IDIAP RESEARCH REPORT

MINING HUMAN LOCATION-ROUTINES USING A MULTI-LEVEL TOPIC MODEL

Katayoun Farrahi Daniel Gatica-Perez

Idiap-RR-28-2010

AUGUST 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Mining Human Location-Routines using a Multi-Level Approach to Topic Modeling

Katayoun Farrahi
IDIAP Research Institute, Martigny
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
Email: kfarrahi@idiap.ch

Daniel Gatica-Perez
IDIAP Research Institute, Martigny
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
Email: gatica@idiap.ch

***Abstract*—In this work we address the problem of modeling varying time duration sequences for large-scale human routine discovery from cellphone sensor data using a multi-level approach to probabilistic topic models. We use an unsupervised learning approach that discovers human routines of varying durations ranging from half-hourly to several hours. Our methodology can handle large sequence lengths based on a principled procedure to deal with potentially large routine-vocabulary sizes, and can be applied to rather naive initial vocabularies to discover meaningful location-routines. We successfully apply the model to a large, real-life dataset, consisting of 97 cellphone users and 16 months of their location patterns, to discover routines with varying time durations.**

I. Introduction

Recently, large datasets on human behaviour are being collected via mobile phones, potentially providing many insights on large-scale human communications, movements, and interactions. Reality Mining is a name coined for this data type [7]. The interpretation of such data, collected over hundreds, sometimes hundreds of thousands of users over many months, represents new challenges in ubiquitous computing due to the nature of the data and depending on the particular research task at hand. In this work, we consider the task of identifying human routines over time considering varying time durations in an unsupervised fashion. There are several difficulties to this problem including various types of "noise", lack of ground truth, and complexity due to the size of the data, the multiple types of data, and the various types of phone users. A fundamental issue in human activity modeling 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, however the effective modeling of many unknown time-durations is an open problem. Previous works always assume a fixed and predefined unit of time, limiting the timescale of routines discovered. In this work we propose a method to discover human routines considering varying time durations, built on probabilistic topic models, and demonstrate its feasibility on the Reality Mining data [7].

In several previous works addressing human activity modeling, a bag of words approach is used for activity discovery from video, wearable devices, or mobile phones [8], [12], [14]. The advantages of a bag approach are the robustness to noise, and the compact representation. The disadvantage with most of these works is that the words are not simple and time duration must be predefined by hand [8], [12] or a supervised method is used requiring a training phase [14]. Furthermore, *previously unknown* timescales, whether single or multiple, are not considered. Effective human activity modeling has a diverse range of applications, ranging from epidemiology to psychology. Human activities in terms of movement and interactions may provide insight to the spread of disease [11], urban planning, and human behaviour understanding such as what influences human opinion change [13], weight change [3], and the spread of happiness [9].

In this work we propose an elegant approach to construct vocabularies of increasingly large n-grams, built on a multi-level topic model. The number of consecutive words (or size of n in the n-gram) increases with each level of the topic model. Our approach is built on the probabilistic topic model, Latent Dirichlet Allocation (LDA) [2]. The maximum number of levels determines the maximum n-gram size considered, and large n-gram sizes can easily be considered with this method. At a given level n, the vocabulary consists of only n-grams. The output of the topic model at level n is used to construct the vocabulary for level $n + 1$. There are many advantages to our technique. Firstly, the vocabulary growth is controlled and does not explode since we only consider a subset of words for concatenation with the vocabulary in forming groups of consecutive words as opposed to the naive n-gram approach which grows exponentially with the size of the vocabulary. Secondly, not only do we consider the most frequently occurring words, but we can capture frequently co-occurring words which is critical in n-gram discovery. For example, if 2 words co-occur alot, they are more likely to form a bigram than if each occurs very frequently, but not together. Further, the initial vocabulary can be very simple, not requiring much hand-craft. Also, each individual level results in unique patterns with routines found for the particular vocabulary at that level. The contribution of this work is a unique approach to identify varying length human location routines and the successful application of this idea to a large-scale, real-life human dataset of over 10118 days and 242800 hours of data.

