What Did You Do Today? Discovering Daily Routines from Large-Scale Mobile Data

Katayoun Farrahi and Daniel Gatica-Perez
Idiap research institute, Martigny, Switzerland
Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
{kfarrahi, gatica}@idiap.ch

ABSTRACT

We present a framework built from two Hierarchical Bayesian topic models to discover human location-driven routines from mobile phones. The framework uses location-driven bag representations of people's daily activities obtained from cell-tower connections. Using 68 000+ hours of real-life human data from the Reality Mining dataset, we successfully discover various types of routines. The first studied model, Latent Dirichlet Allocation (LDA), automatically discovers characteristic routines for all individuals in the study, including "going to work at 10am", "leaving work at night", or "staying home for the entire evening". In contrast, the second methodology with the Author Topic model (ATM) finds routines characteristic of a selected groups of users, such as "being at home in the mornings and evenings while being out in the afternoon", and ranks users by their probability of conforming to certain daily routines.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Human Factors

1. INTRODUCTION

Learning patterns of human behavior from large-scale sensor data is an emerging domain in ubiquitous and media computing aimed towards determining the behavior, habits, and activities of individuals in addition to the structure and dynamics of institutions [2, 5]. In particular, given the massive amount of data that can be captured by cell phones for many individuals over long durations of time, two key research questions are how to discover the emerging behavior of people (including habits and routines) over a long period, and how characteristic mobile sensor data (e.g. location extracted from cell tower information) is of people's routines.

The automatic discovery of people's daily routines is not a trivial problem given the often noisy and incomplete data that can be captured with a cell phone. In addition, the variations in a given person's activities across varying timescales, as well as the differences between many individuals' activities, complicates the task significantly. An unsupervised approach to human routine discovery has the potential of automatic discovery, not requiring training data, and is an ideal starting point for visualization of complex behavioral patterns within and across people and timescales.

In this paper, we develop a novel methodology built on Hierarchical Bayesian models to address the two questions above. Specifically, we use probabilistic topic models, initially designed for text documents [1, 7]. Recently, they have been successfully applied to data sources other than text, such as images [6], video, and genetics, but to our knowledge their use for real-life routine modeling from large-scale mobile phone data is novel. Topic models are generative models that represent documents as mixtures of topics, learned in a latent space, and they allow for clustering and ranking of documents, words, and other entities, like authors. They are advantageous to activity modeling tasks due to their ability to effectively characterize discrete data represented by bags (i.e. histograms of discrete items). These models learn which words are important to a topic as well as the prevalence of those topics within a document, resulting in a rank measure. The fact that multiple topics can be responsible for the words occurring in a single document discriminates this model from standard Bayesian classifiers. They are also useful as a time component can be incorporated into the bag representation. Further, we can take advantage of the bag flexibility to find routines at different temporal granularities. In this paper, we show that topic models prove to be effective in making sense of behavioral patterns at large-scale while filtering out the immense amount of noise in complex real-life data.

Our framework is used to automatically discover location-driven routines from the day in the life of a person without any supervision. The first contribution of this paper is the design of a methodology for the automatic discovery of daily routine patterns with Latent Dirichlet Allocation (LDA) where we discover routines characteristic of all days in the dataset and use this information to discover part of the underlying nature of individuals' life patterns. The second contribution is the extension of our methodology via the Author Topic model (ATM) to discover location-driven routines of a varying sort, this time emphasized on small groups of users' routines.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

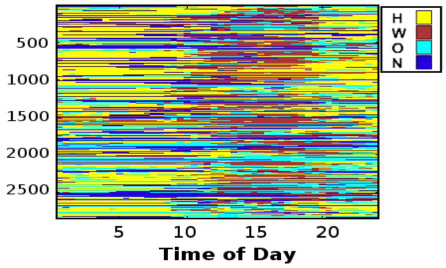


Figure 1: Fine-grain location visualized over all individuals' days (y-axis) in the study. The x axis is the time of day. The legend displays the home(H), work(W), other(O), and no reception(N) labels.

