and 30 minutes respectively, we estimate the next visit at 12:55pm for 35 minutes, averaging both arrival times and duration times.

The main idea behind this algorithm is the assumption that human behavior is strongly determined by daily patterns: the sequence of visit start times is therefore mapped to a 24-hour time interval, focusing only on the start time of each visit. The choice of the value m has an impact on the accuracy of the prediction: in fact, this can be improved by taking into account more visits in order to identify particular patterns that may be present only in certain intervals of time such as specific days.

We can generalize this algorithm to predict not only the next visit to a location, but also successive visits in the future: in fact, we can choose to average together not only the next values of each subsequences but also values that are 2 or more steps ahead. However, the prediction of time series can become inaccurate when adopted to calculate further values in the future [12].

Since we can predict when the future visits to all significant locations will start and for how long they will last, we can design a simple method to predict the location where the user will be at a given time in the future. Let us suppose that at time T we want to predict in which significant location user i will be after ΔT seconds. Then, the following steps are performed:

1. for each location the sequence of the next k visits (starting with $k = 1$) are predicted and a global sequence of all predicted visits $(loc_1, t_1, d_1), \ldots, (loc_n, t_n, d_n)$ is created, with $t_1 \leq \cdots \leq t_n$;
2. if there is a prediction (loc_i, t_i, d_i) which satisfies $t_i \leq T + \Delta T \leq t_i + d_i$, then loc_i is returned as predicted location (in case several predictions exist which satisfy the predicate, we choose at random between them);
3. if no prediction satisfies the condition stated above, there are two cases: if the minimum start time t_1 of the current predicted visits is smaller than $T + \Delta T$, then prediction needs to be extended further in the future in order to find a suitable visit, thus the parameter k is doubled and the algorithm is repeated considering new predicted visits. Otherwise, extending the prediction provides visits which start after $T + \Delta T$ and which cannot be exploited for prediction: thus, the algorithm terminates returning that the user will not be in any significant location.

Note that it is realistic for a user to be predicted as being outside the set of significant places (e.g., maybe transitioning from one to another) and that our technique is also able to predict this state.

2.3 Nonlinear Prediction Framework: Key Concepts and Practical Implementation Issues

In this section we provide a brief overview of the key concepts at the basis of the forecasting framework and we discuss the practical issues in implementing it.

In this work we adopt a prediction technique inspired by *nonlinear time series analysis* [12]. A time series can be seen as a collection of scalar observations of a given system made sequentially in time and spaced at uniform time intervals, albeit this last assumption can be relaxed to allow any kind of temporal measurement pattern [6].

While the scalar sequence of values contained in a time series may appear completely unrelated to the underlying system, it is possible to uncover the characteristics of its dynamic evolution by analyzing sub-sequences of the time series itself. In order to investigate the structure of the original system, the time series values must be transformed in a sequence of vectors with a technique called *delay embedding*.

More formally, a time series $(s_0, s_1, \ldots, s_N)$ can be embedded in a m-dimensional space by defining an appropriate delay ν and then creating a *delay vector reconstruction* for the time series value s_n as follows:

$$\beta_n = [s_{n-(m-1)\nu}, s_{n-(m-2)\nu}, \ldots, s_{n-\nu}, s_n]$$

where all vectors β_n have m components and are defined in a so called *embedding space*. Note that m is the parameter used in the algorithm described in Section 2.2.

The values of the parameters m and ν greatly affect the accuracy of the representation. Nonetheless, a fundamental mathematical result (the so-called *embedding theorem* [12]) ensures that a suitable value for m does exist and is related to the complexity of the underlying system. At the same time, ν might be chosen to represent a suitable time scale of the phenomenon, since consecutive values in the time series should not be too strongly correlated to each other.

An effective predictive model can be generated directly from time series data through the delay embedding. Let us suppose that a prediction for the value $s_{N+\Delta n}$, a time Δn ahead of N, must be made for the time series $(s_0, s_1, \ldots, s_N)$. The steps of the prediction process are as follows:

1. The time series is embedded in a m-dimensional space by defining an appropriate time delay ν and then creating the related embedding space;
2. The embedding space is searched for all the vectors that are close, with respect to some given metric distance, to vector β_N: more formally, a neighborhood $U_\epsilon(\beta_N)$ of radius ϵ around the vector β_N is created;
3. Since determinism involves that future events are set causally by past events, and since all vectors $\beta_n \in U_\epsilon(\beta_N)$ describe past events similar to the past events of β_N, the prediction $p_{N+\Delta n}$ is taken as the average of all the values $s_{n+\Delta n}$

$$p_{N+\Delta n} = \frac{1}{|U_\epsilon(\beta_N)|} \sum_{\beta_n \in U_\epsilon(\beta_N)} s_{n+\Delta n}$$

where $|U_\epsilon(\beta_N)|$ denotes the number of elements of the neighborhood $U_\epsilon(\beta_N)$. The value of ϵ should be chosen in order to obtain a sufficient number of vectors for the prediction.

