

# Classification of Whereabouts Patterns From Large-scale Mobility Data

Laura Ferrari

DISMI, Università di Modena e Reggio Emilia  
Via Amendola 2, 42122 Reggio Emilia, Italy  
laura.ferrari@unimore.it

Marco Mamei

DISMI, Università di Modena e Reggio Emilia  
Via Amendola 2, 42122 Reggio Emilia, Italy  
marco.mamei@unimore.it

## Abstract

*Classification of users' whereabouts patterns is important for many emerging ubiquitous computing applications. Latent Dirichlet Allocation (LDA) is a powerful mechanism to extract recurrent behaviors and high-level patterns (called topics) from mobility data in an unsupervised manner. One drawback of LDA is that it is difficult to give meaningful and usable labels to the extracted topics. We present a methodology to automatically classify the topics with meaningful labels so as to support their use in applications. This mechanism is tested and evaluated using the Reality Mining dataset consisting of about 350000 hours of continuous data on human behavior.*

## 1 Introduction

The recent diffusion of smart phones equipped with localization capabilities allows to collect data about mobility and whereabouts in an economically feasible and unobtrusive way from a large user population [17, 15]. Several institutions, both in academia and industry, are exploiting this technology to create applications to collect logs of people's whereabouts. This information opens new scenarios and possibilities in the development of context-aware services and applications, but several challenges need to be tackled to extract practically useful information from such large mobility datasets. Accordingly, one of the key research is to develop and apply pattern analysis algorithms to such data. However, this kind of research has been impeded until recently by the sheer availability of mobility data from a large user population.

The Reality Mining dataset is a seminal dataset in this area. It collects data about the daily life of 97 users over 10 months. Some pioneering researches started to apply pattern-analysis and data mining algorithm to such mobility dataset in order to extract high-level information and routine behaviors.

A number of these researches focus on two "similar"

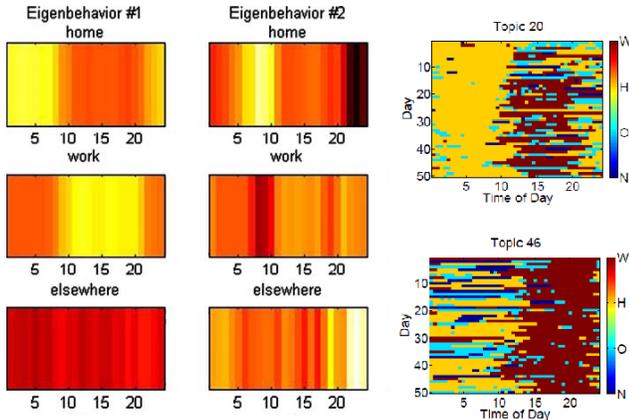
techniques: Principal Component Analysis (PCA) [2] and Latent Dirichlet Allocation (LDA) [3]. The goal of both these techniques is to discover significant patterns and features from the input data. More precisely, from a maximum-likelihood perspective, both these techniques aim at identifying a set of latent variables  $z$  and conditional probability distributions  $p(x|z)$  for the observed variable  $x$  representing users' whereabouts. The latent variables  $z$  are typically of a much lower dimensionality than  $x$ . Thus they encode patterns in a more understandable way with reduced noise. These techniques have been applied to a variety of people mobility datasets [7, 8, 4] with similar modalities.

In the context of the Reality Mining dataset (that is also the target of our work), the approach consists of extracting for each user and for each day a 24-slots array indicating where the user was at a given time of the day (24 hours). User's locations are expressed as either: 'Home', 'Work', 'Elsewhere' or 'No Signal', the latter indicating lack of data (in the remainder of this paper we refer to these locations respectively as 'H', 'W', 'E', 'N'). For example, a typical day of a user could be 'HHHHHHHHH-WWWWWWWWEEEEHHH' expressing the user was at home at night and early morning, then went to work until late afternoon, then went to somewhere else for three hours, and finally went back home (see next Section for further details on the dataset).