2. FRAMEWORK: ROUTINE DISCOVERY

Bag Representations. We use the Reality Mining dataset [3] for which the activities of 100 students and staff at MIT were recorded by Nokia 6600 smart phones over the 2004-2005 academic year. Given a day in the life of a person in terms of where they go, our goal is to discover real routines hidden in the enormous volume and complexity of information. We represent the day in the life of a person in terms of their locations obtained by cell tower connections, and implement a bag of location transitions with dynamic considerations.

Bag of Location Transitions. For a given individual, the dataset contains entries for each connected cell tower, as well as the start and end connection time. Over 32 000 towers are seen by all the people, covering a large geographical area. We classify the cell towers into 3 categories, HOME(H), WORK(W), and OTHER(O), representing towers which correspond to the homes of individuals, MIT work premises, and other towers, respectively. For missing data, we introduce a fourth label, NO RECEPTION(N), when there is no tower connection recorded for a person for a given time (e.g. no connection, no battery, or phone off).

A day in the life of a person can be expressed as a sequence of location labels (H,W,O,N). We begin by constructing a *fine-grain location* representation which is used to visualize the results and from which the bag of location transitions, described in the next paragraph, is constructed. We divide a day into 30-minute timeslots resulting in 48 blocks per day. For each block of time, we chose the single location label which occurred for the longest duration. The result is a day of a person represented as a vector of 48 location labels, visualized over all days and individuals in Figure 1.

The *bag of location transitions* is then built from the fine-grain location representation considering 8 coarse-grain timeslots in a day as follows: 0-7am, 7-9am, 9-11am, 11am-2pm, 2-5pm, 5-7pm, 7-9pm, and 9-12pm. The goal of these coarse-grain timeslots is to remove some of the potential noise due to minor time differences between daily routines (e.g. if a person leaves home at 7:30am as opposed to 8am, we want to capture the important feature of "leaving the house early in the morning").

A *location word* (in analogy with real words in the case of text bags) contains 3 consecutive location labels of the fine-grain representation (corresponding to 1.5 hour intervals) followed by the coarse-grain timeslot label in which it occurred. Thus a location word has 4 components. Location words are computed for each 30 minute period. The bag of location transitions is the histogram of the 48 location words present in a day. In this study, a document is a day of a user and an author (for ATM) is an individual in the study.

2.1 Topic Models for Routine Discovery

LDA [1] is a probabilistic, unsupervised learning model of a collection of bags and of hidden discrete variables called topics. With respect to text modelling, each document may be viewed as a mixture of various topics, where each topic is characterized by a distribution over words. The probability of a given word w_t assuming K topics and W unique words is given by: $P(w_t) = \sum_{k=1}^{K} P(w_t|z_t = k)P(z_t = k)$, where z_k is a latent variable indicating the topic from which the t^{th} word was drawn.

The objective of LDA inference is to determine the word distribution $P(w|z = k) = \phi_w^{(k)}$ for each topic k and the topic distribution $P(z = k) = \theta_k^{(d)}$ for each document d. In LDA, $P(\theta)$ is a Dirichlet(α) and $P(\phi)$ is a Dirichlet(β), where α and β are hyperparameters. The estimation problem in the LDA model is to maximize $P(w|\phi, \alpha) = \int P(w|\phi, \theta)P(\theta|\alpha)d\theta$, which is intractable. We use the approximation derived in [4] through Markov Chain Monte Carlo methods resulting in

$$\phi_k^{(w)} = \frac{n_k^{(w)} + \beta}{n_k^{(\cdot)} + W\beta},\ \theta_k^{(d)} = \frac{n_k^{(d)} + \alpha}{n_{\cdot}^{(d)} + K\alpha}, \tag{1}$$

where $n_k^{(w)}$ and $n_k^{(d)}$ are the number of times word w and document d have been assigned to topic k respectively.

