Extracting Mobile Behavioral Patterns with the Distant N-Gram Topic Model

Katayoun Farrahi
JKU University Linz, Austria
farrahi@pervasive.jku.at

Daniel Gatica-Perez
Idiap Research Institute, Martigny, Switzerland
EPFL, Lausanne, Switzerland
gatica@idiap.ch

Abstract

Mining patterns of human behavior from large-scale mobile phone data has potential to understand certain phenomena in society. The study of such human-centric massive datasets requires new mathematical models. In this paper, we propose a probabilistic topic model that we call the distant n-gram topic model (DNTM) to address the problem of learning long duration human location sequences. The DNTM is based on Latent Dirichlet Allocation (LDA). We define the generative process for the model, derive the inference procedure and evaluate our model on real mobile data. We consider two different real-life human datasets, collected by mobile phone locations, the first considering GPS locations and the second considering cell tower connections. The DNTM successfully discovers topics on the two datasets. Finally, the DNTM is compared to LDA by considering log-likelihood performance on unseen data, showing the predictive power of the model on unseen data. We find that the DNTM consistantly outperforms LDA as the sequence length increases.

1. Introduction

As large scale mobile data on human behavior become more readily available, the need for effective methods and mathematical models for analysis becomes crucial. Research in Reality Mining [5, 7] has led to the need for the development of models that discover patterns over long and potentially varying durations. We address the problem of modeling activity sequences for large-scale human routine discovery from cellphone sensor data. Our objective is to handle sequences corresponding to human routines, based on principled procedures, and to apply them to generic human location data.

There are several difficulties to modeling human activities, including various types of uncertainty, lack of ground truth, complexity due to the size of the data, and the various types of phone users. The fundamental issue motivating this work is that we often do not know (or cannot pre-specify) the basic units of time for the activities in question. We do know that human routines have multiple timescales (hourly, daily etc.), however the effective modeling of multiple unknown time-durations is an open problem.

We focus on probabilistic topic models as the basic tool for routine analysis for several reasons. Topic models are, first and foremost, unsupervised in nature. Their probabilistic generative nature make them attractive over discriminative approaches since we are interested in mining the structure of the data. Topic models are also intuitive and provide opportunity for extensions with approximate methods for inference. They can handle large amounts of uncertainty due to the exchangeability of the bag of words property and process large amounts of data without major computational issues [15]. They can also be extended in various ways to integrate multiple data types [7].

The contributions of this paper are as follows: (1) we propose the distant n-gram topic model (DNTM) for sequence modeling; (2) we derive the inference process using Markov Chain Monte Carlo (MCMC) sampling [13]; (3) we apply the DNTM to two real large-scale datasets obtained by mobile phone location data. The model discovers user location routines over several hour time intervals, corresponding to sequences, and these results are illustrated by differing means; (4) we also perform a comparative analysis with Latent Dirichlet Allocation (LDA) [3], showing that the DNTM performs better in predicting unseen data based on log-likelihood values.

2 Probabilistic Topic Models

Probabilistic topic models were initially developed to manage large collections of text documents [3]. Recently, they have been found to be useful tools in the domain of activity modeling [10], particularly for mining wearable sensor data, such as location [7] and physical proximity data [1, 4, 8]. First we will describe the basic functionality of topic models in terms of text, and then introduce our approach for interpreting them in the context of human activity.

LDA [3] is a generative model in which each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words. By defining a Dirichlet prior on the document/topic (Θ) and word/topic (Φ) distributions, LDA provides a statistical foundation and a proper generative process. The main objective of the inference process is to determine the probability of each word given each topic, resulting in the matrix of parameters Φ, as well as to determine the probability of each topic given each document, resulting in Θ. Formally, the entity termed *word* is the basic unit of discrete data defined to be an item from a vocabulary. In the context of this paper, a word, later referred to as a label $\mathbf{w}$, is analogous to a person's location. A *document* is a sequence of words. In our case, a document is a day in the life of an individual. A *corpus* is a collection of M documents. In this paper, a corpus corresponds to the collection of sensor data to be mined. In the context of text, a *topic* can be thought of as a 'theme', whereas in our analogy, a topic can be interpreted as a human location routine.

Topic models have also been used for n-gram discovery, which can be seen as a method for variable sequence length discovery. The bigram topic model [16], the LDA collocation model [17], and the topical n-gram model [17] are all extensions of LDA to tackle this problem. The topical n-gram model is an extension to the LDA collocation model, and is more general than the bigram model. This approach was developed to be applied to text modeling and retains counts of bigram occurrences and thus could not easily be extended for large n (i.e. $n > 3$) due to parameter dimension explosion. Alternatively, dynamic topic models [2] model the change in the topic dynamics over time, and may be an alternative to model sequences with topics.

