Improving Location Prediction Services for New Users with Probabilistic Latent Semantic Analysis

James McInerney, Alex Rogers, Nicholas R. Jennings
University of Southampton, Southampton, SO17 1BJ, UK
{jem1c10,acr,nrj}@ecs.soton.ac.uk

ABSTRACT

Location prediction systems that attempt to determine the mobility patterns of individuals in their daily lives have become increasingly common in recent years. Approaches to this prediction task include eigenvalue decomposition [5], non-linear time series analysis of arrival times [10], and variable order Markov models [1]. However, these approaches all assume sufficient sets of training data. For new users, by definition, this data is typically not available, leading to poor predictive performance. Given that mobility is a highly personal behaviour, this represents a significant barrier to entry. Against this background, we present a novel framework to enhance prediction using information about the mobility habits of existing users. At the core of the framework is a hierarchical Bayesian model, a type of probabilistic semantic analysis [7], representing the intuition that the temporal features of the new user's location habits are likely to be similar to those of an existing user in the system. We evaluate this framework on the real life location habits of 38 users in the Nokia Lausanne dataset, showing that accuracy is improved by 16%, relative to the state of the art, when predicting the next location of new users.

Author Keywords

Knowledge Representation and Reasoning::Geometric, Spatial, and Temporal Reasoning, Machine Learning::Data Mining, Machine Learning::Machine Learning (General/other, Reasoning under Uncertainty::Uncertainty in AI (General/other), Machine Learning::Unsupervised Learning

ACM Classification Keywords

H.5.2 User/Machine Systems: I.5 Pattern Recognition

General Terms

Experimentation, Performance.

INTRODUCTION

Location prediction of daily life mobility has been a topic of considerable interest in recent years [10, 5]. In general, location data is gathered about an individual user with global positioning system (GPS), cell tower or wifi data, and this history of locations is used to predict the user's future locations. Predicting user mobility gives the promise of many exciting real world ubiquitous services. Better mobile reminders, search results, and advertisements are likely to result from knowing where the user will be [10].

Existing approaches typically assume an adequate history of observations of user mobility in order to train a statistically accurate model of behaviour. Yet, in real life, we know that this will not usually be available for a new user. This presents an important barrier to the success of ubiquitous services. For any service to grow in its number of users, a high proportion of those users will necessarily be new. But performance of the service is at its worst (due to poor predictions) precisely at the time when the user has just started using it. Nevertheless, new users are important precisely because they often have the responsibility of deciding whether to commit to or abandon the service.

This problem suggests the need to exploit the similarities between new and existing users, an approach often used in recommender systems [12] and recently introduced in systems for activity recognition [8]. In the domain of mobility prediction, it is known that many people share a common set of mobility habits [5], such as going to work during weekdays, staying at home on weekend mornings, and going to new places in evenings. These similarities could be used to increase accuracy for new users. While this approach makes sense intuitively, it is hard to exploit in location prediction because it requires a semantic understanding of user location (e.g., home, work, sports centre, or restaurant). That is, a user's history of locations is made up of points in geographical space, but generalizing the habitual elements usually requires conversion to general and meaningful labels before behaviour correlation analysis can begin [5]. However, semantic labelling of locations is challenging to achieve, even for individuals with larger data sets.

Yet, arguably, deriving semantic labels for locations is an unnecessary requirement for the problem of learning user models of mobility. To get the benefits of accurate predictions, we only require that the states (i.e., the locations) of the new user's model be correctly linked to those of the much richer model of a similar established user. For example, if the habit of going to a certain train station to go home after work appears in a user's history, then this pattern may be used for prediction without explicitly knowing the meaning of the locations. Furthermore, if a similar pattern appears in the history of a new user, then linking their respective transport hubs, homes and places of work could enable richer predictions, such as a lower probability of going home from that station at weekends. In short, we can remain indifferent to

UbiComp '12, September 5-8, 2012, Pittsburgh, USA.
ACM 978-1-4503-1224-0/12/09.

the precise interpretation of locations because we primarily care about the dynamics of the model, which can be empirically accurate and generate good predictions even if the underlying semantic information (e.g., home, work) is missing.