The ATM [7] is built from LDA and assumes authors of documents represent a probability distribution over topics where each topic is a probability distribution over words. The probability distribution over topics in a multi-author paper is a mixture of the distributions associated with the authors. Again the estimation problem is intractable and we use the Gibbs approximation in [7] to find $P(w|z = k) = \phi_w^{(k)} \propto \phi_k^{(w)}$, which is the same as Equation 1. The distribution of topics for authors is $\theta_k^{(a)} = \frac{n_k^{(a)}+\alpha}{n_{\cdot}^{(a)}+K\alpha}$, where $n_k^{(a)}$ is the number of times author a has been assigned to topic k.

3. EXPERIMENTS AND RESULTS

From the Reality Mining dataset, we experimented with 30 individuals and 121 consecutive days (from 26.08.04 to 21.12.04). We chose this subset with the goal of analyzing people and days for which the data was reasonably available. Of the people selected, six were business students and the others were Media Lab undergraduate and graduate students and staff. We removed days which had NO RECEPTION the entire day since they contained no useful information. The resulting dataset is still huge, amounting to 2856 days over all people, and over 68 000 user-hours.

3.1 LDA-Based Routine Discovery

We apply our methodology to discover routines with $K = 30$ hidden topics (other values of K produced similar results). The LDA model successfully found latent topics which are mixtures over all users and are found to contain location-routines of people. A short video demo can be found at www.idiap.ch/~kfarrahi/LDADemo/topics.wmv.

In Table 1, we show the resulting top words for 3 topics. Topic 4 and Topic 8 characterize and contrast different work routines, whereas Topic 5 illustrates a going home routine. Topic 4 corresponds to "going to work in timeslot 2" and

Topic 4 - LDA		Topic 8 - LDA		Topic 5 - LDA	
Word	$P(w\|z)$	Word	$P(w\|z)$	Word	$P(w\|z)$
W W W 3	0.453	W W W 5	0.450	H H H 7	0.380
W W W 4	0.348	W W W 6	0.291	H H H 6	0.354
H H W 2	0.066	W W W 4	0.099	H H H 8	0.229
H W W 2	0.048	W H H 8	0.052	O H H 6	0.016
O W W 2	0.015	W W H 8	0.040	W O H 5	0.008
O O W 2	0.013	W W O 7	0.033	W O H 6	0.005
N N W 2	0.013	W O O 7	0.022	N O H 4	0.002
H O W 2	0.010	W H O 8	0.003	N O H 6	0.002
N O W 2	0.008	O W W 8	0.002	H N H 3	0.002

Table 1: LDA Results: Top location words ranked by $P(w|z)$ for topics 4, 8, and 5. Topic 4 captures patterns of "going to work" as well as "being at work in timeslot 3 and 4". Topic 8 illustrates "being at work in the afternoon". Topic 5 captures "being at home in the evening" and various patterns of "going home".

"being at work in timeslots 3 and 4". The two top words are working non-stop in timeslots 3 and 4. The following top words are various patterns that one would follow to arrive to work (e.g. HHW or HOW), all occurring in timeslot 2 (7-9am) with the final destination of work. Topic 8 resulted in "being at work in timeslots 4, 5, 6" followed by various patterns of "leaving work in timeslots 7 and 8". Topic 5 characterizes "being at home from 5pm on" as well as various patterns of getting home (e.g. OHH, WOH).

In Figure 2a, some of the location-driven routines found using LDA are visualized for the 50 top documents per topic, using the property $P(z|d) \propto P(d|z)$. We observe topics 10, 11, 4, and 8 characterize "being at work" patterns for various times of the day. Topics 5, 16, 24, and 30 are representative of "being at home" at various times. Topics 1, 26, and 29 capture "being in other locations", and topic 27 illustrates fluctuating between home and other locations before 10am, and work/other transitions throughout the day. The 30 topics obtained from the LDA experiments illustrate unique location-routines, however due to space constraints we have only selected 12 to display for discussion.