II. Related Work

Previous works have tackled the problem of discovering human activities from mobile phone location data. In [6], characteristic vectors of entire days are found, termed eigenbehaviors, representing the principal components of the dataset, resulting in routines found on the *fixed timescale* of a 24 hour period. In [8], [12], daily human routines are discovered with the use of topic models, though in [12] wearable sensor data is used. However, all of these works are based on fixed time duration words, resulting in "narrow" routine discovery due to inflexible vocabularies.

In text modeling [1], collocation has been tackled extensively, though usually trigram are the maximal n-gram considered. The simplest method is based on counting. In [15], word frequency is combined with linguistic knowledge to discover meaningful phrases. In [16], collocation discovery is based on variance. Also [4], [5] use hypothesis testing to assess whether or not two words occur together more often than by chance. These methods are of particular interest in text analysis, however, none of these methods would be relevent for large n-gram discovery.

Probabilistic topic models have been used for n-gram discovery. The bigram topic model [18], the LDA Collocation model [17], and the Topical n-gram Model [19] are all extensions of LDA to tackle this problem. The topical n-gram model [19], is an extension to the LDA Collocation model, and is more general than the bigram model. It approaches the n-gram sequence discovery problem by generating words in their textual order by, for each word, first sampling a topic, then sampling its status as a unigram or bigram, and then sampling the word from a topic-specific unigram or bigram distribution. This approach retains counts of bigram occurrences and thus could not easily be extended for very large n due to matrix size explosion.

III. Methodology

We use the Reality Mining dataset [7] containing the mobile phone sensor data of 97 subjects over the 2004-2005 academic year. We investigate the location data obtained by cell tower readings which have been semantically labeled by the users. We form very simple location words, capturing location and time information for a user, and use these as input to our multi-level topic model, which finds sequences of varying lengths as topic outputs in an unsupervised way. The components of our methodology are described in detail next.

A. Location Sequences as Words

We represent a day in the life of a mobile phone user's location as a bag of location sequences called location words, where an individual location word is constructed as follows. There are over 32000 cell towers recorded in the dataset, which based on semantic labels provided by the users themselves, and the data collection, can be converted into prototypical location labels: home (H), work (W), and out (O). A fourth semantic label, no reception (N), is also used for cases where the user's phone is off or has no reception. A day is divided into 48 half-hour time intervals, and each 30 minute interval is assigned a location label (the class that occurs the most during this 30 minute interval). The resulting features are 48 time intervals in the day and 48 location labels, one for each of the time intervals. A location word is a location interval in the range $\{1, 48\}$ and location label $= \{$'H', 'N', 'O', 'W'$\}$. This description of a location word is for a unigram and applies to level 1 of the model. The vocabularies for additional levels and n-grams for $n > 1$ are described in the following sections. A bag of location sequences or unigram location words consists of the 48 unigram location words in the day. The bag representation we are using is simple and does not require much handcraft of time intervals as in [8].

B. Latent Dirichlet Allocation

Topic models can be used to discover a set of underlying (or latent) topics from which a corpus of documents is composed via unsupervised learning. They are generative models initially invented for text, though have been used for other data such as video [14] and wearable sensor data [12]. The Latent Dirichlet Allocation (LDA) model [2] used in this paper is graphically illustrated in Figure 1.

In LDA, a word w is generated from a convex combination of topics, z, where the probability of term t is

$$p(w = t) = \sum_k p(w = t|z = k)p(z = k), \sum_k p(z = k) = 1. \tag{1}$$

A corpus is a collection of M documents d. The number of latent topics, K, must be chosen by the user. Words are considered to be exchangeable, meaning they are independent given topics.

The objective of LDA inference is to obtain (1) the term distribution $\phi_k = p(t|z = k)$ for each topic, and (2) the topic distribution $\theta_m = p(z|d = m)$ for each document, where $\Phi = \{\phi_k\}_{k=1}^K$ and $\Theta = \{\theta_m\}_{m=1}^M$. The distributions Φ and Θ are assumed to have Dirichlet priors with hyperparameters β and α, respectively.

When considering location data, location words can be seen as analogous to text words and a day in the life of a user is analogous to a document. Further, latent topics are analoguous to human routines, where Φ gives an indication of how probable topics are given days, and Θ results in a distribution of location words given topics.