Applying PCA or LDA to a set of these arrays allows to extract some low-dimensions latent variables (eigenvectors and LDA-topics respectively) representing underlying patterns in the data, and offering conditional probability distributions for the observed arrays (i.e., days). Figure 1 illustrates some eigenvectors and LDA-topics extracted from the Reality Mining dataset.

Eigenvectors, leftmost part of Figure 1, encode the probability of the user being at a given location: 'H', 'W', 'E'. In the picture the lighter the color, the higher the probability.

Similarly, LDA-topics encode the probability of the user being at a given location (in a different representation format – see next Section for details). The rightmost part of Figure 1 shows the most probable days according to the con-



**Figure 1. (left) The top two eigenbehaviors for Subject 4 of the Reality Mining dataset. (right) Exemplary LDA-topics – images respectively taken from [7] and [8]**

ditional probability distribution of a given topic. These days are thus a representation of the topic itself.

Assigning a *meaning* to the extracted latent variables is a difficult task, that has been typically performed by visually inspecting the latent variable itself or the days that strongly correlate to the variable (i.e., most probable days given the latent variable) [7, 8, 4]. For example in [7], by visually inspecting the first eigenbehavior represented in Figure 1, authors conclude that it relates to the typical weekday behavior consisting of being at work from 10:00 to 20:00 and being at home in the remaining part of the day. The second eigenbehavior corresponds to typical weekend behavior where the subject remains at home past 10:00 in the morning and is out in the town (elsewhere) later that evening. Similarly, the rightmost part of Figure 1 reports some LDA-topics obtained in [8]. By visually inspecting the days strongly correlated to the extracted topics, authors conclude that topic 20 means “at home in the morning” while topic 46 means “at work in the afternoon until late in the evening”. Looking at these examples, it is clear how difficult it is to give a meaning to the extracted patterns and how difficult it is to evaluate the quality of the given meaning (i.e., label).

The need to associate meaningful labels to topics have been also considered in text-mining applications. The work presented in [16] tries to classify latent topics extracted from text corpora. Although this work applies to a completely different scenario, the fact that topic understanding is an important research challenge also in other communities further motivates our work.

**The contribution** of this paper is to present a methodology to automatically classify the extracted topics without

UserID	Begin	End	CellID
22	2004-08-27 14:00	2004-08-27 16:00	102
22	2004-08-27 16:30	2004-08-27 17:00	122
...	...	...	...

UserID	CellID	Label
22	102	Home
22	121	Work
...	...	...

**Figure 2. Tables used in the Reality Mining dataset.**

any visual inspection or user involvement. Once meaningful labels are given to the topics, the extracted pattern becomes readily understandable and usable in applications. For example, life-log applications [10] could readily use the extracted label to automatically create an entry in the user blog. Similarly, analyzing city-wide mobility patterns, applications could identify routine behaviors affecting city-life and communicate such information to local government and city planners. These tasks are simplified once a proper label is assigned to the discovered patterns, while they are very difficult starting from the extracted eigenbehaviors and topics in Figure 1.

The proposed classification methodology is a powerful tool in that it allows to express what is going to happen to the user (i.e., which topic is going to be expressed) with high-level meaningful labels.

Despite the fact that the presented approach is generalizable both to PCA and LDA, in the following of this paper we focus only on the LDA application. The probabilistic model realized by LDA is better suited at extracting different patterns from complex datasets [3, 8].

## 2 Data Preprocessing and LDA

In this section we present the data and algorithms representing the starting point of our work.

### 2.1 Data Preprocessing

The work presented in this paper is based on the GSM-localization part of the Reality Mining dataset. This dataset basically consists of two big tables (see Figure 2). For each user are recorded several time-frames and the GSM towers where the user was connected. Tables have missing data (time-frames in which no information has been logged) due to data corruption and powered-off devices. On average logs account for approximately 85% of the time that the data collection lasted. Another table records the labels

given by the users to the different GSM towers. Not all the towers are labeled. The dataset comprises 32628 GSM towers and only 825 are labeled (2.5%). Fortunately labeled cells are those in which users spend most of the time so overall 75% of the dataset happens to be in labeled cells. Still, identifying where the users have been in the remaining 25% of the time is an important issue to improve the data.