3 Distant N-Gram Topic Model

We introduce a new probabilistic generative model for sequence representation. The model is built on LDA, with the extension of generating sequences instead of single words as LDA does. The limiting criteria is to avoid parameter dimension explosion. We define a sequence to be a series of N consecutive labels or words. We represent a sequence as follows: $\mathbf{q} = (\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_N})$, where $\mathbf{w}$ denotes a label. In the context of this paper, a label $\mathbf{w}$ corresponds to a user's location obtained from a mobile phone sensor, though in general a label can correspond to any given feature in a series. The sequence $\mathbf{q}$ is then a sequence of locations occurring over an interval of time. The interval of time is defined by the duration over which each label occurs times the number of elements N in the sequence. The distant n-gram topic model (DNTM) generates a corpus of sequences. The maximum length of the sequence N is predefined. In existing n-gram models [17], a label in a sequence is assumed to be conditionally dependent on all previous labels in the sequence, thus making large sequences (longer than 3 labels) infeasible to manage due to an exponential number of dependencies as the sequence length grows. In contrast here, we integrate latent topics and assume a label in the sequence to be conditionally dependent only on the first element, the distance to this label, and the corresponding topic, removing the dependency on all other labels, and thus removing the exponential parameter growth rate.

The underlying concept and the novelty of our method is to obtain a distribution of topics given the first element in a sequence, represented by $\mathbf{\Phi_{1_z}}$. Then for each position j in the sequence, where $j > 1$, the distribution of topics given the j^{th} position in the sequence is obtained, depending on both the first element and the topic, represented by $\mathbf{\Phi_{j_{z,w_1}}}$. With this logic, our parameter size grows linearly with the sequence length N. Note that our approach for label dependency on $\mathbf{w_1}$ is the simplest case for which a label is always present. More advanced methods, including determining the number of previous labels for dependency are the subject of future work. We apply this model to location data to discover activities over large durations considering intervals of up to several hours. Next we define the generative process and introduce the learning and inference procedure. More derivation details can be seen in the Appendix, and the full derivation can be found in [6] where our model was referred to with a slightly different acronym.

3.1 The Probabilistic Model

Table 1. Symbol description

N	The length of the sequence
$\mathbf{q}$	A sequence of N consecutive labels ($\mathbf{w_1}$, ..., $\mathbf{w_N}$)
m	An instance of a document (a day here)
S_m	The total number of sequences $\mathbf{q}$ in document m
M	The number of documents in the corpus
T	The number of latent topics
z	A latent topic (a location routine here)
V	The vocabulary size
$\mathbf{\Theta}$	The distribution of topics given documents
$\mathbf{\Phi}$	The distribution of sequences given topics, where $\mathbf{\Phi} = \{\mathbf{\Phi_{1_z}}, \mathbf{\Phi_{2_{z,w_1}}}, ..., \mathbf{\Phi_{n_{z,w_1}}}\}$
$\mathbf{\Phi_{1_z}}$	The distribution of $\mathbf{w_1}$ given topics
$\mathbf{\Phi_{j_{z,w_1}}}$	The distribution of $\mathbf{w_j}$ given $\mathbf{w_1}$ and topics

The graphical model for our distant n-gram topic model is illustrated in Figure 1. We use a probabilistic approach where observations are represented by random variables, highlighted in gray. The latent variable z corresponds to a topic of activity sequences. The model parameters are defined in Table 1.

The generative process is defined as follows:

1. Initialization:

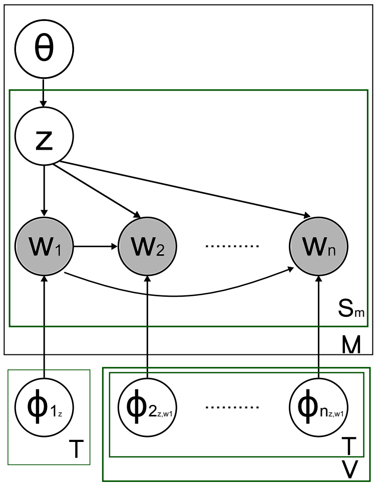


Figure 1. Graphical model of the Distant N-Gram Topic Model (DNTM). A sequence q is defined to be N consecutive locations $\mathrm{q} = (\mathrm{w_1}, \mathrm{w_2}, ..., \mathrm{w_N})$. Latent topics, z, are inferred by the model and can be interpreted as the different routines found to dominate the sensor data. There are M days (or documents) in the dataset. Θ is a distribution of days given the routines, and Φ_j is a distribution of location sequences given routines.