Intuitively, this algorithm searches the past history to find sequences of values that are very similar to the recent history: assuming that the evolution is ruled by deterministic patterns, a given state will always be followed by the same outcome.

In our implementation we have chosen $\nu = 1$, since we do not have to deal with particular time scales which require to skip some values of our time series. As suggested in [12], the radius ϵ of the vector neighborhood is chosen in order to be 10% of the standard deviation of each time series: this value allows us to obtain enough vectors to perform prediction and, at the same time, filters out vectors that are not close to β_N.

We note that for each prediction all vectors in the embedding space have to be considered and searched. For this reason, it is wise to use an efficient method to find nearest neighbors in the embedding space: the main computational burden is the calculation of the neighborhood $U_\epsilon(\beta_N)$ and the asymptotic complexity $O(N^2)$ can be reduced to $O(N \log N)$ with binary trees or even to $O(N)$ with a box-assisted search algorithm [25], which is the method we implement.

3 Validation of the Prediction Framework using Real-world Measurements

In this section we introduce the datasets used in our analysis and we describe how we process them in order to extract significant places. Then, we investigate the predictability of the time series extracted from sequences of visits of each user to his/her significant locations, using standard metrics adopted in time series analysis. Finally, we compare NextPlace prediction performance against other prediction methods.

3.1 Datasets

For the evaluation of our approach we choose four different datasets of human movements:

1. **Cabspotting** This dataset is composed of movement traces of taxi cabs in San Francisco, USA, with GPS coordinates of approximately 500 taxis collected over 30 days in the San Francisco Bay Area. Each vehicle is equipped with a GPS tracking device that is used by dispatchers to efficiently reach customers [24]. The average time interval between two consecutive GPS measurements is less than 60 seconds.
2. **CenceMe GPS** This dataset was collected during the deployment of CenceMe [21], a system for recreational personal sensing, at Dartmouth College. The GPS data was collected by means of 20 Nokia N95 phones carried by postgraduate students and staff members from the Department of Computer Science and the Department of Biology.
3. **Dartmouth WiFi** This dataset was extracted from the SNMP logs of the WiFi LAN of the Dartmouth College campus. The compact nature of the campus means that the signal range of interior APs extends to most of the campus outdoor areas. Between 2001 and 2004 data about traffic in the access points was collected through three techniques: syslog events, SNMP polls, and network sniffers [9, 15].

Dataset	N	V	P	p	v	D [s]	Trace length	Significant time
Cabspotting	252	150612	6122	24.29	597	231	23 days	7.27%
CenceMe GPS	19	3832	225	11.84	201	696	12 days	14.74%
Dartmouth WiFi	2043	772217	539	17.87	377	2094	60 days	11.24%
Ile Sans Fils	804	142407	173	3.61	177	5296	370 days	0.18%

Table 1: Properties of the different datasets: total number of users N, total number of visits V, total number of significant places P, average number of significant places per user p, average number of visits per user v, average residence time in a place D (seconds), total trace length and average proportion of time spent by each user in significant places.

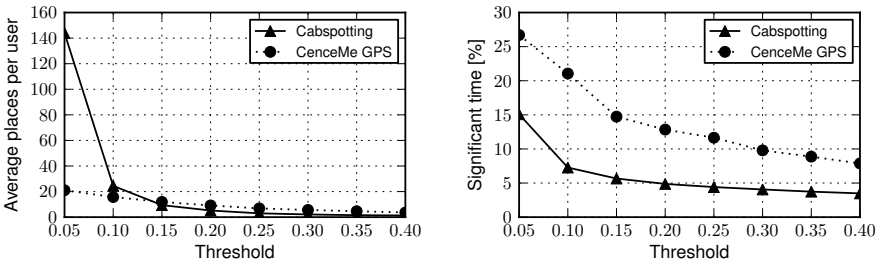
4. **Ile Sans Fils** Ile Sans Fils [18] is a non-profit organization which operates a network of free WiFi hotspots in Montreal, Canada. It now counts over 45,000 users with 140 hotspots located in publicly accessible spaces. These hotspots are deployed mostly in cafes, restaurants and bars, libraries, but also outdoor to cover parks and sections of popular commercial streets.

We choose a subset of regularly active users for each original dataset, filtering out all the users that appear only a few times and for which prediction may be worthless. In Table 1 we report some important characteristics and metrics of the resulting datasets.