C. A Multi-level LDA

We propose a multi-level approach, as illustrated in Figure 1, based on LDA, where the input vocabulary V is redefined at each level. The corpus C is input to the LDA model, consisting of M documents and a bag of words taken from a vocabulary of unigrams of size V_1. LDA inference at level 1 results in a ranking of words given topics at level 1, Φ_1, and topics given documents Θ_1. We concatenate the T most probable words given topics, Φ_1^T with the initial vocabulary V_1 to form bigrams as the new vocabulary to level 2, denoted V_2. The originality of our method lies in the formulation of the words, which are guided by the topic outputs. More

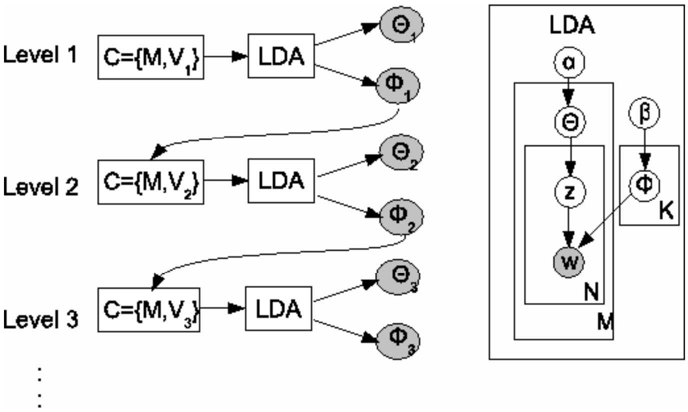

Fig. 1. Overview of our methodology. At level 1, the corpus C of M documents and words from vocabulary V_1 are input to the LDA model, whose graphical model is expanded on the right. Output from LDA at level 1, results in a ranking of words given topics Φ_1 and topics given documents Θ_1. We concatenate the top ranked Φ_1^T with the initial vocabulary V_1 to form bigrams as the new vocabulary to level 2, V_2. This process can continue for large n different levels, resulting in output sequences of length n from the LDA model.

generally, the T most probable words given topics, Φ_i^T at level i, are concatenated with the initial vocabulary of unigrams V_1 to form the new vocabulary for the next level of LDA V_{i+1}. Essentially, we are pruning and extending our vocabulary based on topic relevance.

Considering text, there are certain sequences of words that often belong together to give a particular meaning. For example, the expression "Markov Chain Monte Carlo" is a sequence of 4 consecutive words which has a different meaning than simply "Markov Chain". The problem is that we do not know how many consecutive words should belong together, and the number of consecutive words belonging together varies greatly from expression to expression. If we consider an example in terms of text related to a collection of machine learning papers, at level 1 possible word outputs for a topic related to Bayesian statistics could be "chain" or "Markov". Given the top words for topics at level 1, Φ_1^T, these are then concatenated with the original vocabulary. Effectively, this grows the vocabulary size by a factor of $T * K * V_{i-1}$ instead of V_{i-1}^2, which would be the case if we took all possible pairs of unigrams to form bigrams. For a large vocabulary, one could limit the new vocabulary to only bigrams contained in the corpus. Again LDA inference is performed with the bigram vocabulary, and this time Φ_2 (at level 2) will contain a distribution over bigrams given topics which are concatenated with V_1 to form a set of trigrams. If this is continued over several levels, for example "Markov Chain Monte Carlo" could be a potential output at the fourth level. This process can continue for many levels forming very large sequences since the input vocabulary size will not explode.

IV. Experiments and Results

A. Dataset

The Reality Mining dataset [7] collected by Eagle and Pentland at MIT, contains the recorded activities of 97 subjects (both engineering and business students and staff) over the 2004-2005 academic year. There are 491 consecutive days of data recording, which are from January 1, 2004 to May 5, 2005. All privacy concerns of the individuals in the study have been addressed by the collectors of the data. For experiments, we remove days which contain entirely N (no reception) labels since they do not provide any useful information. The resulting dataset remains very large, containing 10118 days and over 242800 hours of data. This amounts to just over 21% of the days containing at least a single location label. For the multi-level LDA model, we set the number of topics K_i to 50 for all levels i. The LDA hyperparameters are set to $\beta_i = 0.01$ and $\alpha_i = 50/T$ for all levels i. These hyperparameters are chosen based on standard values used for text analysis [10]. 12 levels are considered for experiments, resulting in routines discovered based on vocabularies ranging from a half-hour (at level 1) to six hour (at level 12) intervals.