- **(1)** For each document m in the corpus draw a distribution over topics $\boldsymbol{\theta_m} \sim$ Dirichlet(α).
- **(2)** For each document m in the corpus:
  - **(2.1)** For each sequence **q** in document m:
    - **(2.1.1)** Draw a distribution over labels $\mathbf{\Phi_{1_z}} \sim$ Dirichlet(β_1) for the first element in the sequence.
    - **(2.1.2)** For each consecutive label $\mathbf{w_j}$ in the sequence:

      Draw a distribution over labels $\mathbf{\Phi_{j_{z,w_1}}}$ $\sim$ Dirichlet(β_j). Here $\mathbf{\Phi_{j_{z,w_1}}}$ captures the dependency with $\mathbf{z}$, $\mathbf{w_1}$, as well as the distance from the first label. Note the sequence length is defined by the user and is fixed.

2. Sequence generation procedure.

- **(1)** For each document m in the corpus:
  - **(1.1)** For each sequence **q** of the S_m sequences in document m:
    - **(1.1.1)** Draw a topic $\mathbf{z}$ $|m \sim$ Multinomial$(\theta_\mathbf{m})$.
    - **(1.1.2)** Draw the first label in the sequence $\mathbf{w_1}|\mathbf{z} \sim$ Multinomial$(\mathbf{\Phi_{1_z}})$.
    - **(1.1.3)** For $j = 2$ to N:

      Draw the j-th label in the sequence $\mathbf{w_j}|\mathbf{w_1}, \mathbf{z} \sim$ Multinomial$(\mathbf{\Phi_{j_{z,w_1}}})$ for $1 < j \leqq N$.

In summary, in the generative process for each sequence, the model first picks the topic $\mathbf{z}$ of the sequence and then generates all the labels in the sequence. The first label in the sequence is generated according to a multinomial distribution $\mathbf{\Phi_{1_z}}$, specific to the topic $\mathbf{z}$. The remaining labels in the sequence, $\mathbf{w_j}$ for $1 < j \leqq N$, are generated according to a multinomial $\boldsymbol{\Phi_{j_{z,w_1}}}$ specific to the current label position j, the topic $\mathbf{z}$ as well as the first label of the sequence $\mathbf{w_1}$. Note j is the j-th label in the sequence, but it can also be viewed as the distance between label j and 1.

We define the following notation; n_m^k is the number of occurrences of topic k in document m; $n_m = \{n_m^k\}_{k=1}^T$; $n_k^{w_1}$ is the number of occurrences of label w_1 in topic k, $n_k = \{n_k^t\}_{w_1=1}^V$; finally $n_{k^{`}_j}^{(w_1,w_2)_j}$ is the number of occurrences of label $w_2$ occurring $j$ labels after $w_1$ in topic $k$ and $n_{k^{`}_j} = \{n_{k^{`}_j}^{(w_1,w_2)_j}\}_{w_1=1,w_2=1}^{V,V}$.

We assume a Dirichlet prior distribution for $\mathbf{\Theta}$ and $\mathbf{\Phi} = \{\mathbf{\Phi_{1_z}}, \mathbf{\Phi_{2_{z,w_1}}}, ..., \mathbf{\Phi_{n_{z,w_1}}}\}$ with hyperparameters α and $\beta = \{\beta_1, \beta_2, ..., \beta_n\}$, respectively. We assume symmetric Dirichlet distributions with scalar parameters α and β such that $\alpha = \sum_{k=1}^T \frac{\alpha_k}{T}$, $\beta_1 = \sum_{v=1}^V \frac{\beta_{1,v}}{V}$, and $\beta_j = \sum_{w_1=1}^V \sum_{w_2=1}^V \frac{\beta_{(w_1,w_2)_j}}{V^2}$ for $1 < j \leqq N$. Note the parameters α_k, $\beta_{1,v}$, and $\beta_{(w_1,w_2)_j}$ are the components of the hyperparameters α, β_1, and β_j, respectively in the case of non-symmetric Dirichlet distributions.

Like LDA, the optimal estimation of model parameters is intractable. The model parameters are derived based on the MCMC approach of Gibbs sampling [9]. The model parameters can then be estimated as follows [6]:

$$\boldsymbol{\theta}_m^k = \frac{n_m^k + \alpha}{\sum_{k=1}^T (n_m^k + \alpha)} \quad (1)$$

$$\boldsymbol{\phi}_{1,k}^t = \frac{n_k^t + \beta_1}{\sum_{w_1=1}^V (n_k^{w_1} + \beta_1)} \quad (2)$$

$$\boldsymbol{\phi}_{j,k}^{(w_1,w_2)_j} = \frac{n_k^{(w_1,w_2)_j} + \beta_j}{\sum_{w_1=1}^V \sum_{w_2=1}^V (n_k^{(w_1,w_2)_j} + \beta_j)} \quad (3)$$

4 Data and Pre-Processing

The DNTM can be potentially applied to any type of data with discrete valued labels in a sequence, for example text, video, or mobile sensor data. We are interested in mobile location data over time. As stated in Section 2, we make an analogy with LDA where a document is an interval of time in a person's daily life. Here we always consider a document to be a day in the life of a user. A label $\mathbf{w} = (t, l)$ is composed of a location $l \in L$ which occurred over a 30 minute interval and a time coordinate of the day $t \in T = \{1, 2, 3, ..., tt\}$. We consider two different datasets for experiments. The representations for each are detailed below.