To address this shortcoming, we present the first approach to model such *functional mappings* between existing and new users, to significantly improve location prediction for the latter group, without requiring any semantic labelling. In doing so, we make several contributions:

- We develop a hierarchical Bayesian model, based on probabilistic latent semantic indexing [7], for matching the locations of a new user with those of existing users. The model makes no cultural-specific assumptions about habits (such as going to work on weekdays) and uses no extra information about locations.
- We show how this mapping can be used to increase prediction accuracy for new users, as compared to a model that does not use the mapping. Specifically, it is possible to transform the transition matrix of a Markov model representing an established users' mobility, to a mobility model that approximates well the habits of the new user. In general, the benefit of inferring the mapping between the locations of new and established users is that it does not require commitment to any specific prediction model.
- We validate our approach using the location histories of 38 real world users from the Lausanne Nokia dataset. We simulate the arrival of new users to a location prediction system by truncating their location history and find that prediction accuracy is improved 16% relative to a state of the art predictor using our approach.

When taken together, these contributions open the way for improved performance for new users in predictive systems with a minimal amount of assumed knowledge.

In the remainder of this paper, we first introduce our model of functional mapping for locations in Section 2, illustrating how such mappings can be used to transform a Markov model from that of the established user to the new user (Section 3), and then applying it to the Nokia dataset to validate the approach (Section 4). Finally, Section 5 concludes.

MODEL OF FUNCTIONAL LOCATIONS

To formalise the observed temporal similarity between the mobility patterns of users, we assume a fixed number, T, of time slots, each with a probability distribution over L significant locations of the user. We represent the probability that a single user will be in significant location i at time t with a $T \times L$ probability matrix $\boldsymbol{M}$, responsible for generating the actual observations $\boldsymbol{X}$, of which we assume there are N. Hence, $\boldsymbol{X}$ is an $N \times L$ binary matrix, with one row for each observation. Clearly, the user can be in only one location at a time, resulting in a 1-of-L choice at each time step (i.e., for each row in $\boldsymbol{X}$). A natural assumption for such categorical variables is therefore that the probability distribution over the set of locations (i.e., presence) for each time slot t follows a multinomial distribution [3]. The *sufficient statistics* of the multinomial distribution, that is, the only information required from the raw observations, is the $T \times L$ matrix of integer counts, $\boldsymbol{C}$, representing the sum of presence counts for each time slot of the history $\boldsymbol{X}$, where element c_{ti} is the number of times the user was observed at location i at time slot t.

The likelihood of an observed history of presence counts $\boldsymbol{X}$ can therefore be found via the definition of a multinomial distribution [3], assuming that all observations are independently and identically distributed:

$$p(\boldsymbol{X}|\boldsymbol{M}) \propto \prod_{t=1}^{T} \prod_{i=1}^{L} m_{ti}^{c_{ti}} \tag{1}$$

The total number of observations (i.e., $N = \sum_{t=1}^{T} \sum_{i=1}^{L} c_{ti}$) will clearly be higher for an established user than for a new user. Our model attempts to address this disparity in the number of observations by explicitly linking the two users' $\boldsymbol{M}$ parameters. Throughout, we refer to the random variables specific to the established user (i.e., $\boldsymbol{X'}$, $\boldsymbol{C'}$ and $\boldsymbol{M'}$) with the apostrophe modifier, to distinguish them from those of the new user (i.e., $\boldsymbol{X}$, $\boldsymbol{C}$ and $\boldsymbol{M}$).

To address the disparity in the number of observations between the new and established user, we make the key assumption that the location behaviour of the new user is generated entirely from the probability distribution of the established user, *subject to some transformation of locations*. More formally, we assume that $\boldsymbol{M'R} = \boldsymbol{M}$, where $\boldsymbol{R}$ is an unknown transformation matrix. This work focuses primarily on inferring $\boldsymbol{R}$ from the observed location histories of both users, and using it to enhance prediction.

In more detail, $\boldsymbol{R}$ can be interpreted as the mapping of locations between two specific users, A and B. In general, there are two types of mapping. The first is one-to-many, in which user A's presence in a single location tends to co-occur with presence in multiple locations for user B. For example, user A may work in a single office, while user B may spend half her time working in a lab and the other half in the library. As another example, user A might tend to visit a single cafeteria for lunch, while user B might visit multiple sandwich shops nearby. The inverse relationship, many-to-one, is also possible. Multiple locations for user A can co-occur with just a single location for user B. Clearly, the one-to-one mapping, in which both users have a single location in which they tend to be at the same time (e.g., home) is just a special case of either of these transformations. Inferring this mapping goes beyond simply smoothing the probability densities of sparse presence observations [6] because it enables the reuse of rich densities from other real world users.