B. Multi-Level Topics

The 50 topics at each level revealed human activities in terms of their locations for varying durations. The results are evaluated in terms of the most probable words for topics and by the top ranked days for the topic. We select several topics at various levels, and plot the 50 most highly ranked days in terms of their probability for a given topic, and list the most probable words given the topic in tables. In general, we observe over most topics, as the level increases, the routines discovered occur over longer durations (as expected) though this duration is not explicitly modeled. Further, as the level increases, the routines become more refined and discriminant over the day. These findings are explained in more detail in the sections that follow.

Figure 2 illustrates results seen at various levels. The plots show the top 50 ranked days for the selected topics. Each plot visualizes the locations as a function of the time of day (x-axis). Each row is a day in the life of a user, and the users can be any of the 97 in the study. The legend (in Figure 3) shows the location colour scheme, which is consistent throughout the paper. We pick some topics at level 4, which consisted of a vocabulary of 4-grams, occurring on 2 hour intervals. Looking at Figure 2 topic 40 at level 4, we can see a "home from midnight to approximately 3am" routine. Most of the days are also followed by a "no reception" for a few hours. Topic 44 at level 4 are days with a "work - out - work" sequence for a few hours around lunch, and topic 14 at level 4 are days with a work routine for approximately 5 hours in the late afternoon. Three topics are visualized at level 7 showing "out", "home", and "work" routines for several hours at various times of the day. The topics displayed for level 12 all show location sequences occurring for at least a 6 hour duration. The $6+$ hour routines discovered at level 12 are not discovered at lower levels. Topic 1 at level 12 shows an "out in the morning" routine occurring from midnight to 9am on a 9 hour interval. The out in the morning routine in topic 31 at level 7 occurs on a 5 hour interval.