4.1 Nokia Smartphone Data

We use real life data from 2 volunteers using a Nokia N95 smart-phone from 2009.10.01 to 2010.07.01 corresponding to a 9 month period of the Lausanne Data Collection Campaign [11]. Users live in two different small cities. The phone has an application that collects location data on a quasi-continuous basis using a combination of GPS and WiFi sensing, along with a method to reduce battery consumption. Place extraction was performed using the algorithm proposed in [14], that reported good performance on similar data. For data representation, we create $\mathbf{w}$ where $tt = 8$, (i.e., the day is divided into 8 equivalent time intervals), $L = \{l_0, l_1, l_2, ...l_{MAX}\}$, where MAX is the number of detected places determined by [14] and l_i is the user-specific index of the place. If $l_i = 0$, there is no detected place, either due to no location being sensed, or due to the user moving or not staying at the location for very long. All places $l_i > 0$, are indexed according to their frequency of occurrence. Note that each user has a differing set of places and for this data collection topics are discovered on an individual basis. For user 1, $MAX = 101$ places and for user 2, $MAX = 108$ places, which gives an idea of the diversity of the location patterns of these users.

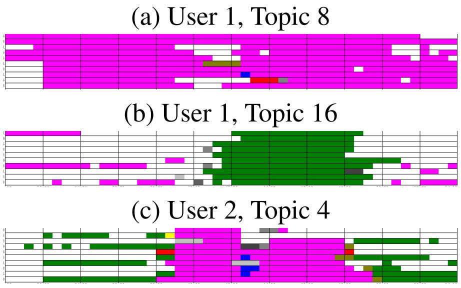
4.2 MIT Reality Mining (RM) Data

The MIT RM data collected by Eagle and Pentland [5] contains the data of 97 users over 16 months in 2004-2005. This data contains no detailed location information, but we define four possible location categories for a user collected via cell tower connections. The towers are labeled as 'home', 'work', 'out', or 'no reception', making the labels consistent over all the users. This corresponds to $L = \{'H', 'W', 'O', 'N'\}$. For this we set $tt = 48$.

5 Experiments and Results

5.1 Nokia Smartphone Data

For experiments with the smartphone data, we remove days which do not have at least one place detected. The results shown here are for $T = 25$, $\beta_j = 0.1$, $1 \leqq j \leqq N$ and $\alpha = 0.1$ selected heuristically. We consider $N = 12$ corresponding to six-hour sequences for the topics displayed here. Note that a range of values of T give similar results, the difference being that when T is small, the overall most occurring topics are discovered, and when T is larger, more activities are found. The constraint on the hyperparameters β_j and α are that they be smaller than the order of label/topic and document/topic counts.

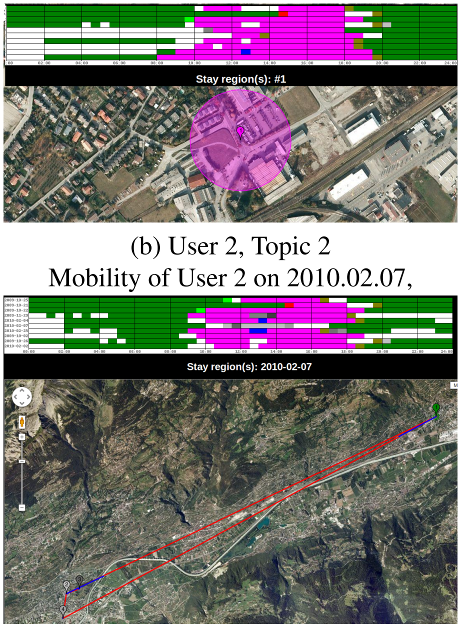

(a) User 1, Topic 8

(b) User 1, Topic 16

(c) User 2, Topic 4

Figure 2. Topics discovered for N=12. The x-axis is the time of day, the y-axis are the 10 most probable days for the topic ranked from top to bottom (output as Θ by the DNTM). Each unique colour represents a unique place. Our model discovers sequences of locations which dominantly co-occur in a user's mobility patterns. For example, topic 8 for user 1 corresponds to being at home (magenta) throughout the day. Topic 16 for user 1 corresponds to being at work (green) for several hours in the afternoon.