Although some works [8] try to extract patterns directly from such data (considering as “Elsewhere” all the unlabeled towers), we opted to run some preprocessing to complete missing values. In particular, we train a SVM to infer the labels of all the GSM towers. SVM computations have been performed with the LIBSVM library [5], on the basis of the following procedure:

1. We create training and testing set as:

Day of Week	Weekend	Hour	CellID	Label
Tuesday	no	14	150	Work
Saturday	yes	17	950	Home
Wednesday	no	15	155	?

The table associates the *label* to be identified to a feature vector consisting of the day of the week when the tower is visited, whether it is a weekend or not, the hour of the visit, and the cell ID.

2. We conduct a simple scaling on the data, converting all the values to a [0,1] scale. This is to avoid attributes in greater numeric ranges dominating the others.
3. Following other examples in the SVM literature [5], we consider Radial Basis Function (RBF) kernel since it well-applies to a vast range of classification problems. In addition, we use cross-validation to find the best parameters  $C$  and  $\gamma$  of SVM and RBF respectively. Basically we try all the combinations of the parameters with an exponentially-growing grid search. Parameters producing best cross-validation accuracy are selected for the final model.
4. We use the best parameters to train the SVM model on the whole training set and to classify the testing set.

SVM classification produces results with an overall 86% accuracy in cross validation (accuracy being defined as the proportion of correct results in the population). After SVM-classification, the dataset is much more representative of the *plausible* whereabouts of the users (missing groundtruth information, we can not make assertions on their *actual* whereabouts). For example, without SVM, a lot of days of the users are described by being always “Elsewhere” (i.e., not at home nor work). This is rather unrealistic and in fact SVM corrects this unbalance by restoring, for example, the being-at-home-at-night behavior.

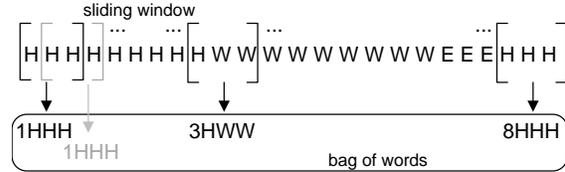


Figure 3. Sliding windows approach.

Following [7, 8], we organize the dataset into a sequence of days each consisting of 24 time-slots lasting 1 hour. Each time slot is labeled after the cell where the user spends most of his time. If no information are present for that time slot, the cell is marked as ‘No-Signal’.

To apply the LDA algorithm described in the following section, the dataset has been further processed. Each day is divided into a sequence of *words* each representing a 3 hours time-slot. A 3-hours sliding window runs across the day, each word is composed of an integer value in (1,8) (we refer to this value as *time-period*) and the 3 (‘H’, ‘W’, ‘E’ or ‘N’) labels in the sliding window. The *time-period* abstracts the time of the day, it is 1 if the sliding window starts in 0:00am - 3:00am, 2 in 3:00am - 6:00am, and so on (see Figure 3).

The fact of using *time-periods* of 3 hours each (in contrast with some previous work [8] in which different *time-periods* are skewed) improves the resulting dataset in that the number of words for each time slot is not biased by its length, but better reflects the actual user behavior.

The resulting bag of words summarizes the original dataset and is the input data structure for the LDA algorithm described in the next subsection.

## 2.2 LDA Algorithm

LDA is a probabilistic generative model [3] used to cluster documents according to the topics (i.e., word patterns) they contain. The work in [8] proposes using this model to extract mobility patterns from a mobility dataset.