Note since topics are discovered for most co-occurring

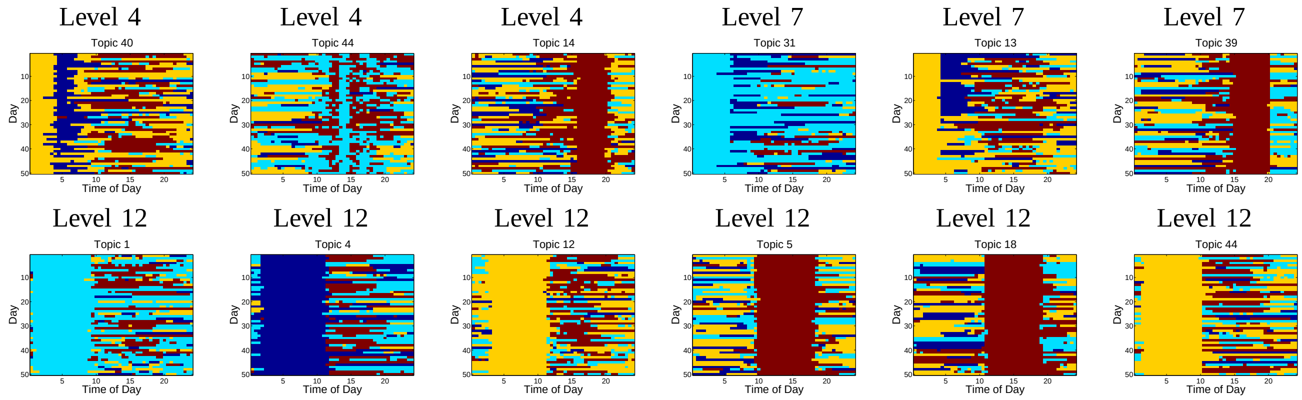

Fig. 2. The plots visualize the 50 most probable days for topics output at various levels of the multi-level model. In each plot, a row corresponds to the day in the life of an individual, where the x-axis corresponds to the time of day (in hours). The colour scheme is visualize in the legend in Figure 3. The first three figures labeled as "Level 4" visualize topics 40, 44, and 14 at level 4. The vocabulary at this level consists of 4-grams, corresponding to 2 hour location sequences, since a single word (or 1-gram) is a 30 minute location. The following 3 plots illustrate 3 topics output at level 7. The vocabulary at this level consists of 7-grams corresponding to 3.5 hour location sequences. The bottom row of 6 plots illustrate a set of topics output at level 12. The vocabulary at this level consists of 12-grams, occurring on 6 hour consecutive intervals. The routines obtained increase in time duration as the level increases. The routines at level 12 correspond to 6 hours of more of "being out in the morning" (topic 1), "having no reception in the morning" (topic 4), and so on. It is important to note that the routines discovered at level n do not necessarily correspond to routines of exactly $n/2$ hours since topic outputs correspond to highest *co-occurring* words.

words in a set of days and we do not explicitly model time durations in the topic model, the routines do not necessarily correspond to exactly $n/2$ hours of a routine at level n. However, each level contains a set of topics which are unique and can not be discovered at other levels due to vocabulary size constraints revealing varying dominant routines. For example, the "at home for 3-4 hours in the morning at level 7 (topic 13) never occurs at levels greater than 7.

C. Comparing Topics Across Levels

In Figure 3, we plot selected topics as well as an evolution of a similar topic after several levels. The plots in FIgure 3 show the top ranked days for the selected topics, and the table in Figure 4 shows the top ranked words for the same selected topics and levels (i.e. (a) corresponds to level 1 and topic 14 in both Figure 3 and 4. The tables (Figure 4) list the 2 most probable words for the topic at the specified level, as well as the probability of this word for the topic.

We observed some unique routines occurring at each level. We also found that the output over various levels sometimes revealed similar routines, with varying time durations. The results show that increasing the sequence length of the input vocabulary can result in routine disambiguation as well as more precisely "filtered" output. For instance consider Figure 3 plot (a) for topic 14 in level 1 characterized by the most probable words "Work 5:30-6pm" and "Work 6-6:30pm" and comparing this to plot (b) for topic 18 at level 12 characterized by words "Work 12-6pm" and "Work 12:30-6:30pm". We can see that at level 12 this work routine was found to occur on a larger time duration of 6 hours or more, whereas at level 1, it will occur with high probability (for highly ranked days) between 5:30-6:30, but not necessarily before and after this time interval. At level 12, the days with work routines non-stop between 12-6:30pm are "filtered" from the large number of days with varying types of work routines.

Considering plot (c) for topic 1 at level 1 and comparing it to plot (d) for topic 16 at level 10, both capture "home in the morning" routines as seen by the top words and the plot of the top 50 days. However, the routine captured by the level 10 vocabulary (plot (d)) shows a sharper transition before 9am since at this level longer location sequences are considered. The same is observed for the "home after 7pm" routine seen in plots (e) and (f). Again, a sharper transition is observed in the "at home" routine at level 4 than at level 1. The same phenomenon is also seen in plots (g) and (h). This is a nice characteristic of our method, which retrieves days with specific dominant patterns disambiguating general routines as the level increases. Essentially, we are finding more refined and meaningful routines over levels and the vocabulary becomes more discriminant as the level and the n-gram size increases. In terms of text, this could be seen as the word "Monte" co-occurring with many other words such as "Cristo", "Method", and "Lake" at level 1, but co-occurring with words related to "Monte Cristo" in one topic and "Monte Carlo Method" in another topic as the level increases, disambiguating the meanings of the word "Monte" with topics.