Several of the topics discovered by the DNTM for the smartphone data are shown in Figures 2 and 3. The first parameter the model returns is Θ, containing a probability distribution of each day in the corpus for each topic. We rank these probabilities for each topic and visualize the 10 most probable days, illustrating which days in the data had the highest probability of the location sequences for the given topic. In Figure 2, the three figures illustrate the 10 most probable days (i.e. max(θ_m^k) for a given topic k). The x-axis corresponds to the time of day, the y-axis corresponds to days, and each unique colour corresponds to a unique place. For both users, magenta corresponds to home, green to work, etc. White indicates that no place was observed during that time interval. We can see that sequences of places occurring over particular intervals of the day are discovered by the model. For example, topic 8 for user 1 corresponds to place 1 (home in magenta) occurring over most of the day. In Figure 3 we visualize details in geographic terms. In Figure 3(a) we show topic 19 of user 2. We also visualize the GPS coordinates of the place as displayed below the topic. The circle indicates the location of place 1 on a satellite map view. We display the satellite view for anonymity. In Figure 3(b) we show topic 2 for user 2.



(a) User 2, Topic 19
Satellite view of place 1 (magenta)
Note Satellite view displayed for anonymity

(b) User 2, Topic 2
Mobility of User 2 on 2010.02.07,

Figure 3. Topics and location details for user 2. (a) The satellite view of place 1 is displayed. (b) The mobility for day 2010.02.07 is displayed. The colours of the places displayed on the map correspond to those displayed in the topic. Note 2010.02.07 is one of the 10 most probable days for user 2 discovered in topic 2 and involved transitions between 3 places.

Below the topic we display the mobility traces for the day 2010.02.07, which is one of the 10 most probable days for topic 2. On the satellite view, each colour corresponds to a unique location, coordinated with the colour scheme of the topic displayed.

5.2 MIT RM Data

For experiments with the MIT dataset, we remove days which contain entirely no reception (N) labels. We experimented with many values of T and plot selected results for $T = 20$. We plot results for the same values of α and β as in Section 5.1. We consider up to $N = 14$ corresponding to seven-hour sequences.

We first visualize a set of 6 topics corresponding to activity sequences for various N. Note the location colorbar. Figure 4 corresponds to dominant sequences discovered for $N = 3$ (Figure 4 (a)-(c)), and $N = 13$ (Figure 4 (d)-(f)). We plot the results in terms of the 20 most probable days given topics, θ_m^k. The x-axis of the figures corresponds to the time of the day, the y-axis are days, and the legend of the colours are shown to the right of the plots. In general, we can see emerging location patterns discovered for specific subsets of days in the corpus. For example, in Figure 4 (a) there is 'N' (no reception) in the morning. In (b) there is 'W' (work) after roughly 10 am, with 'O' (out) several hours later, followed by 'W' again. These results resemble the type of results that standard LDA would extract, however, we are able to obtain precise sequence information in our output and "push" the model to output sequences by searching for results at distance d from the first label in the sequence. As N increases, we generally discover longer duration location patterns, which are defined in the output parameters of the DNTM model as shown in Tables 2 and 3. Note these tables show the sequences that defined the topics displayed in Figure 4.

In Table 2, we display the DNTM results in terms of the most probable sequence components given topics. The table shows the model output for $N = 3$, where the sequence is as follows $q = (w_1, w_2, w_3)$. The top ranked sequence components given topics k are displayed: $w_2|w_1$ obtained by $\phi_{2,k}^{(w_1,w_2)_2}$ and $w_3|w_1$ obtained by $\phi_{3,k}^{(w_1,w_2)_3}$ along with their probabilities. We do not display w_1 obtained by $\phi_{1,k}^{w_1}$ since it is inherent in the previous two parameters. We can see the sequence O-O-O starting at 8 pm is discovered in (a) for topic 3 ($N = 3$). The notation '*' represents any possible location, i.e. O-*-H indicates that $w_1 = O$, $w_3 = H$, with any possible location label for w_2.

In Table 3 we show the two most probable sequences for the topics displayed in Figure 4(d)-(f). Here, due to the larger value of $N = 13$, the actual sequences $\mathbf{q}$ are displayed. For large N, we can observe that some of the sequences output are separated in time, for example sequence 2 in (a) $N = 13$ topic 2. Since we do not force the output to always be a sequence of length N, there may be more than one sequence of duration less than N output by the model where the sum of the durations of the sequences output results in N. Constraints could be imposed to always force length N sequence as output, though the relaxation of this dependency in our model can be viewed as an advantage. We may in fact be discovering the durations of the dominantly co-occurring sequences. This characteristic is further discussed in the limitations section of the paper. We can see the output obtained by our model contains sequence information, since we obtain probabilities for the labels j up to distance N whereas LDA would simply output a probability for each individual label, without any sequence information.