LDA is based on the Bayesian network depicted in Figure 4. A word  $w$  is the basic unit of data, representing user location at a given *time-period* (see bag of words in Figure 3). A set of  $N$  words defines a day of the user. Each user has a dataset consisting of  $M$  documents. Each day is viewed as a mixture of topics  $z$ , where topics are distributions over words (i.e., each topic can be represented by the list of words associated to the probability  $p(w|z)$ ). For each day  $i$ , the probability of a word  $w_{ij}$  is given by  $p(w_{ij}) = \sum_{t=1}^T p(w_{ij}|z_{it})p(z_{it})$ , where  $T$  is the number of topics.  $p(w_{ij}|z_{it})$  and  $p(z_{it})$  are assumed to have Multinomial distributions with hyperparameters  $\alpha$  and  $\beta$  respectively. LDA uses the EM-algorithms [2] to learn the model

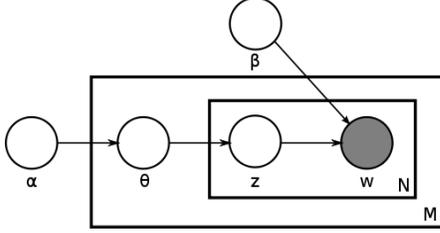


Figure 4. Plate notation of the LDA model.

parameters. In our implementation we use the library Mallet (<http://mallet.cs.umass.edu>) to perform these computations.

Once the model parameters have been found, Bayesian deduction allows to extract the topics best describing the routines of a given day (rank  $z$  on the basis of  $p(d|z)$ ). However, as already introduced, since  $z$  are just distributions over words, it is difficult to give them an immediate meaning useful in applications. The next section contains the main contribution of this paper: giving a meaningful label to topics  $z$ .

### 3 LDA-Topic Classification

#### 3.1 Method

In extreme summary, our approach consists of identifying a set of labels describing the main trends of a user typical day. For example, ‘Work 9:00 - 18:00’ represents a day in which the user is at work from 9:00 to 18:00. We then identify the LDA-topics representing such a label (we refer to these topics as *label-defined* topics). LDA-topics extracted from the Reality Mining dataset are labeled after the most similar *label-defined* topics. Thus, rather than being described only in terms of probability distributions over words, topics get a compact description like ‘Work 9:00 - 18:00’.

More in detail, our methodology is based on these key points:

1. We create a set of 30 predefined labels each composed of a place (‘H’, ‘W’, ‘E’ or ‘N’; we refer to this places as *pattern-label*) and a time-frame (we refer to it as *time-frame-label*). For example, the label ‘W 9:00 - 18:00’ represents the pattern where the user is at work from 9:00 to 18:00 while the label ‘H 12:00 - 14:00’ represents the pattern where the user is at home at lunch time. We choose these labels by visually inspecting the recurrent patterns in the Reality Mining users’ days.
2. For each predefined label, we create a set of 15 sample days representing the corresponding daily behavior

(each day is represented as a 24-slots array indicating where the user was at a given time of the day). The different days keep the pattern indicated by the label constant, and change the remaining part of the day.

3. For each block of 15 days, we compute one LDA-topic. The final result is a set of 30 *label-defined* topics, each representing one of the predefined labels. For example, recalling that topics are distributions over words, the days associated with the ‘W 9:00 - 18:00’ pattern will create a topic in which the  $p(w|z)$  of words like 3WWW, 4WWW, 5WWW will be high compared to other words probabilities. We verified that the resulting topics are not strongly affected if being produced by a number of days smaller or greater than 15.
4. Topics extracted from the Reality Mining dataset are classified using  $k$ -Nearest Neighbor (kNN) and the Kullback-Leibler divergence as distance metric from the above *label-defined* topics.
5. The final result is that days of a users can be described as easily-understandable labels (e.g., ‘W 9:00 - 18:00’), rather than with just probability distributions over words that are much more complex to be interpreted.