In Figure 5, we visualize the progression of 2 routines over levels. The results of levels $1-2$ and 6 also show other interesting routines following the trend displayed. The 2 rows in Figure 5 capture an "at work followed by out" routine, which is fluctuating at level 3 and 4, and appears to turn into a work-out-work routine, possibly corresponding to "lunch or break" at level 5, since the "out" occurs for a slightly longer duration of about 1-2 hours depending on the topic. Level 7 to level 9 routines still capture a "work followed by out" or "out-work-out" routine though now occurring for longer durations with varying patterns and time durations depending on the topic. As the level increases and the vocabulary size changes we recover days with new and different dominating sequences

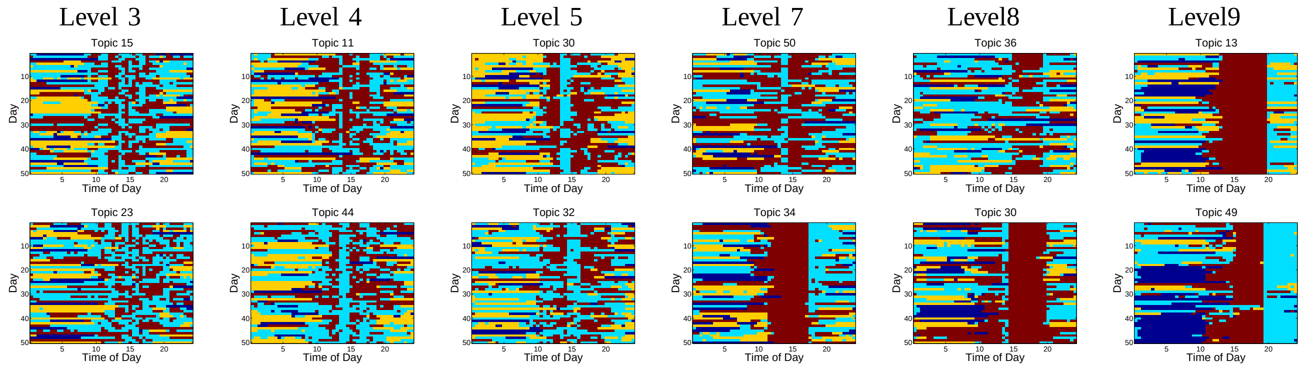

Fig. 5. Topic dynamics over levels. Each column is a different level, and each row is a similar routine which is changing as the vocabulary size changes to recover days with differing transition pattern as the level increases. We capture an "at work followed by out" routine occurring for varying time durations revealing all sorts of "work-out-work" routines in the data.

of location transitions.

D. Vocabulary Analysis

One question that arises is whether the T top words Φ_i^T are not simply the most frequent words. To investigate this we compare the number of overlapping words between the T most probable words given topics at level 1, Φ_1^T, and the most frequently occuring words. In Figure 6, we plot the percentage of overlapping words between the most frequently occurring words and the top words Φ_1^T output by LDA at level 1 (single words). We consider varying numbers of top words $p \cdot T, p = 1, 2, \ldots$ and most frequently occurring words, which is on the x-axis of the plot. The y-axis shows the percentage of overlapping words. Considering up to 20 top words output by LDA topics, and up to the 20 most occurring words in the corpus, only 1 word overlaps, which explains the downwards trend in the plot up to 20. After 20, the number of overlapping words starts to increase.

This plot illustrates that we are not simply capturing the most frequently occurring words. By taking the top words output by topics to form bigrams, we are capturing something new, namely, words that co-occur often enough to form clusters, which is what LDA inference achieves at each level.

Limitations Though the multi-level topic model revealed interesting results, we also discovered a few limitations. The first one is at each level only a single number of consecutive words (or n-gram size) is considered. More specifically, at level 2, the vocabulary only consists of bigrams, and at level 3, the vocabulary only consists of trigrams. So the topics discovered are limited by co-occurring bigrams at level 2 and co-occurring trigrams at level 3, and so on. If we could combine all unigram, and bigrams, up to n-grams, into a single vocabulary at level n, the routines discovered may reveal new dominating human routines in the data. Another limitation of this work is the pre-specification of the number of topics at each level.

V. Conclusion

In this work we devise a probabilistic multi-level topic model to discover human routines of semantic locations. We apply the model successfully to a large, real-life human location dataset consisting of 97 users over 1 year, and discover unique trends, such as "working in the afternoon" occurring over 3 hour durations and 6+ hour durations. An unsupervised method to discover human routines considering varying length time durations has never been modeled, to our knowledge. The technique used in this paper can be easily applied for n-gram vocabulary construction for large n (previous methods will break down due to vocabulary size explosion after $n = 3 - 4$). These techniques may be integrated to a wide variety of applications.