In Figure 5, we plot the perplexity of the DNTM over varying number of topics computed on 20% unseen test data. Note, perplexity is a measure in text modeling of the ability of a model to generalize to unseen data; it is defined as the reciprocal geometric mean of the likelihood of a test corpus given a model. The experiments are conducted for

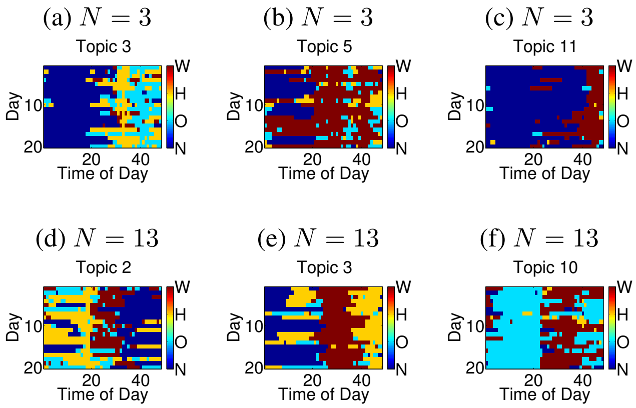


Figure 4. Topics discovered using our model with N=3, N=13. We plot the results in terms of the 20 most probable days given topics. In general, we can see emerging location patterns discovered within subsets of days in the corpus.

(a) $N = 3$, Topic 3

$w_2\|w_1$	$p(w_2\|w_1)$	$w_3\|w_1$	$p(w_3\|w_1)$
8 pm O-O	0.25	8 pm O-*-O	0.23
5 am N-N	0.21	5 am N-*-N	0.21

(b) $N = 3$, Topic 5

$w_2\|w_1$	$p(w_2\|w_1)$	$w_3\|w_1$	$p(w_3\|w_1)$
3:30 pm W-W	0.15	3:30 pm W-*-W	0.14
1:30 pm W-W	0.13	1:30 pm W-*-W	0.12

(c) $N = 3$, Topic 11

$w_2\|w_1$	$p(w_2\|w_1)$	$w_3\|w_1$	$p(w_3\|w_1)$
12:30 pm W-W	0.16	12:30 pm W-*-W	0.15
5:30 am N-N	0.14	5:30 am N-*-N	0.14

Table 2. Topics discovered using the DNTM corresponding to those displayed in Figure 4, expressed in terms of the most probable sequence components for topics. We show the top ranked sequence components given topics with the probabilities.

a sequence length of $N = 8$. We can see the perplexity drops to a minimum at around $T = 50$ topics. We therefore use $T = 50$ topics in order to compare the performance of our model to LDA. The perplexity results illustrate that for a large number of topics, T, the model does not seem to overfit the data, since the perplexity does not increase, but remains stable.

Table 3. Continuation of Table 2. The results in this table are for N=13 displayed as the sequence q.

(a) $N = 13$, Topic 2

Sequence 1	9 am	H-H-H-H-H-H-H-W-W-W
Sequence 2	5 pm	N-N-N-N-N
Sequence 2	9 am	H-*-*-*-*-W-W-W-W-W

(b) $N = 13$, Topic 3

Sequence 1	3 pm	W-W-W-W-W-W-W
Sequence 1	1:30 pm	W-*-*-*-*-*-*-W-W-W-W-W
Sequence 1	4:30 am	O-*-*-*-*-*-*-*-*-*-*-*-O
Sequence 2	1:30 pm	W-W-W-W-W-W-*-*-*-*-*-W
Sequence 2	3 pm	W-*-*-*-*-*-W-W-W-W
Sequence 2	4:30 am	O-*-*-*-*-*-*-*-*-*-O-O

(c) $N = 13$, Topic 10

Sequence 1	4 pm	W-W-W-W-W-W
Sequence 1	4 am	O-*-*-*-*-*-O-O-O-O-O-O-O
Sequence 2	4 am	O-O-O-O-O-O
Sequence 2	4 pm	W-*-*-*-*-*-W
Sequence 2	5 am	O-*-*-*-*-*-*-O-O-O-O-O-O

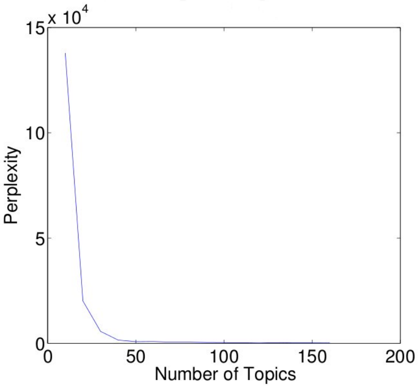


Figure 5. Perplexity of the DNTM over the number of topics on 20% unseen days (documents).