The topic distributions for 5 random weekends and weekdays (Figure 2b and c respectively) are visualized for users 7 and 19. We plot the most influential topics composing at least 50% of the days' activities. User 7's weekend routines are characterized best by topics 10, 11, and 24 which are "working before 10am", "working after 3pm" and "being at home midday" respectively. This user's weekdays are a mixture of several topics though topics 4, 8, and 11 dominate, corresponding to "working non-stop from 10am-roughly 6pm", "working non-stop in the afternoon-8pm" and "working until late in the evening", respectively. We see this user works a lot with work routines dominating most of the days. Further, the work patterns on weekends differ than those on weekdays, with a midday break for weekends only. User 19 seems to have less work routines than user 7, and works mostly on weekdays, as seen by topic 10 dominating the weekday patterns. User 19's weekends are dominated by topics 1, 26, 27, and 29 which contain "being in other locations" various times of the day, though topic 27 contains the "working with breaks" routine.

In Figure 3a, we plot a histogram of the number of topics composing over 50% of the probability mass of a day, over all days in the study. We can see that some days are very well characterized by a few topics (3-4), whereas other days require 10-11 topics to characterize their routines. On average, 7 topics can be used to describe the day in the

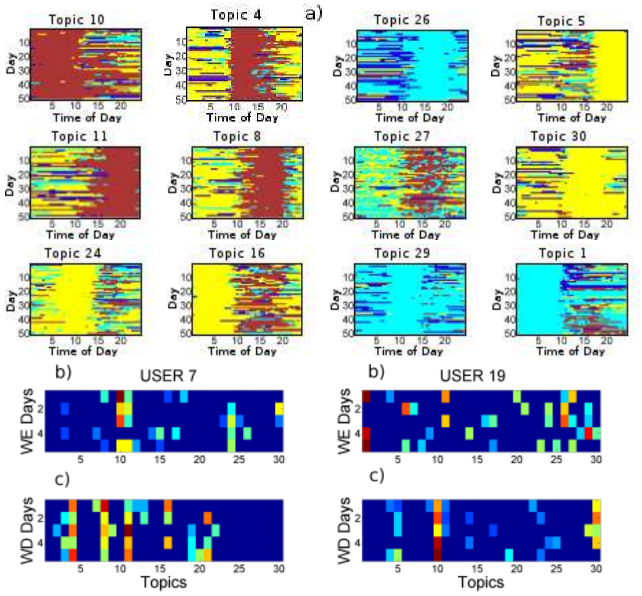


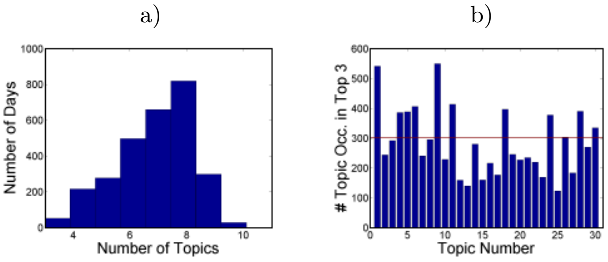
Figure 2: a) Some topics discovered from LDA are visualized (refer to the legend in Figure 1). b) Topic distributions for 5 weekends (WE) plotted for users 7 (left) and 19 (right). c) Topic distributions for 5 weekdays (WD) plotted for users 7 and 19. These plots illustrate the key topics (routines) composing selected days of a user. User 7's WEs are characterized best by the W routines of topics 10 and 11, though the WDs have different work routines, characterized by topics 4 and 8. User 19's WEs are characterized by "going out" routines and WDs are topic 10 (W routines) and topic 30 (H routines).

life of a user. As expected, increasing the threshold on the probability mass increases the number of topics. Figure 3b illustrates the number of topic occurrences for the top 3 topics of a day for all days and users. We can see the days are truly characterized by a mixture of topics. The most significant topics, which occur above the red line in 3b, have their corresponding routines described in the table beneath.