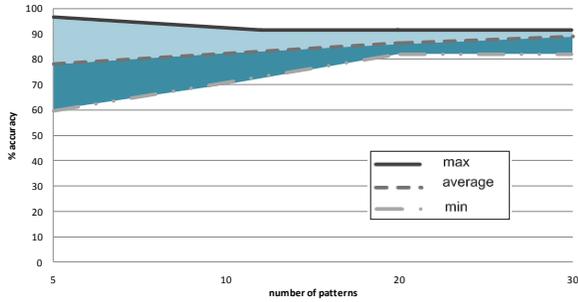
#### 3.2 Experiments and Discussion

We conduct some experiments to test the above approach. First, we experiment with an artificially-created dataset where we get groundtruth information, and thus classification accuracy can be precisely evaluated. Then, we test the system with the Reality Mining data that misses groundtruth information.

The first set of experiments studies classification accuracy. Starting from the above described 30 predefined labels expressing user patterns, we create a testing set on which to extract and classify topics. More in detail, we select  $L$  labels. For each label we create 15 days (following the same procedure described above) and we stack the  $15 \cdot L$  days together to create an artificial dataset of user’s whereabouts. The  $L$  labels represent the groundtruth user’s whereabouts patterns.

We extract  $L$  topics from this dataset and classify the topics with the kNN algorithm (in this experiment with just use  $k=1$ ). The expected result is to classify the extracted topics with the same labels used to create the dataset. For each experiment we compute classification accuracy as:

$$\frac{|\{\text{classified labels}\} \cap \{\text{groundtruth label}\}|}{|\{\text{classified labels}\}|}$$



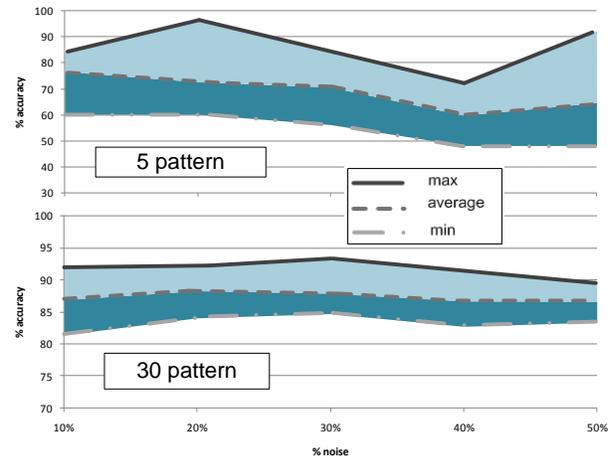
**Figure 5. Classification accuracy and number of patterns used to create the testing datasets. Minimum and maximum accuracies give a fair measure of the variance in our results.**

All the results are averaged over 100 runs of the experiment in which we generate random groundtruth topics and random days containing that topics.

Figure 5 shows the average classification accuracy as a function of the number of patterns used to create the testing dataset. The loss in the accuracy classification obtained with a little number of patterns is due to the fact that our algorithm tends to misclassify labels representing short *time-frame-label*. For example, if the predefined label is ‘H 12:00 - 14:00’, a testing day could be ‘EEEHHHH-HWWWHHHWWWWHHHHH’ which is better represented by a topic described by the label ‘W 9:00 - 18:00’ or ‘H 20:00 - 24:00’. This fact is also reflected by the higher variation between the minimum and maximum classification accuracy obtained with a little number of patterns.

To further test the approach in this experimental setting, we add artificial noise to the testing dataset. The idea is to corrupt the underlying pattern to see that what extent the LDA algorithm and our classification mechanism are able to generalize the pattern. In particular, once a day has been created, we change *noise%* of the labels in the pattern time-slots with random other labels. For example, if the label expressing the testing set days is ‘E 18:00 - 24:00’ a sample day for this experiment could be ‘HHHHHHHH-WWWWWWWWEEEEHHH’ expressing the user was somewhere else only until 22:00 and then went back home (*noise* = 3/7 = 42%).

Figure 6 shows the classification accuracy obtained at different noise levels in the case of datasets composed of 5 and 30 patterns respectively. As above mentioned, the variation between minimum and maximum classification accuracy is higher with a little number of patterns than with a lot of them. In addition, a higher number of pattern results less affected by the introduction of artificial noise in the testing



**Figure 6. Average classification accuracy as a function of the noise introduced. The x-axis represents the hours’ percentage (with respect to the time slot length indicated by the label) that has been randomly changed.**

dataset.