3.2 Practical Issues

In order to extract significant places for each user in the Cabspotting and CenceMe GPS datasets, which are composed of GPS measurements, we need to choose a suitable threshold T for the frequency map. Thus, we investigate how the average number of significant places per user changes as a function of the threshold itself. As reported in Figure 2(a), the average number of places decreases as the threshold increases: for the Cabspotting dataset a suitable choice is $T = 0.10$, where the curve changes its slope, which denotes the transition from a situation with many unimportant significant areas to a situation with less but probably more important places. However, in the case of the CenceMe GPS dataset such transition does not occur: hence, we investigate how the percentage of time spent in significant locations changes with T, as reported in Figure 2(b): this percentage quickly decreases with T but the steepness of the curve changes at $T = 0.15$. Hence, we choose the value of $T = 0.15$ for this dataset. These values of T result in an average number of about 24 and 12 places per user for the Cabspotting and CenceMe GPS datasets, respectively.

When dealing with GPS measurements, the duration of a visit can be computed as the difference between two consecutive GPS samples. However, the GPS measurement process usually involves a periodic sampling of the location. When the user is located for a long time interval inside the same region, this results in several successive short visits being recorded, whose length depends on the adopted sampling interval. The same problem may occur with WiFi association logs: since WiFi connectivity may be intermittently available and handoff mechanisms are in place in this type of network infrastructure, a long residence time may be split in several shorter sessions.



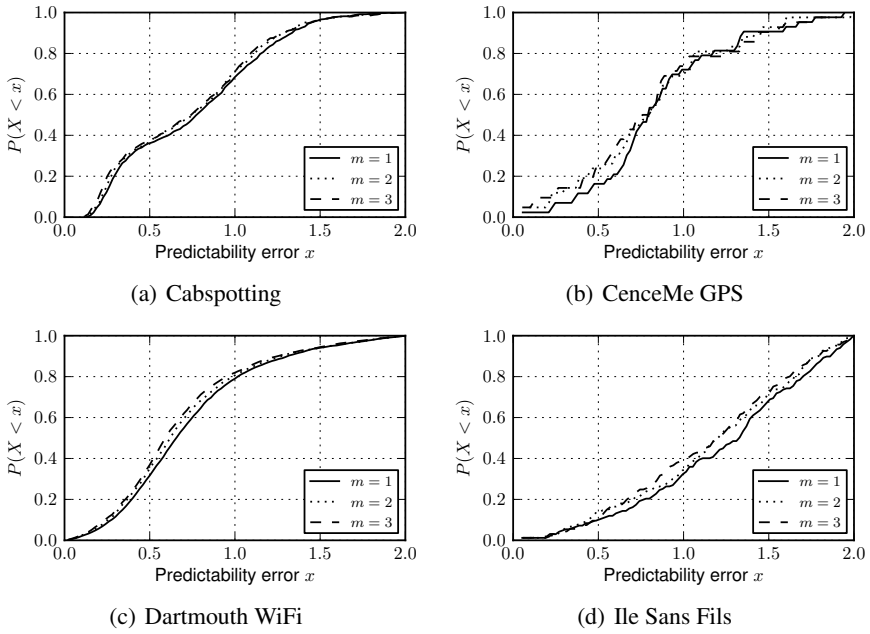
(a) (b)

Fig. 2: Average number of significant places per user (a) and percentage of time spent in significant locations (b) as a function of the threshold T of the frequency map for the Cabspotting dataset and the CenceMe GPS dataset.

In order to infer a more accurate residence time of the user in a certain region, we apply a merging procedure to the dataset of the sequence of visits. Given a sequence of visits to the same location $(t_1, d_1), (t_2, d_2), \ldots (t_n, d_n)$, if the end time of a visit is close to the start time of the next one, that is if $t_{i+1} - (t_i + d_i) \leq \delta$, we merge them in a new visit starting at t_i and ending at $t_{i+1} + d_{i+1}$. In this way the visits obtained are more likely to mimic the real patterns of presence of users, thus improving prediction. We adopted the value of $\delta = 60$ seconds for the Cabspotting dataset and $\delta = 180$ seconds for the CenceMe GPS dataset, since these are the values of the scanning period for the GPS data acquisition. On the other hand, we apply the same merging procedure to WiFi association logs in the Dartmouth and Ile Sans Fils datasets with a value of $\delta = 300$ seconds, in order to filter out casual disconnections from the access point which may last for few minutes.

