IDIAP RESEARCH REPORT

Discovering Human Routines from Cell Phone Data with Topic Models

Katayoun Farrahi [1] Daniel Gatica-Perez [1]

IDIAP–RR 08-32

July 2008

to appear in
ISWC08

[1] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, {kfarrahi,gatica}@idiap.ch

IDIAP Research Report 08-32

Discovering Human Routines from Cell Phone Data with Topic Models

Katayoun Farrahi Daniel Gatica-Perez

July 2008

to appear in
ISWC08

Abstract. We present a framework to automatically discover people's routines from information extracted by cell phones. The framework is built from a probabilistic topic model learned on novel bag type representations of activity-related cues (location, proximity and their temporal variations over a day) of peoples' daily routines. Using real-life data from the Reality Mining dataset, covering 68 000+ hours of human activities, we can successfully discover location-driven (from cell tower connections) and proximity-driven (from Bluetooth information) routines in an unsupervised manner. The resulting topics meaningfully characterize some of the underlying co-occurrence structure of the activities in the dataset, including "going to work early/late", "being home all day", "working constantly", "working sporadically" and "meeting at lunch time".

Contents

1 Introduction

Human activity modeling from large-scale sensor data is an emerging domain in ubiquitous computing towards determining the behaviour and habits of individuals and the structure and dynamics of organizations [3, 1]. In particular, given the massive amount of data that can be captured by cell phones for many individuals over long periods of time, fundamental questions to address through automatic analysis include: How characteristic is mobile sensor data (e.g. location extracted from cell tower information, proximity measured by Bluetooth devices) of people's daily routines? Further, what are these routines and how can we discover them? The applications of this analysis range from tools to support social science research to self-assessment tools.

Although most people follow certain daily routines, their identification is not a trivial problem given the often noisy and partial data that can be captured with a cell phone in terms of location or interaction. For automatic analysis, a supervised learning approach to activity recognition would require prior knowledge regarding the activities in question. In contrast, an unsupervised learning approach has the potential of automatic discovery of routines, not requiring training data.

In this paper, we develop a novel methodology built on topic models [5, 2] to address the questions above. Topic models are powerful tools, initially designed for text documents [5, 2]. Recently, they have been successfully applied to querying, clustering, and retrieval tasks for data sources other than text, such as images, video, and genetics [6]. Topic models are generative models, that can be used to represent documents as mixtures of topics, to learn a latent space, and they allow for clustering. Topic models are advantageous to activity modeling tasks due to their ability to effectively characterize discrete data represented by bags (i.e. histograms of discrete items). A time component can be incorporated into the bag representation. Further, we can take advantage of the bag to find routines at different temporal granularities. Further, topic models prove to be effective in filtering out the immense amount of noise in complex real-life data. They can be applied to a wide variety of data, given a bag can be constructed to represent a person's daily observations.

Our framework is used to automatically discover proximity- and location-driven routines from the day in the life of a person. The topic model characterizes the underlying co-occurrence structure in the location and proximity datasets well, with the discovery of activities including "going to work/home late/early", "working constantly", "working sporadically", and "meeting at lunch time".

The first contribution of this paper is the design of a methodology for discovery of daily routine patterns from cell phone data based on topic models. The second contribution is the evaluation of this methodology on the massive, complex, real-life Reality Mining dataset.

2 A Topic Framework for Routine Discovery

2.1 Bag Representations

We use the Reality Mining dataset [1] for which the activities of 100 subjects were recorded by Nokia 6600 smart phones over the 2004-2005 academic year at MIT. This comprises over 800 000 hours of data on human activity. The subjects are students and staff of MIT that live in a large geographical area covered by over 32000 cell towers. They work in offices with computers that have Bluetooth devices which can sense in a 5-10m radius [1]. The privacy concerns of individuals in the study have been accounted for by the collectors of this dataset.

Given a day in the life of a person in terms of where they go or who they're in proximity with, the goal is to automatically discover real routines hidden in the enormous volume and complexity of information. In the first part of this work, 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 time considerations. For the second part of the paper, we represent a day for a person in terms of who they were in proximity with, keeping the time of proximate interaction as an additional source of information using coarse-grain time considerations.