In order to compare our DNTM to LDA, we adapt the vocabulary used for LDA to have a comparable format to that used in the DNTM. The vocabulary we use for LDA consists of a pair of locations, a timeslot, as well as the distance between the locations. This results in a competitive comparison since the key attributes of the DNTM are taken into the vocabulary for LDA. The log-likelihood results on 20% unseen test data, are plotted in Figure 6. We plot the log-likelihood, averaged over all the test documents. The log-likelihood results reveal that for small N, LDA performs

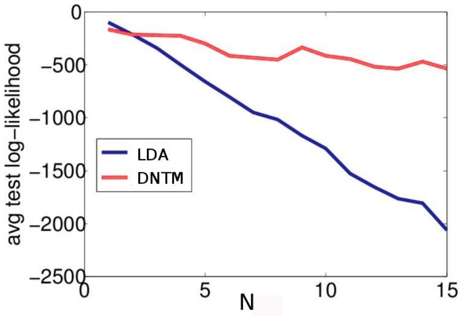


Figure 6. Average loglikelihood of the DNTM versus LDA on 20% unseen days (documents).

slightly better. However, as N increases, the DNTM consistantly has better generalization performance.

5.3 Discussion

Though only selected results are presented for the discussion here, most topics generally correspond to human routines. There are topics corresponding to noise, though this is interesting in itself and does not dominate the extracted routines.

One evaluation criteria in determining the quality of a model is its predictive power. In order to run experiments in which the predictive power of the DNTM could be computed, a prediction or classification task would need to be defined, and the performance of the model prediction would be very much dependent on the task at hand. In Section 5.2 we considered the average loglikelihood of the model on previously unseen data. This is a very general measure giving insight into the predictive capabilities of the model for data that was not previously seen by the model, and the results from Figure 6 are promising for the DNTM.

There are two main limitations of our model. The first one is that there is no constraint forcing the output components to be in sequence. More specifically, a valid output could be $w_2|w_1, z$ and $w_3|w_{1'}, z$ where $w_1 \neq w_{1'}$. In our experiments, we found that this effect did not occur often in the output. This can also be an advantage in that the output generates varying length sequences and determines the actual sequence lengths of the activities since they may not necessarily be exactly N. We would have to add some constraints to the model in order to always force the output to be sequences of length N. Another potential limitation is that the output can contain overlapping components. For example, a valid sequence output for a topic may be 3:30 pm H-H and 3 pm H-*-H. Here, the sequence output is not of length 3. To address this problem, again, some constraints should be imposed regarding the time component in the feature construction.

6. Conclusions

We propose the distant n-gram topic model to model long sequences for activity modeling and apply it in the context of human location sequences. Considering two real life human datasets, collected via mobile phone location logs, we test our model firstly on locations obtained by smartphones based on GPS and wifi and secondly by cell tower location features. The patterns extracted by our model are meaningful. We evaluate our model against LDA considering log-likelihood performance on unseen data and find the DNTM outperforms LDA for most of the studied cases.

There are several future directions for this work. The first direction is to further improve the model. One could improve the DNTM by taking into account the limitations mentioned and imposing application-specific constraints. One can also further investigate the dependence problem and consider methods to model dependence among labels as opposed to always having the label dependent on the first element, though this could quickly lead to parameter size explosion. For example, there may be effective hierarchical methods for determining the number of previous labels that a given label in a sequence should depend on. The second direction of extensions would be to consider other types of data, for example in the context of other wearable data and activities. Finally, one other relevant line of work future work is a comparison of our method with Hidden Markov Models learned in an unsupervised setting, imposing structure to learn long-term sequential patterns.

Acknowledgements
This research was funded by the SNSF HAI project and the LS-CONTEXT project funded by Nokia Research. K. Farrahi also acknowledges the Socionical project and the Pervasive Computing Group at JKU, Linz. We thank Olivier Bornet (Idiap) for help with place data processing and visualization, Gianpaolo Perruci (Nokia) for insights on routine visualization, and Trinh-Minh-Tri Do (Idiap) for discussions on sequence modeling methods.

7. Appendix

From the graphical model in Figure 1, we can determine the following relationship:

$$\begin{aligned} p(\mathbf{z}, \mathbf{q}|\alpha, \beta) &= p(\mathbf{z}, \boldsymbol{w_1}, ..., \boldsymbol{w_N}|\alpha, \beta) \qquad (4) \\ &= p(\mathbf{z}|\alpha) p(\boldsymbol{w_1}|\mathbf{z}, \beta_1) \prod_{j=2}^{N} p(\boldsymbol{w_j}|\mathbf{z}, \boldsymbol{w_1}, \beta_j) \end{aligned}$$

The joint probability of observations and latent topics can be obtained by marginalizing over the hidden parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$. These relations are then used for inference and parameter estimation where $p(\mathbf{z}|\alpha)$, $p(\boldsymbol{w_1}|\mathbf{z},\beta_1)$, and $p(\boldsymbol{w_j}|\boldsymbol{w_1},\beta_j)$ are derived in [6] resulting in the following.