From a statistical point of view, these datasets show different characteristics, as reported in Table 1: while Cabspotting, Dartmouth WiFi and Ile Sans Fils contain measurements for hundreds or thousands of users, CenceMe GPS consists of data related to a smaller group of moving users. On average about 12 significant locations have been recorded for each user in the CenceMe GPS dataset. In the Dartmouth WiFi and Cabspotting datasets the number of significant places is 18 and 24, respectively. On the other hand, in the Ile Sans Fils dataset we have less than 4 significant locations per user. This is due to the fact that the Ile Sans Fils dataset contains association logs with access points located in public spaces, thus, a large portion of individuals are seen just in few locations. In fact, public access point are not likely to capture some important places for a given user, such as his/her home and working place. There are also differences in the residence time of users in their significant locations: while for Ile Sans Fils and Dartmouth WiFi the average residence time is about 90 and 30 minutes, in the Cabspotting and CenceMe GPS datasets it is about 5 and 10 minutes.

Finally, the amount of time spent in significant locations is crucial to the investigation of the performance of the location prediction technique. While in the CenceMe GPS and in the Dartmouth WiFi datasets each user spends on average 14.74% and 11.24% of their time in a significant location, this value drops to 7.27% in the Cabspot-



(a) Cabspotting

(b) CenceMe GPS

(c) Dartmouth WiFi

(d) Ile Sans Fils

Fig. 3: Cumulative Distribution Function of the predictability error of the time series of the start instants extracted from the various datasets. We report the results for different values of the embedding dimension m adopted in the prediction method.

ting dataset and to 0.18% in the Ile Sans Fils dataset, since it covers a longer period of time (more than one year) and many of its users are present less regularly than in the other datasets.

3.3 Time Series Predictability Test

In order to exploit time series techniques to predict user behavior, we first need to investigate if determinism is present in the extracted time series. In other words, we want to evaluate the *predictability* of these time series.

Let us consider a time series $(s_0, s_1, \ldots, s_N)$. If a real measurement for s_{N+1} is given, the prediction error is the difference between s_{N+1} and the predicted value p_{N+1}. Given a prediction technique, it is possible to obtain predicted values $(p_0, p_1, \ldots, p_N)$ for the whole time series. Then, the *mean quadratic prediction error* can be evaluated as $\varepsilon = \frac{1}{N} \sum_{n=1}^{N} (s_n - p_n)^2$. Large values of ε indicate that the prediction is not accurate and the time series is not predictable.

The evaluation of ε is based on the comparison to the variance σ^2 of the time series: thus, a convenient way of deciding whether ε is small or large is to take the ratio $\frac{\varepsilon}{\sigma^2}$, which is the *predictability error*: if this ratio is close to 1, then, the mean quadratic prediction error is large, while if it is close to 0, the mean quadratic prediction error is small. We refer to this ratio as the *predictability error* of a prediction algorithm. The

absolute error value ε may be meaningless if not compared to the average amount of fluctuations a time series exhibits: by dividing by the variance of the series we can normalize the error and compare the prediction accuracy for different time series.

We exploit this metric to evaluate whether the time series extracted from user visits in the different datasets are predictable. We divide each dataset in two halves: we use the first half to build the model and we compute predicted values of the second half and vice versa. A value equal to 1 means that no determinism is present in the time series, since in this case the predictor has the same accuracy of the simple average value, whereas a value closer to 0 indicates a high degree of determinism.

In Figure 3 we show the Cumulative Distribution Function of the predictability error for the time series of the visit start times for different values of the embedding dimension m. We have also investigated the predictability error for the time series of visit end times, obtaining similar results, which we do not show due to space limitations. On average, a large proportion of users exhibit predictability: in the Dartmouth WiFi dataset 80% of the time series show predictability error smaller than 1, whereas in the CenceMe GPS and Cabspotting datasets the same figure is 70% and it drops to 40% in the Ile Sans Fils dataset, which show less predictability than the others. This is due to the fact that visits may not occur every day with the same pattern for access points in public places, since different individuals are likely to show less regularity in public space than in more personal locations as living or working places, which are not present in this dataset. Moreover, in all datasets the predictability error is lower for higher values of the embedding dimension m: this confirms that nonlinear methods improve prediction quality, since they are able to capture and recognise specific patterns of visits and to estimate when the next visit will be. However, we have noticed that values of $m \geq 4$ show worse performance because we do not have sufficient statistics in order to make a correct prediction.

Interestingly, we expected to observe a lower degree of regularity in the Cabspotting traces, since the movements of a taxi are related to the destinations of the different customers and these destinations can be hardly predictable. Nonetheless, we were able to identify a set of places among which taxis move with more regular patterns. These places correspond to areas where taxi drivers periodically go and wait for new customers, such as touristic locations, shopping malls, cinemas, and they tend to exhibit regular and predictable patterns.

3.4 Evaluating Prediction Accuracy

We compare the performance of NextPlace with those of other two methods: a state-of-the-art Markov-based spatio-temporal predictor and a modified version of NextPlace, where time series of visits are predicted with linear methods rather than with nonlinear algorithms.