Acknowledgments This research has been supported by the Swiss National Science Foundation through the MULTI project. We thank Nathan Eagle (Santa Fe) and Alex Pentland (MIT) for sharing the data.

References

[1] F. Bimbot, R. Pieraccini, E. Levin, and B. Atal. "Variable-length sequence modeling: multigrams," *IEEE Signal Processing Letters, 2(6), 111-113*, 1995.

[2] D. Blei, A. Ng and M. Jordan. "Latent Dirichlet Allocation," *JMLR 3*, 2003.

[3] N. Christakis and J. H. Fowler. "The Spread of Obesity in a Large Social Network over 32 Years," *N Engl J Med, 357(4), 370-379*, 2007.

[4] K. Church and W. Gale. "Concordances for parallel text," *Proc. of the 7th Annual Conf. of the UW Centre for the New OED and Text Research, 40-62*, 1991.

[5] T. E. Dunning. "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics, 19(1):61-74*, 1993.

[6] N. Eagle and A. Pentland. "Eigenbehaviors: Identifying Structure in Routine," *Behavioral Ecology and Sociobiology 63:7, 1057-1066*, 2009.

[7] N. Eagle, A. Pentland, and D. Lazer. "Inferring Social Network Structure using Mobile Phone Data," *PNAS, 106(36) 15274-8*, 2009.

[8] K. Farrahi and D. Gatica-Perez. "What Did You Do Today? Discovering Daily Routines from Large-Scale Mobile Data," *ACM MM*, Vancouver, Canada, 2008.

[9] J. H. Fowler, N. Christakis. "Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study," *BMJ, 337*, 2008.

[10] T.L. Griffiths and M. Steyvers. "Finding Scientific Topics," *PNAS 101:5228-5235*, 2004.

[11] R. Huerta, and Lev S. Tsimring. "Contact tracing and epidemics control in social networks," *Phys. Rev. E, 66:5 056615-19*, November, 2002.

[12] T. Huynh, M. Fritz, and B. Schiele. "Discovery of activity patterns using topic models," *UbiComp, 10-19*, Seoul, Korea, 2008.