In work by Eagle and Pentland [3], which is the closest to ours, the structure in daily human behavior has been represented by principal component analysis (PCA), resulting in location-driven vectors termed eigenbehaviors. We propose a different framework for activity discovery based on two different topic models. Unlike PCA, topic models are probabilistic, and thus have advantages with respect to clustering and ranking days. Further, we have designed novel bag representations for routine discovery with more sophisticated data representations to consider location dynamics on both fine-grain and coarse-grain timescales.

3.2 ATM-Based Routine Discovery

For ATM, we apply our methodology to discover routines with $K = 30$ hidden topics in order to compare with results obtained in Section 3.1. The model returns $P(w|z)$ identifying the probability of words given topics, which is also obtained by LDA. In addition, ATM associates probabilities of authors given topics (from $P(a|z) \propto P(z|a)$) where an author is an individual in the study; we use this result to rank people in our dataset.



Routines for Highest Topic Occurrences	
Topic	**Routine**
1	OTHER in morning (before 10am)
4	WORK non-stop midday (9am-∼3pm)
5	HOME in evening (∼5pm-midnight)
6	OTHER in evening (∼5pm-midnight)
9	NO RECEPTION in morning (midnight-∼10am)
11	WORK all afternoon/evening (∼12-midnight)
18	NO RECEPTION midday (∼midnight-∼3pm)
24	HOME in morning (∼midnight-∼3pm)
28	NO RECEPTION in evening (∼3pm-midnight)
30	HOME almost all day

Figure 3: a) Histogram of the number of 'dominating' topics composing more than 50% of the probability mass of all the days in the study. b) The number of topic instances for the top 3 topics for each day in the study. The table at the bottom describes the routine type for the top topics (topics occurring above the red line in b).

Topic 21 - ATM		Topic 18 - ATM		Topic 20 - ATM	
Author	$P(w\|z)$	**Word**	$P(w\|z)$	**Author**	$P(w\|z)$
W W W 4	0.286	W W W 5	0.235	H H H 7	0.248
W W W 5	0.272	W W W 6	0.217	H H H 8	0.163
W W W 3	0.179	H H H 1	0.132	H H H 6	0.111
H H H 1	0.077	W W W 4	0.121	H H H 1	0.090
N N N 1	0.065	W W W 7	0.117	O H H 6	0.037
H H H 8	0.030	H H H 2	0.069	O O O 3	0.033
O W W 3	0.019	H H W 3	0.019	O W W 5	0.033

Topic 10 - ATM		Topic 12 - ATM		Topic 30 - ATM	
Word	$P(w\|z)$	**Word**	$P(w\|z)$	**Word**	$P(w\|z)$
W W W 1	0.358	H H H 1	0.430	W W W 3	0.204
W W W 8	0.129	W W W 4	0.146	H H W 2	0.057
W W W 7	0.113	W W W 5	0.125	H W W 2	0.054
W W W 6	0.072	W W W 6	0.080	N N N 1	0.050
W W W 4	0.062	H H H 2	0.042	W W H 7	0.047
W W W 2	0.052	H W W 3	0.030	W W H 4	0.042
W W W 3	0.039	H H H 8	0.029	W H H 8	0.041

Topic 10 - ATM		Topic 12 - ATM		Topic 30 - ATM	
Author	$P(z\|a)$	**Author**	$P(z\|a)$	**Author**	$P(z\|a)$
7 M	0.124	9 M	0.125	14 B	0.188
27 M	0.124	3 M	0.108	11 B	0.086
9 M	0.104	6 M	0.097	16 M	0.084
16 M	0.092	5 M	0.079	7 M	0.081

Table 2: Author Topic Model Results: Top location words ranked by $P(w|z)$ for selected topics. The bottom row displays the top ranked users by $P(z|a)$ for selected topics.