Methodology Firstly, we compare NextPlace with a more sophisticated *spatio-temporal Markov predictor* derived by extending the techniques presented in [26]. To the best of our knowledge, this is the most accurate algorithm that has been presented in the literature for this class of prediction problems, because it combines spatial and temporal

dimensions to estimate both next location and handover time for users in a cellular network.

Consider a user visit history among several locations $H = (t_1, d_1, l_1), \ldots, (t_n, d_n, l_n)$, where t_i is the time when the user arrived at location l_i and d_i is the residence time in that location. Then, from H we extract the location history $L = l_1, \ldots, l_n$ and the order-k location context $L_k = L(n-k+1, n) = l_{n-k+1}, \ldots, l_{n-1}, l_n$. The history L is searched for instances of the context L_k and, for each destination that follows an instance, we examine the duration of the previous residence time. More formally, we extract the following set of inter-arrival times A_x and set of durations D_x for each possible destination x:

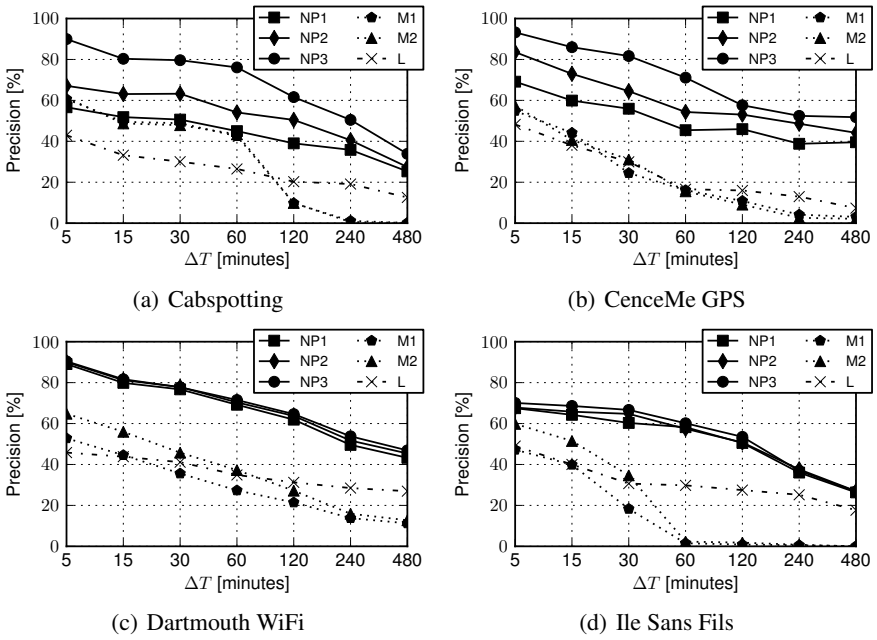
$$A_x = \{t_{i+1} - t_i \quad \text{if} \quad L(i-k+1, i+1) = (L_k, x)\}$$
$$D_x = \{d_{i+1} \quad \text{if} \quad L(i-k+1, i+1) = (L_k, x)\}$$

Then, we compute the estimated time when the user will move to location x and the estimated residence time in x by using a CDF predictor with probability $p = 0.8$ [26]. Moreover, a Markov predictor of order k is used to assign the probability of transition between the current location and the possible destinations. Finally, spatial and temporal information are combined to obtain the predicted location. In order to predict not only the next location but also the subsequent ones, we extend this approach taking the predicted location as the current one and computing again the next movement. We refer the interested reader to the original paper for further details [26].

To understand how largely NextPlace relies on the performance of the nonlinear time series predictor, we can design a linear version of our prediction technique. We use an *order-k running average predictor* instead of a nonlinear method to estimate the future values of a time series: given the sequence of previous visits of a user to a location, the last k visit duration times and k intervals between visits are averaged to obtain a prediction of future visits. Then, the future location is chosen among several predicted locations according to the same algorithm at the basis of the nonlinear predictor (presented in Section 2.2). However, this simplistic time series predictor ignores how user behavior changes over time, since high heterogeneity can be observed in visits occurring during different times of the day. Focusing only on recent data and not investigating these temporal aspects may not be sufficient to obtain accurate estimates.

Results We now evaluate the performance of NextPlace with the nonlinear predictor presented in Section 2.2 compared to the other predictors previously described.