To find $\boldsymbol{R}$, the naïve derivation of the new user's probability distribution over locations uses $\boldsymbol{R}$ to directly transform the established user's probability matrix:

$$p(\boldsymbol{X}|\boldsymbol{R},\boldsymbol{M}) = \prod_{t=1}^{T}\prod_{i=1}^{L}\left(\sum_{j=1}^{L} r_{ji}m_{ti}\right)^{c_{ti}} \quad (2)$$

A full Bayesian approach then seeks the posterior distribution of the $\boldsymbol{R}$ parameter, i.e., $p(\boldsymbol{R}|\boldsymbol{X},\boldsymbol{M})$. However, we see that this cannot be done tractably, due to the inner summation in Equation 2. This is true even with a maximum likelihood approach (i.e., if we try to maximize Equation 2 without a prior). However, as is common for such situations, by introducing a set of latent variables (one for each observation vector $\boldsymbol{x_n}$) we can derive a tractable joint distribution over both observed and latent variables.

Let latent variable $\boldsymbol{z_n}$ be a binary vector of length L representing the (unobserved) location of the established user at time n (and $\boldsymbol{Z}$ the matrix in which these vectors correspond to the rows). Therefore, under our model of mobility, $\boldsymbol{z_n}$ has a multinomial distribution. This latent variable in turn is used to select which row of $\boldsymbol{R}$ is used as the generative probability distribution of the new user's location at time n. Given that we are dealing with small numbers of observations, the maximum likelihood approach is likely to overfit the data [3]. We therefore choose prior distributions over the multinomial random variables $\boldsymbol{M}$ and $\boldsymbol{R}$. The natural choice of prior is the Dirichlet distribution, which is conjugate to multinomials [3]. Conjugacy means that the posterior and prior have the same form, with respect to the likelihood function, and is extremely useful because it limits complexity. The Dirichlet has a hyperparmeter representing the prior observation count. $\boldsymbol{M}$ and $\boldsymbol{R}$ are assumed to have hyperparameters α and $\boldsymbol{\beta}$, respectively.

The resulting generative hierarchical model of new user location behaviour is presented graphically in Figure 1 and summarized in Algorithm 1. It is similar to probabilistic latent semantic indexing [7], which models text documents as distributions over topics, which in turn generate bags of words in the training set. Documents are analogous to the time slots of the established user, where each time slot has a different distribution over locations (i.e., topics).

Where our approach differs is that we care a lot about the sparsity of observations of new user locations (i.e., words), so assume that the generating conditional probabilities R also follow a Dirichlet distribution. Additionally, we have extra information that we need to integrate, in the form of observations, $\boldsymbol{X'}$, of the topics themselves (i.e., established user locations).

Our model also has strong similarities with latent Dirichlet allocation (LDA) [4], which also models documents, topics, and bags of words hierarchically, but is concerned with the problem of *unseen* documents, and so assumes that the distributions over topics are themselves randomly generated, forming a three-level hierarchy. In contrast, we are comfortable with a predefined set of time slots that repeatedly generate observed locations, as this conforms to strong daily and weekly periodicities found in human location behaviours [5].

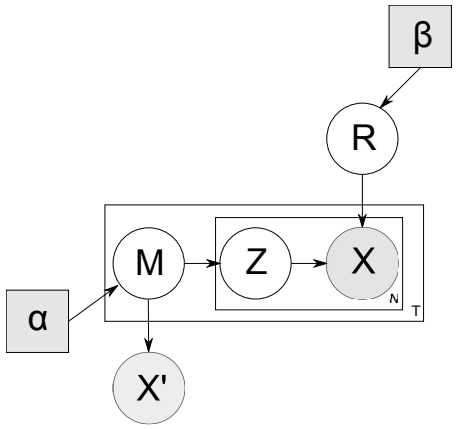

Figure 1. Hierarchical Bayesian model for new user behaviour. Shaded nodes represent observed variables.