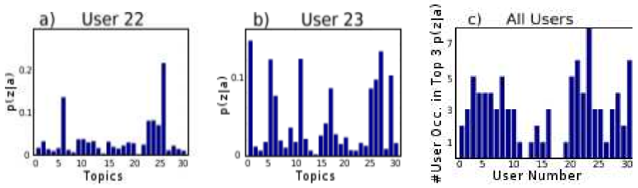


Figure 4: a-b) Topic distributions of users 22 and 23. We can see that user 22's location routines are primarily driven by 2 topics, whereas user 23's routines can be explained by a combination of several topics indicating perhaps a highly varying lifestyle. c) For every latent topic $P(z|a)$ distribution, we consider the top 3 users and plot them in a histogram. We can see that most of the individuals in the study can be depicted strongly by one of the 30 topics.

The first row in Table 2 shows $P(w|z)$ for topics 21, 18, and 20, corresponding closest to the LDA results in Table 1 to topics 4, 8, and 5, respectively. All of the topics display similar patterns, such as "being at work and leaving work", "arriving to work and being at work" as well as "going home and being at home", however, these routines seem noisier. The topics discovered with ATM contain routines which are more characteristic of selected users' routines. For example, users 7 and 27 work a lot, as seen in topic 10 with the top words all containing work routines. Users 9 and 3 work in timeslots 3-6, and are at home in timeslots 1 and 8. Users 14 and 11 go to work in timeslot 2, go home in timeslots 4, 7, 8 and are at work in slot 3. The users and their student types (B: business, M: Media Lab) are shown for topics 10, 12, and 30 in Table 2. We can see topic 30's top 2 users are business students and their routines contain less work than the Media Lab students in topic 10. Also the individuals whose documents are highly ranked for topic 12 (users 9, 3, 6 and 5) have more work routines than the top users of topic 30, however less than those of topic 10. The users ranked highly for topic 10 may live in work locations, explaining the work routines in timeslots 1 and 8.

The distribution of topics for users 22 and 23 is plotted in Figure 4a and b. We can see that user 22's location routines are primarily driven by 2 topics, whereas user 23's routines can be explained by a combination of several topics. User 22 likely lives a non-varying lifestyle in terms of location routines, explained well by topics 6 and 26 whereas user 23 likely lives a highly varying lifestyle. Further we can discover that most of the individuals in the study have been characterized well by the latent topics by plotting a histogram of the top 3 users for each latent topic in Figure 4c.

4. CONCLUSIONS

The results we have presented display human location-driven routines discovered using two Hierarchical Bayesian models from a massive dataset collected by mobile phones. We have proposed a methodology for location bag construction, and incorporated this into LDA and ATM. The resulting distributions of words for latent topics, as well as topics given days, and topics given users, reveal the successful discovery of routines and characteristic features of days as well as individuals in the study. In the future, we plan to design other topic-based models to discover other routine patterns (e.g. based on proximity information).

Acknowledgements This research has been supported by the Swiss National Science Foundation through MULTI. We thank Nathan Eagle (MIT) for sharing the data and helping with various aspects of the collection structure.

5. REFERENCES

[1] D. Blei, A. Ng and M. Jordan. "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3*, 2003.
[2] T. Choudhury and A. Pentland. "Sensing and Modeling Human Networks using the Sociometer," *Proc. ISWC*, 2003.
[3] N. Eagle and A. Pentland. "Eigenbehaviors: Identifying Structure in Routine," *Behavioral Ecology and Sociobiology (in submission)*, 2007.
[4] T.L. Griffiths and M. Steyvers. "Finding Scientific Topics," *PNAS 101:5228-5235*, 2004.
[5] L. Liao et al. "Location-Based Activity Recognition," *Proc. NIPS*, 2005.
[6] F. Monay and D. Gatica-Perez. "Modeling Semantic Aspects for Cross-Media Image Retrieval," *IEEE Trans. on PAMI*, 2007.
[7] M. Steyvers et al. "Probabilistic Author-Topic Models for Information Discovery," *Proc. KDD*, 2004.