We use the following definition of correctness: if we predict, at time T, that the user i will be at location l at time $T_P = T + \Delta T$, the prediction is considered correct only if the user is at l at any time during the interval $[T_P - \theta, T_P + \theta]$, where θ is the error margin. It is important to note that each prediction algorithm can also estimate if the user will not be in any of her significant places: thus, a prediction may be correct whether the user is predicted to be in a particular location l and then he/she is in l or if the user is predicted not to be in any significant location and then, in fact, she is not. However, as reported in Table 1, the fraction of time that on average users spend



(a) Cabspotting (b) CenceMe GPS

(c) Dartmouth WiFi (d) Ile Sans Fils

Fig. 4: Prediction precision as a function of time interval ΔT for the different datasets and for different predictors: NextPlace with nonlinear predictor for different values of embedding dimension $m = 1, 2, 3$ (NL1-NL2-NL3), first-order and second-order Markov-based (M1-M2) and NextPlace with linear predictor (L). Error margin is $\theta = 900$ seconds.

in their significant locations ranges between 14.74% in the CenceMe GPS dataset and only 0.18% in the Ile Sans Fils dataset. Hence, it is not easy to understand if predictions are accurate because a method is performing well or because, on average, it is just easier to predict the user outside of all her significant locations.

Therefore, we introduce an accuracy metric that takes into account this issue. We define the *prediction precision* as the ratio between the number of correct predictions and the number of all attempted predictions which forecast the user to be in a significant location. We do not consider for the evaluation any predictions which forecast the user outside her significant locations.

We report the performance of different predictors: we test NextPlace with different values of the embedding parameter $m = 1, 2, 3$, two order-1 and order-2 Markov-based predictors and the linear version of NextPlace with a running average predictor considering the last $m = 4$ values. For each dataset, we use the first half to build a prediction model and then we compute predictions during the second half and, for each user, we make 1000 predictions at uniformly distributed random instants. Finally, prediction precision is computed and we investigate how it changes with ΔT, using an error margin $\theta = 900$ seconds. All results are averaged over 20 runs with different random seeds.

We see in Figure 4 that for all datasets, NextPlace with its nonlinear predictor is always outperforming the other methods. We also note that using a higher value of

m improves prediction quality, as it can be appreciated especially in the GPS-based Cabspotting and CenceMe GPS datasets. Similarly, Markov models are able to provide correct predictions when ΔT is smaller than 1 hour: however, except for the Ile Sans Fils dataset, the performance of the nonlinear NextPlace are at least about 50% better of the Markov-based predictors, since they reach a maximum precision of 60% while NextPlace achieves a precision higher than 90%. Moreover, when ΔT increases, the precision of Markov predictors decreases rapidly and the performance gap with the nonlinear approach widens. This can be explained by the fact that Markov predictors are generally employed to predict the next movement and, thus, when predictions are extended in the future, movement after movement, a large error is accumulated.

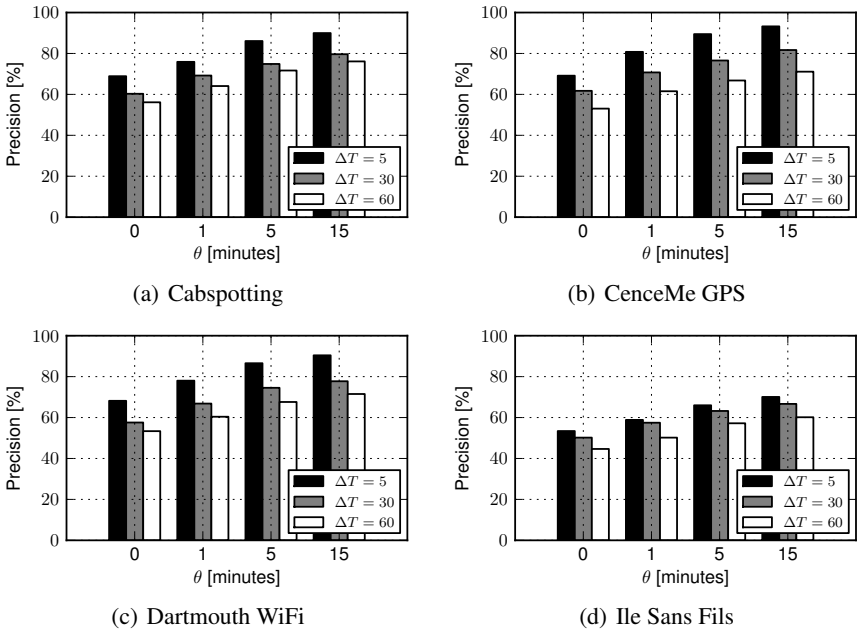
If we substitute the nonlinear predictor in NextPlace with a linear one, we observe a similar trend but precision is considerably lower, since errors on time series prediction are larger and, hence, affect the location prediction. However, NextPlace with both nonlinear and linear predictors is less dependent on ΔT than Markov models, which show a lower precision when predictions are extended in the future. Again, this demonstrates how NextPlace, which focuses only on temporal information of visits in significant places, is more robust for long-term predictions.