$$\begin{aligned} p(\mathbf{z}|\alpha) &= \prod_{m=1}^{M} \frac{B(n_m+\alpha)}{B(\alpha)} && (5) \\ p(\boldsymbol{w_1}|\mathbf{z},\beta_1) &= \prod_{k=1}^{T} \frac{B(n_k+\beta_1)}{B(\beta_1)} && (6) \\ p(\boldsymbol{w_j}|\boldsymbol{w_1},\mathbf{z},\beta_j) &= \prod_{k=1}^{T} \frac{B(n_{k^{‘}_j}+\beta_j)}{B(\beta_j)}, 1 < j \leqq n && (7) \end{aligned}$$

We then derive the model parameters based on the MCMC approach of Gibbs sampling [9].

$$p(z_i = k|\mathbf{z}_{-i},\mathbf{q},\alpha,\beta) = \frac{p(\mathbf{z},\mathbf{q}|\alpha,\beta)}{p(\mathbf{z}_{-i},\mathbf{q}|\alpha,\beta)} \tag{8}$$

using the knowledge $\mathbf{z}_{-i},$ or $\mathbf{w}_{x_{-i}}$ indicate that token i is excluded from the topic or label $\mathbf{w}_x$

$$\propto \quad (n^k_{m,-i}+\alpha) \cdot \frac{n^{w_1}_{k,-i}+\beta_1}{\sum_{w_1=1}^{V} n^t_{k,-i}+\beta_1} \cdot \tag{9}$$

$$\prod_{j=2}^{n} \frac{n^{(w_1,w_2)_j}_{k,-i}+\beta_j}{\sum_{w_1=1}^{V}\sum_{w_2=1}^{V} n^{(w_1,w_2)_j}_{k,-i}+\beta_j}$$

where $n^{(y)}_x = n^{(y)}_{x,-i}+1$ if $x = x_i$ and $y = y_i$ and $n^{(y)}_x = n^{(y)}_{x,-i}$ in other cases.

where $n_k = \{n^{w_1}_k\}^V_{w_1=1}$ and $n_{k'_j} = \{n^{(w_1,w_2)_j}_{k'}\}^{w_1=V,w_2=V}_{w_1=1,w_2=1}$.

The model parameters can then be estimated as follows:

$$\begin{aligned} \boldsymbol{\theta}^k_m &= \frac{n^k_m+\alpha}{\sum_{k=1}^{T}(n^k_m+\alpha)} && (10) \\ \boldsymbol{\phi}^t_{1,k} &= \frac{n^t_k+\beta_1}{\sum_{w_1=1}^{V}(n^{w_1}_k+\beta_1)} && (11) \\ \boldsymbol{\phi}^{(w_1,w_2)_j}_{j,k} &= \frac{n^{(w_1,w_2)_j}_k+\beta_j}{\sum_{w_1=1}^{V}\sum_{w_2=1}^{V}(n^{(w_1,w_2)_j}_k+\beta_j)} && (12) \end{aligned}$$

References

[1] T. Bao, H. Cao, E. Chen, J. Tian, and H. Xiong. An unsupervised approach to modeling personalized contexts of mobile users. In *IEEE International Conference on Data Mining (ICDM)*, pages 38–47, 2010.

[2] D. Blei and J. Lafferty. Dynamic topic models. In *Proc. of the 23rd International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, USA, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[4] T. Do and D. Gatica-Perez. Groupus: Smartphone proximity data and human interaction type mining. In *Proc. IEEE Int.Symp. on Wearable Computers (ISWC)*, San Francisco, USA, June 2011.

[5] N. Eagle and A. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[6] K. Farrahi. *A Probabilistic Approach to Socio-Geographic Reality Mining*. PhD thesis, Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland, 2011.

[7] K. Farrahi and D. Gatica-Perez. Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*, 4(4):746–755, 2010.

[8] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology, Special Issue on Intelligent Systems for Activity Recognition*, 2(1):3:1–3:27, January 2011.

[9] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS USA*, 101 Suppl 1:5228–5235, April 2004.

[10] T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In *Ubiquitous computing (UbiComp)*, pages 10–19, Seoul, Korea, 2008.

[11] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. ACM Int. Conf. on Pervasive Services (ICPS)*, Berlin, Germany, 2010.

[12] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 787–794, Vancouver, Canada, 2006.

[13] D. J. C. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[14] R. Montoliu and D. Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proc. ACM Int. Conf. on Mobile and Ubiquitous Multimedia (MUM)*, Limassol, Cypress, Dec. 2010.

[15] J. Petterson, A. J. Smola, T. S. Caetano, W. L. Buntine, and S. Narayanamurthy. Word features for latent dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1921–1929, 2010.

[16] H. Wallach. Topic modeling: beyond bag-of-words. In *Proc. of the International Conference on Machine Learning (ICML)*, Pittsburgh, USA, 2006.

[17] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *IEEE International Conference on Data Mining (ICDM)*, pages 697–702, Washington, USA, 2007.