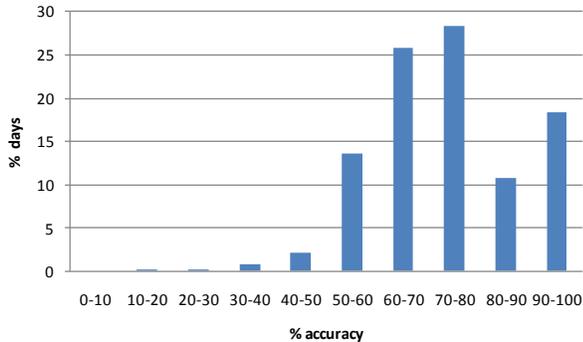
In a second group of experiments, we test our classification method on the Reality Mining dataset.

We experiment with 36 individuals and 121 consecutive days (from 26-08-2004 to 21-12-2004). We chose this subset of days with the goal of analyzing people and days for which the data is reasonably complete and with the goal of comparing our results with those presented in [8] taking into consideration the same subset of days. We also complete missing values with the SVM mechanism described in the previous section.

Since groundtruth information regarding user topics are not available, our experiments on the Reality Mining dataset focus on two main aspects:

1. We first evaluate whether the predefined labels associated to the user’s topics are a good representation of her days. In other words we verify whether the labels are informative enough to reconstruct the day of the user.
2. We evaluate if there are other labels describing user days better than the ones selected by our approach.

With regard to the former aspect, we extract 100 LDA-topics from all the days of each user taken into consideration. For each day  $d$  we rank the topic  $z$  according to  $p(d|z)$ . Starting from the topic  $z$  with higher  $p(d|z)$ , we reconstruct the day  $d$  according to the *pattern-label* and the *time-frame-label* associated to  $z$ . If a part of the day has



**Figure 7. Average day reconstruction accuracy computed over all users and days.**

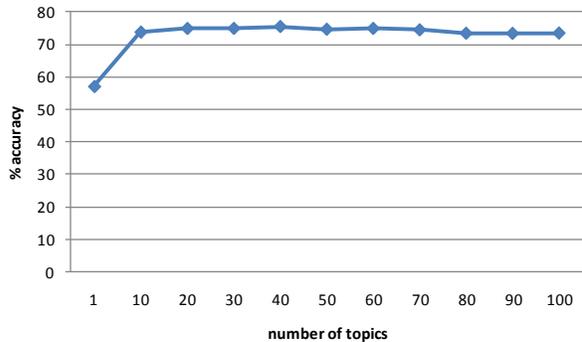
been already reconstructed by a previous (more probable) topic, the mechanisms just left it unchanged. Parts of the day that do not appear in the considered topics labels are not reconstructed.

We then compare the real and the reconstructed day. For each time-slot (hour) we assign an error equals to 1 if the reconstructed label is wrong. While an error equals to 0.5 if that hours is not reconstructed. The idea is that it is better for the algorithm not to reconstruct a part of the day rather than reconstructing it wrong. Figure 7 shows the distribution of days reconstruction accuracy: an average of 80% is obtained.

Figure 8 shows days reconstruction accuracy as a function of the number of LDA-topics extracted from users’ days. The low number of LDA-topics necessary to obtain high accuracy level can be explained by the limited number of users’ days available and by their repetitiveness.

With regard to the latter aspect of the Reality Mining experiments, we extract 100 LDA-topics from all the days of each user taken into consideration. For each day  $d$ , we want to find the topic that best describes that day. We rank topics  $z$  according to both  $p(d|z)$  and the length of the *time-frame-label* assigned to the topic. The idea is that topics explaining a bigger part of the user day are to be preferred.