As discussed in Section 3, the Ile Sans Fils dataset exhibits less predictability. This is confirmed by the analysis of prediction precision, which shows the lowest figures among all the datasets. The other datasets score a precision equal to about 90% for $\Delta T = 5$ minutes and around 70% for $\Delta T = 60$ minutes. We also investigate the impact of the error margin θ on prediction results: prediction precision is lower for smaller error margins, but it shows the same trends for all predictors and for all the datasets. In Figure 5 we report how prediction precision of our nonlinear approach with $m = 3$ is affected by different error margins for some values of ΔT. Even with $\theta = 0$, which represents the worst case scenario, prediction precision is between 50% and 60% after $\Delta T = 60$ minutes for all datasets except Ile Sans Fils, where it is below 50%.

From a general point of view, our evaluation shows how NextPlace achieves high prediction accuracy, even for long-term predictions made some hours in advance. Furthermore, these results also show how focusing on spatial movements, as Markov models do, may be useful only for short-term predictions. Instead, focusing just on temporal information about recurrent patterns in significant places proves to be more robust both for short-term and long-term predictions, since NextPlace outperforms Markov models even for small values of ΔT.

4 Related work

Pioneering work [3,4] has focused on the analysis of mobility traces in order to gain insight about human mobility patterns. Key papers in this area include studies on mobility and connectivity patterns, such as [5, 13]. The main findings are that contact duration and inter-contacts time between individuals can be represented by means of power-law distributions and that these patterns may be used to develop more efficient opportunistic protocols [10]. In addition, temporal rhythms of human behavior have been studied and modeled to discover daily activity patterns, to infer relationships and to determine significant locations [7]. This related body of work concentrates on the *statistical char-*



(a) Cabspotting (b) CenceMe GPS

(c) Dartmouth WiFi (d) Ile Sans Fils

Fig. 5: Prediction precision of NextPlace with nonlinear predictor with $m = 3$ as a function of error margin θ and for different values of ΔT.

acterization of temporal behavioral patterns of groups of users, whereas we concentrate on prediction of single users.

The evaluation of prediction techniques applied to the problem of forecasting the next location (but not the arrival time to that location and the corresponding residence time) are presented in [27]. A prediction framework based on spatio-temporal patterns in collective mobility trajectories has been presented in [22]: this method attempts to predict the next location of a moving object by matching a new trajectory to a corpus of global frequent ones. While this prediction technique is more general, as it captures dependencies between visits at different places, our method includes time-of-day information and does not rely on global patterns, allowing prediction to be made also for users who deviate from collective behavior. In [2] the authors present a model of user location prediction from GPS data. A simple first-order Markov model to predict the transitions between significant places is used, albeit in this work temporal aspects are not taken into consideration. In [19] the significant places are extracted by means of a discriminative relational Markov network; then, a generative dynamic Bayesian network is used to learn transportation routines. Another system for the prediction of future network connectivity based on a second-order Markov model is BreadCrumbs [23]. Again, this system is able to predict only the next location of the user and not the time of the transitions and the interval of time during which users reside in that specific location. Similarly, Markov based techniques have also been applied to the prediction of the destinations (geographical locations) of vehicles using for example partial trajec-

tories [16]. As we have shown in the evaluation section, this class of models is able to provide precise predictions only for instants of time close in the future, given the inherent memorylessness of Markov predictors.

5 Conclusions

In this paper we have presented NextPlace, a new approach to spatio-temporal user location prediction based on nonlinear analysis of the time series of start times and duration times of visits to significant locations. To the best to our knowledge, this is the first approach that not only allows to forecast the next location of a user, but also his/her *arrival* and *residence time*, i.e., the interval of time spent in that location. Moreover, existing models are not able to extend predictions further in the future, since they mainly focus on the next movement of a user.

We have evaluated NextPlace comparing it comparing it with a version based on a linear predictor and a probabilistic technique based on spatio-temporal Markov predictors over four different datasets. We have reported an overall prediction precision up to 90% and a performance increment of at least 50% over the state of the art. We have showed how the adoption of a nonlinear prediction framework can improve prediction precision with respect to other techniques even for long-term predictions.

As future work, there is a number of potential improvements that can be pursued. Regular collective human rhythms can be exploited to refine the prediction and a probabilistic framework can be used to choose between equally promising next locations, giving more flexibility to applications. Finally, we are interested in the investigation of prediction models which take into account human rhythms on a weekly basis, in order to better capture regular human behavior on a longer time scale.