Algorithm 1 Generative probabilistic model of new user mobility

1. $\boldsymbol{z_n} \leftarrow Dirichlet(\boldsymbol{m_t} + \boldsymbol{\alpha})$
2. $\boldsymbol{r} \leftarrow \boldsymbol{z_n}\boldsymbol{R}$
3. $\boldsymbol{x_n} \leftarrow Dirichlet(\boldsymbol{r} + \boldsymbol{\beta})$

Finding the MAP of $\boldsymbol{R}$ in the model shown in Figure 1 is done by maximizing $p(\boldsymbol{R}|\boldsymbol{X},\boldsymbol{X'},\boldsymbol{Z},\boldsymbol{M},\alpha,\boldsymbol{\beta})$. This can be achieved with a widely-used algorithm for models involving latent variables, namely, expectation-maximization (EM) [3]. The steps of the EM algorithm, as applied to our model, are given in Algorithm 2.

Algorithm 2 Expectation-maximization algorithm for estimating the maximum *a posteriori* of parameter $\boldsymbol{R}$

$\boldsymbol{R^{old}} \leftarrow$ initialize randomly
$\boldsymbol{M^{old}} \leftarrow$ initialize randomly
repeat
  E-step $\boldsymbol{\gamma} \leftarrow \mathbb{E}[\boldsymbol{z}]$ (Equation 3)
  M-step $\boldsymbol{M^{new}} \leftarrow \arg\max_{\boldsymbol{M}} p(\boldsymbol{M}|\boldsymbol{X'},\boldsymbol{Z},\alpha)$ (Equation 6)
  M-step $\boldsymbol{R^{new}} \leftarrow \arg\max_{\boldsymbol{R}} p(\boldsymbol{R}|\boldsymbol{M},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\beta})$ (Equation 7)
  $d \leftarrow \mathbf{1}\mathbf{1}^{\mathbf{T}} abs(\boldsymbol{R^{new}} - \boldsymbol{R^{old}})$
  $\boldsymbol{R^{old}} \leftarrow \boldsymbol{R^{new}}$
  $\boldsymbol{M^{old}} \leftarrow \boldsymbol{M^{new}}$
until $d \leq \epsilon$

We now detail the derivation of the equations used in the two steps of Algorithm 2, specifically, the three equations for the iterative updates of $\boldsymbol{M}$, $\boldsymbol{R}$, and $\boldsymbol{Z}$. Starting with the E-step, the expectation of the latent variable $\boldsymbol{Z}$ can be found by applying Bayes' theorem and the standard updating expression of the Dirichlet distribution over $\boldsymbol{R}$ (one of the assumptions of the model):

$$\begin{aligned}\mathbb{E}(z_{nk}) &= \frac{p(z_{nk}=1)p(\boldsymbol{x_n}|z_{nk}=1)}{\sum_{j=1}^{L_1} p(z_{nj}=1)p(\boldsymbol{x_n}|z_{nj}=1)} \\ &= \frac{m_{(n \mod T)k} \prod_{i=1}^{L_2} r_{ki}^{x_{ni}+\beta_i-1}}{\sum_{j=1}^{L_1} m_{(n \mod T)j} \prod_{i=1}^{L_2} r_{ji}^{x_{ni}+\beta_i-1}} \quad (3)\end{aligned}$$

The M-step consists of maximizing the posterior distributions of $\boldsymbol{M}$ and then $\boldsymbol{R}$ *as though* $\boldsymbol{Z}$ *had been observed*. Starting with $\boldsymbol{M}$, we can factorize the posterior distribution by remembering that all observations are independently and identically distributed. Thus, the observations of the established user, $\boldsymbol{X'}$, essentially contribute towards the prior counts for the Dirichlet distribution $\boldsymbol{M}$ (representing the probability distribution over locations for the established user):

$$\begin{aligned}p(\boldsymbol{M}|\boldsymbol{X'},\boldsymbol{Z},\alpha) &= p(\boldsymbol{M}|\boldsymbol{X'},\alpha)p(\boldsymbol{M}|\boldsymbol{Z},\alpha) \\ &\propto p(\boldsymbol{X'}|\boldsymbol{M},\alpha)p(\boldsymbol{M})p(\boldsymbol{Z}|\boldsymbol{M},\alpha)p(\boldsymbol{M}) \\ &= \prod_{k=1}^{L_1}\prod_{t=1}^{T} m_{tk}^{v_{tk}} \quad (4) \\ \text{where } v_{tk} &= \sum_{w=1}^{\lfloor \frac{N}{T} \rfloor} z_{(wt)k} + c'_{tk} + 2\alpha - 2\end{aligned}$$