The most probable topic  $z_{top}$  is selected for describing the day. We then evaluate how good the label assigned to  $z_{top}$  describes the day. In particular, for each time-slot in the *time-frame-label* we assign an error equals to 1 if the reconstructed label is wrong. For each time-slot that is not in the *time-frame-label* we give an error of 0.5. The idea is to lower the performance of topics associated to short *time-frame-label*, thus describing only a fraction of the day. Finally, we evaluate with the same error measure if there exists another label, better describing the day. We obtain that the labels associated to the selected topics are better than



**Figure 8. Average day reconstruction accuracy as a function of the number of topics extracted from the user’s dataset.**

any other label in the 80% of the cases.

All these results support the use of our classification mechanism. Our classification can notably simplify the comparison of users’ routines characteristic, in that it allows direct comparison between topics meaningful labels.

## 4 Related Work

The availability of affordable localization mechanisms and the recognition of location as a primary source of context information has stimulated a wealth of work. In particular, the core problem addressed in this paper: understanding users’ whereabouts.

### 4.1 Identifying Places

Several researches tackle the problem of understanding people whereabouts by trying to extract and identify those places that matter to the user. Mainstream approaches are either based on segmenting and clustering GPS-traces to infer what are the places relevant to the user [1, 18], or on detecting places and mobility on the basis of nearby RF-beacons such as WiFi and GSM towers [13, 7]. These approaches require the user to run a special software on her device to collect and analyze the log of GPS or RF-beacons available. Thus experiments with these mechanisms are usually conducted with a relatively small user population (the Reality Mining dataset used also in this paper is by far one of largest datasets in this category).

Starting from these results, another important area of research concerns the problem of converting from places described in terms of geographical coordinates or abstract IDs (e.g., GSM tower ID) to semantically-rich places such as: “home” or “favorite pub”. The work described in [1, 14]

adopt a probabilistic model to automatically attach semantic labels to visited places. These works rely on the fact that semantic information can be added by exploiting the structure of a person’s daily routine. For example, the place where the user usually spends the night can be tagged semantically as “home”, while the place where the user usually goes from 8:00 to 18:00 can be tagged as “work”. In [1] further information are extracted by geocoding the place and mining the Web in search for relevant information.

In summary, these approaches allow to represent and understand users’ mobility patterns as a sequence of places being visited at different times of the day. This representation resembles the “Home, Work, Elsewhere, No Signal”-representation used in the Reality Mining dataset.

Accordingly, while this paper focuses on higher-level abstractions (the sequence of places being visited is our starting point), the above approaches are the fundamental elements to apply our algorithms to other mobility datasets.

## 4.2 Identifying Routes

A number of related work deals with the problem of identifying the routes the user takes to move from one place to another. These approaches run data mining algorithms to identify recurrent patterns in the GPS tracks from multiple users. Works in this area can be divided in 2 broad categories: “geometric-based” approaches [9] apply pattern matching to the sequence of geographical coordinates composing the tracks. We call it geometric in that they use the physical “shape” of the path to compute the matching among routes. “String-based” approaches, instead, create a symbolic representation of the path (e.g., by considering only the names of the areas crosses by the path) and apply pattern-matching on that list of symbols [11]. In both cases, the extracted routes can be used to classify the user current and past whereabouts.

The work presented in this paper is similar to ‘string-based’ approaches, in that the geographic information about the places visited by the users are lost in favor of the more compact “Home, Work, Elsewhere, No Signal”-representation. However, there are two important differences. On the one hand, our representation is even more abstract than the one discussed in [11] and similar works. Labels in our dataset (e.g., Home) are completely detached from their physical location and, in fact, different users will label as “Home” completely different places. On the other hand our classification algorithm could extend also the above related work to obtain a more descriptive label for the extracted routes.

## 4.3 Identifying Routines

Even more high-level than extracting places and routes, is the problem of identifying routine whereabouts behaviors.

The CitySense project (<http://www.citysense.com>) is based on GPS and WiFi localization and has the goal of monitoring and describing the city’s nightlife. In particular, the application identifies hot-spots in the city (e.g., popular bars and clubs) and compares the number of people located in a given area in real time with past measures, to determine the “activity-level” of a given night. In a similar work based on extremely large anonymized mobility data coming from Telecom operators authors were able to extract the spatio-temporal dynamics of the city, highlighting where people usually go during the day. Authors were able also to identify the most visited areas by tourists during the day and the typical time of the visit [4, 12].