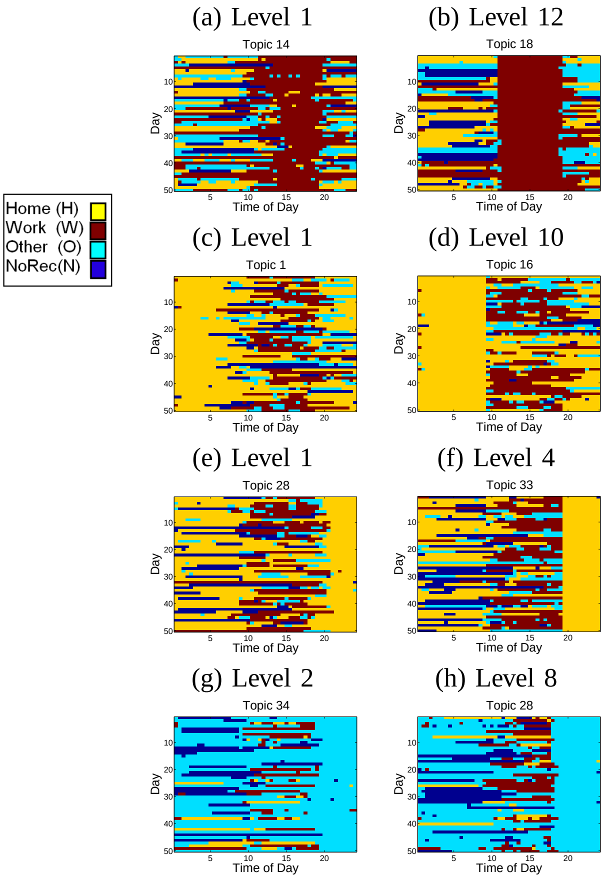

Fig. 3. Selected topics, illustrating the difference between highly ranked days for topics at lower levels and topics at higher levels. The corresponding most probable words for these topics visualized are listed in Figure 4. Comparing plot (a) to (b), where (a) corresponds to topic 14 at level 1 and (b) to topic 18 at level 12. (Note the topic numbers do not need to match at varying levels.) We can see the 50 most probable days for this "working" activity are more refined and consistent over time in level 12 then in level 1 due to the vocabulary being on a smaller scale at level 1. In level 12 the work routine occurs consistently from about noon to 6:30pm (as seen by the top words in Figure 4), whereas in level 1, the work is not consistently occurring during this time interval, though it occurs consistently from 5:30-6:30pm. A similar pattern can be seen from (c) to (d), where (c) is an "at home in the morning" routine captured for unigrams, and (d) is an at home until 8:30am routine. A similar phenomenon can be seen from (e) on level 1 to (f) on level 4 and (g) on level 2 to (h) on level 8. Overall, we observe the topics discovered reveal more refined and meaningful routines and the vocabulary becomes more discriminant as the level increases.

(a) Level 1 - Topic 14	
Work 5:30-6pm	0.120
Work 6-6:30pm	0.120
(b) Level 12 - Topic 18	
Work 12-6pm	0.184
Work 12:30-6:30pm	0.165
(c) Level 1 - Topic 1	
Home 2:30-3am	l 0.105
Home 1:20-2am	0.102
(d) Level 10 - Topic 16	
Home 3-8am	0.129
Home 2:30-7:30am	0.117
(e) Level 1 - Topic 28	
Home 11:30-12pm	0.177
Home 11-11:30pm	0.163
(f) Level 4 - Topic 33	
Home 10-12pm	0.183
Home 10:30-12:30pm	0.180
(g) Level 2 - Topic 34	
Out 10:30-11:30pm	0.141
Out 10-11pm	0.129
(h) Level 8 - Topic 28	
Out 8:30-12:30pm	0.214
Out 8-12pm	0.211

Fig. 4. The two most probable location words which are the n-grams for a topic at level n corresponding to the topics and levels of the plots in Figure 3. For example, in (a) the two most probable words for topic 14 at level 1 are "Work from 5:30-6pm" and "Work from 6:6:30pm" and their corresponding probabilities are both 0.12. The 50 highest ranked days for this topic are plot in Figure 3(a). The top words in the tables that follow (b) to (h) all follow the same structure, with their corresponding plots in Figure 3 labeled as (b) to (h), respectively.

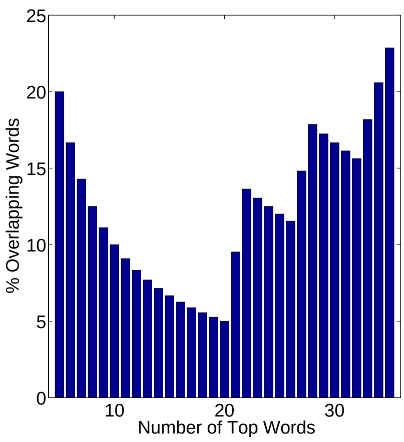

Fig. 6. The percentage of word overlap between the most frequently occurring words and the top words output by LDA plot as a function of the number of words considered. The x-axis is the number of top words and the number of most frequently occurring words, and the y-axis is the percentage of overlap. We can see these do not overlap greatly, indicating we are not simply capturing highly frequent words, but co-occurrence, which grows at each level.

[13] A. Madan, D. Lazer and A. Pentland. "Social Sensing to Model the Evolution of Political Opinions", (in submission).
[14] J.C. Niebles, H. Wang, and L. Fei-Fei. "Unsupervised learning of human action categories using spatial-temporal words," *IJCV, 79(3) 299-318*, 2008.
[15] J. S. Justeson and S. M. Katz. "Technical terminology: some linguistic properties and an algorithm for identification in text," *Natural Language Engineering, 1*, 1995.
[16] F. Smadja. "Retrieving collocations from text: Xtract," *Computational Linguistics, 19*, 1993.
[17] M. Steyvers and T. Griffiths. "Matlab topic modeling toolbox 1.3," 2005.
[18] H. Wallach. "Topic modeling: beyond bag-of-words," *ICML*, 2006.
[19] X. Wang, A. Mccallum, and X. Wei. "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," *ICDM*, 2007.