Maximizing $p(\boldsymbol{M}|\boldsymbol{X'},\boldsymbol{Z},\alpha)$ with respect to $\boldsymbol{M}$ is achieved by maximizing its logarithm, with a Lagrangian added to constrain the rows of $\boldsymbol{M}$ to sum to 1 (giving Equation 5). G is then differentiated to find the set of multipliers $\boldsymbol{\lambda}$. The logarithm is used simply to make the differential equation easier to solve:

$$\begin{aligned}G &= \sum_{t=1}^{T}\sum_{k=1}^{L_1} v_{tk} \ln m_{tk} + \sum_{t=1}^{T} \lambda_t \left(1 - \sum_{k=1}^{L_1} m_{tk}\right) \\ \implies \lambda_t &= \frac{v_{tk}}{m_{tk}} \quad (5)\end{aligned}$$

Solving for m_{tk} gives us the MAP:

$$m_{tk} = \frac{c'_{tk} + \alpha - 1 + \sum_{w=1}^{\lfloor \frac{N}{T} \rfloor} \left(z_{(tw)k} + \alpha - 1\right)}{\sum_{k=1}^{L_1} \left(c'_{tk} + \alpha - 1 + \sum_{w=1}^{\lfloor \frac{N}{T} \rfloor} \left(z_{(tw)k} + \alpha - 1\right)\right)} \quad (6)$$

The derivation of the posterior probability of $\boldsymbol{R}$ follows the same procedure to give:

$$r_{ab} = \frac{\left(\sum_{t=1}^{T} z_{ta}x_{tb}\right) + \beta_b - 1}{\left(\sum_{t=1}^{T} z_{ta}C_t\right) + \left(\sum_{i=1}^{L_2} \beta_i\right) - L_2} \quad (7)$$

We now have all three equations necessary for running the EM algorithm on the model. The key output of the procedure is the functional mapping, $\boldsymbol{R}$, between locations of the established user and those of the new user. We next detail how this can be used to enhance prediction.

ENHANCING PREDICTION

Enhancing prediction requires the selection of an established user who has similar location habits to the new user. We do this by finding the established user with the highest posterior probability $p(\boldsymbol{X},\boldsymbol{X'}|\boldsymbol{R},\boldsymbol{Z},\boldsymbol{M},\alpha,\boldsymbol{\beta})$, evaluated on the observed data of the new user. The model of this established user is then mapped, using matrix $\boldsymbol{R}$, to a model approximating the habits of the new user. We next briefly discuss the choice of this mobility model.

There is a wide choice of models for user mobility, including eigendecomposition [5], non-linear time series analysis [10], and Markov models [1, 2]. In general, we leave open the question of which method to use, as this should be specific to the data and intended application. However, to give a concrete example, we detail how our model works with Markov models, which have been applied to mobility modelling in previous settings [1, 2].

A first-order Markov model is represented by a transition matrix of size $L \times L$, indicating the transition probabilities between a given context and the next possible L locations. To represent the transition probabilities between locations of a new user, we use the inferred $\boldsymbol{R}$ matrix to map both the columns and rows to the configuration personal to that user:

$$\boldsymbol{Y_{new}} = \boldsymbol{R}^{\mathbf{T}}\boldsymbol{Y_{est}}\boldsymbol{R} \quad (8)$$

where $\boldsymbol{Y_{new}}$ is the approximated transition matrix for the new user, and $\boldsymbol{Y_{est}}$ is the transition matrix of the established user who best matches that new user. For each observation, the transition probabilities of the next location can be found from the row in $\boldsymbol{Y_{new}}$ corresponding to the current context (i.e., current observed location). The location with the highest probability is always selected as the predicted next location.