References

1. L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala. Bluetooth and WAP Push Based Location-aware Mobile Advertising System. In *Proceedings of MobiSys '04*, pages 49–58, 2004.
2. D. Ashbrook and T. Starner. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Journal of Personal and Ubiquitous Computing*, 7(5):275–286, October 2003.
3. A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan. Characterizing User Behavior and Network Performance in a Public Wireless LAN. In *Proceedings of SIGMETRICS '02*, 2002.
4. M. Balazinska and P. Castro. Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network. In *Proceedings of MobiSys '03*, San Francisco, CA, May 2003.
5. A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of Human Mobility on Opportunistic Forwarding Algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, June 2007.
6. C. Chatfield. *The Analysis of Time Series: An Introduction - Fifth Edition*. Chapman & Hall/CRC, London, July 1995.
7. N. Eagle and A. S. Pentland. Reality Mining: Sensing Complex Social Systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 2006.

8. M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding Individual Human Mobility Patterns. *Nature*, 453(7196):779–782, June 2008.
9. T. Henderson, D. Kotz, and I. Abyzov. The Changing Usage of a Mature Campus-wide Wireless Network. In *Proceedings of MobiCom '04*, pages 187–201, New York, NY, USA, 2004.
10. S. Jain, K. Fall, and R. Patra. Routing in a Delay Tolerant Network. In *Proceedings of SIGCOMM '04*, 2004.
11. J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello. Extracting Places from Traces of Locations. *SIGMOBILE Mobile Computing Communication Review*, 9(3):58–68, 2005.
12. H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, 2004.
13. T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovic. Power Law and Exponential Decay of Inter-contact Times Between Mobile Devices. In *Proceedings of MobiCom '07*, pages 183–194, 2007.
14. M. Kim, D. Kotz, and S. Kim. Extracting a Mobility Model from Real User Traces. In *Proceedings of INFOCOM'06*, April 2006.
15. D. Kotz, T. Henderson, and I. Abyzov. CRAWDAD trace dartmouth/campus/movement/01_04 (v. 2005-03-08). Downloaded from http://crawdad.cs.dartmouth.edu/, March 2005.
16. J. Krumm and E. Horvitz. Predestination: Inferring Destinations from Partial Trajectories. In *Proceedings of UbiComp '06*, September 2006.
17. A. Lamarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, J. Tabert, P. Powledge, G. Borriello, and B. Schilit. Place Lab: Device Positioning Using Radio Beacons in the Wild. In *Proceedings of Pervasive '05*, May 2005.
18. M. Lenczner, B. Gregoire, and F. Roulx. CRAWDAD data set ilesansfil/wifidog (v. 2007-08-27). Downloaded from http://crawdad.cs.dartmouth.edu/ilesansfil/wifidog, August 2007.
19. L. Liao, D. J. Patterson, D. Fox, and H. Kautz. Building Personal Maps from GPS Data. In *Proceedings of IJCAI Workshop on Modeling Others from Observation*, 2005.
20. N. Marmasse and C. Schmandt. Location-Aware Information Delivery with ComMotion. In *Proceedings of HUC '00*, pages 157–171, London, UK, 2000. Springer-Verlag.
21. E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. Sensing Meets Mobile Social Networks: the Design, Implementation and Evaluation of the CenceMe Application. In *Proceedings of SenSys '08*, pages 337–350, New York, NY, USA, 2008. ACM.
22. A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. WhereNext: a location predictor on trajectory pattern mining. In *Proceedings of SIGKDD '09*, KDD '09, pages 637–646, New York, NY, USA, 2009. ACM.
23. A. J. Nicholson and B. D. Noble. BreadCrumbs: Forecasting Mobile Connectivity. In *Proceedings of MobiCom '08*, pages 46–57, New York, NY, USA, 2008. ACM.
24. M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. CRAWDAD trace set epfl/mobility/cab (v. 2009-02-24). Downloaded from http://crawdad.cs.dartmouth.edu/epfl/mobility/cab, Feb. 2009.
25. T. Schreiber. Efficient Neighbor Searching in Nonlinear Time Series. *International Journal on Bifurcations and Chaos*, 5:349–358, 1995.
26. L. Song, U. Deshpande, U. C. Kozat, D. Kotz, and R. Jain. Predictability of WLAN Mobility and its Effects on Bandwidth Provisioning. In *Proceedings of INFOCOM '06*, April 2006.
27. L. Song and D. Kotz. Evaluating Location Predictors with Extensive Wi-Fi Mobility Data. In *In Proceedings of INFOCOM '04*, pages 1414–1424, 2004.
28. C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. Discovering Personally Meaningful Places: An Interactive Clustering Approach. *ACM Trans. Inf. Syst.*, 25(3):12, 2007.