In this context the works that most directly compare to our proposal are [7, 8]. They use PCA and LDA algorithms to extract routine behavior from the Reality Mining dataset. Our work provides a further classification step to give meaningful labels to the extracted routines.

## 5 Conclusion and Future Work

In this paper we presented a methodology to automatically classify the routine whereabouts extracted from a large mobility dataset with meaningful labels.

Our future work in this area will target 3 main directions:

1. We will apply the presented approaches to “live” datasets such as those that can be acquired from Google Latitude (<http://www.google.com/latitude>) and Flickr [12].
2. We will develop mechanisms to add/modify topic labels at run time, so as to enable the use of the system in a wide range of scenarios.
3. We will develop Web-based visualization mechanisms to inspect and communicate whereabouts behaviors in an effective way [6].

The ultimate goal will be to create a real live Web application allowing different classes of users to see, understand and predict their own and other users’ whereabouts.

## 6 Acknowledgements

We thank Katayoun Farrahi for clarifications on their use of the LDA algorithm on the Reality Mining dataset.

## References

- [1] N. Bicocchi, G. Castelli, M. Mamei, A. Rosi, and F. Zambonelli. Supporting location-aware services for mobile users with the whereabouts diary. In *International Conference on MOBILE Wireless MiddleWARE, Operating Systems, and Applications*, Innsbruck, Austria, 2008.
- [2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.
- [3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022, 2003.
- [4] F. Calabrese, J. Reades, and C. Ratti. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *IEEE Pervasive Computing*, 9(1):78–84, 2010.
- [5] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [6] A. Clear, R. Shannon, T. Holland, A. Quigley, S. Dobson, and P. Nixon. Situvis: A visual tool for modeling a user’s behaviour patterns in a pervasive environment. In *International Conference on Pervasive Computing*, Nara, Japan, 2009.
- [7] N. Eagle and A. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.
- [8] K. Farrahi and D. Gatica-Perez. Learning and predicting multimodal daily life patterns from cell phones. In *International Conference on Multimodal Interfaces (ICMI-MLMI)*, Cambridge (MA), USA, 2009.
- [9] J. Froehlich and J. Krumm. Route prediction from trip observations. In *Intelligent Vehicle Initiative, Technology Advanced Controls and Navigation Systems, SAE World Congress and Exhibition*, Detroit (MI), USA, 2008.
- [10] J. Gemmell, G. Bell, and R. Lueder. Mylifebits: a personal database for everything. *Communications of the ACM*, 49(1):88–95, 2006.
- [11] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Trajectory pattern mining. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose (CA), USA, 2007.
- [12] F. Girardin, J. Blat, F. Calabrese, F. D. Fiore, and C. Ratti. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing*, 7(4):36–43, 2008.
- [13] D. Kim, J. Hightower, R. Govindan, and D. Estrin. Discovering semantically meaningful places from pervasive rf-beacons. In *International Conference on Ubiquitous Computing*, Orlando (FL), USA, 2009.
- [14] L. Liao, D. Fox, and H. Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *International Journal of Robotics Research*, 26(1):119–134, 2007.
- [15] M. Massimi, K. Truong, D. Dearman, and G. Hayes. Understanding recording technologies in everyday life. *IEEE Pervasive Computing, pre-print*, 2010.
- [16] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *ACM International Conference on Knowledge Discovery and Data Mining*, San Jose (CA), USA, 2007.
- [17] S. Patel, J. Kientz, G. Hayes, S. Bhat, and G. Abowd. Farther than you may think: An empirical investigation of the proximity of users to their mobile phones. In *International Conference on Ubiquitous Computing*, Orange County (CA), USA, 2006.
- [18] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories for mobile users. In *International World Wide Web Conference*, Madrid, Spain, 2009.