REAL-LIFE DATA ANALYSIS

Applying the transformation of Equation 8 to the Markov models of 38 real people allowed us to assess the effectiveness of the approach. We trained a first-order Markov model on the real life location data from the Lausanne Nokia data set [9], which recorded the locations of 38 individuals over the course of a year. To simulate performance of the system on a new user, we truncated the history of a designated 'new' user to the first H hours of observed locations only. Running the EM algorithm (Algorithm 2)[1] and finding the posterior probability $p(\boldsymbol{X},\boldsymbol{X'}|\boldsymbol{R},\boldsymbol{Z},\boldsymbol{M},\alpha,\boldsymbol{\beta})$ allowed us to find the

[1] with hyperparameters α and $\boldsymbol{\beta}$ all set to 1.5

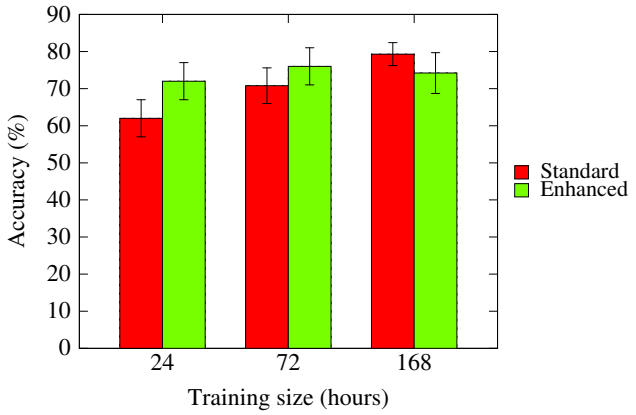

Figure 2. Comparison of performance of the baseline model to the model enhanced with our framework. Error bars indicate the 95% confidence range.

best matching established user from the remaining set of 37 users (whose histories were not truncated).

A first-order transition matrix was then learnt from the best matching established user. We then transformed this matrix, using mapping $\boldsymbol{R}$, to an approximate transition matrix for the new user (with Equation 8). The performance of the approach was evaluated by checking the accuracy of predictions on the rest of the available data for the new user. The same process of simulating a new user was repeated over all individuals in the data set.

As a benchmark, we trained a standard transition matrix on the small amount of truncated data of the new user, which allows us to determine the performance of the Markov model without our framework. Lower order Markov models were previously found to work best on low amounts of training data [11], making this a reasonable benchmark.

Figure 2 shows the results of this procedure for $H = 24$, 72 and 168, i.e., one day, three days, and seven days of observed behaviour as training data, respectively. We see that our framework performs better for very sparse observations (with 24 and 72 hours), implying that our approach is indeed effective at approximating the location habits of new users under these extreme conditions. At 168 hours, no additional improvement is observed in our framework. In contrast, the baseline model improves gradually as the training data size increases. This implies that our framework rapidly reaches its upper limit of performance after only a few days, so should be abandoned after sufficient training sets are made available. Intuitively, the best indicator of future behaviour of an individual is their own past behaviour, once a sufficient history has been gathered.

CONCLUSIONS AND FUTURE WORK

We introduced a new model of location behaviour, capturing the assumption that new users of location prediction services are similar to existing users, subject to an unknown transformation of locations. We applied this model to enhance the accuracy of a first-order Markov model in successfully predicting the next location of real people after just 24 hours of observations.

In future work, we intend to explore ways of making the pairwise choice of new and established users more efficient. Specifically, in a large database of established users, we need to be able to quickly retrieve a relatively small sample of established users most similar to the new user, before applying our training method.

REFERENCES

1. Bapierre et al. A variable order markov model approach for mobility prediction. In *STAMI Workshop at IJCAI*, Barcelona, Spain, 2011.
2. A. Bhattacharya and S. K. Das. LeZi-update: an information-theoretic approach to track mobile users in PCS networks. In *Proc. MobiCom '99*, pages 1–12, 1999.
3. C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.
4. Blei et al. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
5. N. Eagle and A. S. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.
6. Gao et al. Mobile location prediction in spatio-temporal context. In *Nokia Mobile Data Challenge Workshop*, 2012.
7. T. Hofmann. Probabilistic latent semantic indexing. SIGIR '99, pages 50–57, 1999.
8. Lane et al. Enabling large-scale human activity inference on smartphones using community similarity networks. In *UbiComp*, pages 355–364, 2011.
9. Laurila et al. The mobile data challenge: Big data for mobile computing research. In *Proc. Mobile Data Challenge by Nokia Workshop*, 2012.
10. Scellato et al. Nextplace: A spatio-temporal prediction framework for pervasive systems. In K. Lyons, J. Hightower, and E. Huang, editors, *Pervasive Computing*, volume 6696 of *LNCS*, pages 152–169. Springer, 2011.
11. Song et al. Evaluating next-cell predictors with extensive wi-fi mobility data. *IEEE Transactions on Mobile Computing*, 5(12):1633–1649, 2006.
12. X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:1–19, 2009.