Bag of Location Transitions

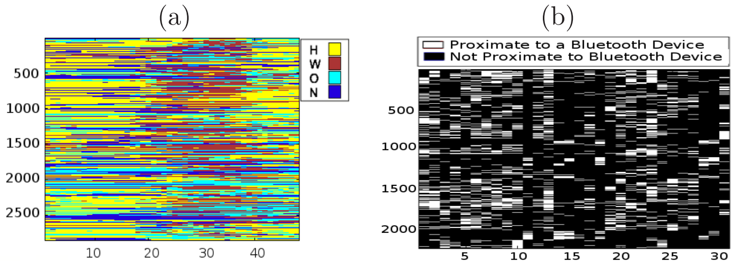


Figure 1: a) Fine-grain location visualized over all individuals' days (y-axis) in the study. The x axis corresponds to the 48 half-hour intervals in a day. b) Proximity displayed over all people and days (y-axis). The x-axis corresponds to the 30 individuals considered.

For a given individual, the dataset contains entries for each connected cell tower, as well as the start and end connection date and time. Over 32 000 towers are seen by all the people and 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 DATA(N), when there is no tower connection recorded for a person for a given time (eg. no battery, phone off or no reception).

A day in the life of a person can be expressed as a sequence of location labels (H,W,O,N). We divide a day into fine-grain, 30 minute timeslots, resulting in 48 blocks per day. For each block of time, we chose a 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 1a.

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 the house at 7:30am as opposed to 8am, we want to capture the important feature of "leaving the house early in the morning").

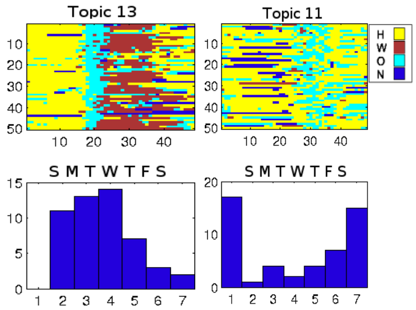
A *location word* (in analogy with real words in the case of text bags) contains 3 consecutive location labels of the fine-grain representation (corresponding to 1.5 hour intervals) followed by the coarse-grain timeslot in which it occurred. Thus a location word has 4 components, 3 location labels followed by a coarse timeslot label. 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.

Bag of Proximity Words

The proximity dataset includes 2 Bluetooth device IDs and the date and time of interaction. The visualization in Figure 1b, over all the individuals and days (y-axis), illustrates if a proximate interaction occurs with an individual (x-axis) within the day (disregarding time). A *proximity word* contains the 2 people whose devices have been in proximity, and the coarse-scale timeslot (of the 8 possibilities described before) in which the interaction took place. Thus a proximity word contains 3 components: 2 individuals and a timeslot. The bag is the histogram of the proximity words in the day.

2.2 Topic Models for Routine Discovery

For topic models applied to text, documents are represented as mixtures over hidden topic variables, where each topic is characterized by a distribution over words [5, 2]. Probabilistic Latent Semantic Analysis (PLSA) is one such model in which each observed word w_j is conditionally independent of the document d_i it belongs to given a topic z_k. The term-document joint probability, assuming K topics, is given by: $P(w_j, d_i) = P(d_i) \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$.



WORD Location Slot		P(W|Z13)
HHH	1	0.2489
HHO	2	0.1373
OWW	3	0.1322
OOO	3	0.1090
HOO	2	0.0849
OOW	3	0.0801
OOO	2	0.0368

WORD Location Slot		P(W|Z11)
HHH	6	0.1578
OHH	6	0.1249
OHH	5	0.1045
OOH	5	0.1030
HHO	6	0.0913
HOO	5	0.0851
OOH	6	0.0771

Figure 2: Routines characteristic of weekdays (Figure Topic 13) and weekends (Topic 11) visualized for top 50 days (ranked by $P(d|z)$). The corresponding bottom histograms are days of the week, SMTWTFS, for topics 13 and 11. Tables display top words ranked by $P(w|z)$ for topics 13 and 11.

The maximum likelihood parameters are estimated using Expectation-Maximization. By ranking $P(z|d)$, we can rank top documents by topics, resulting in the most characteristic location- or proximity-days per topic. We can also rank $P(w|z)$, resulting in the top words per topic. In routine discovery, topics can be used to find where the largest sources of noise are coming from (certain topics will show top documents for noise sources eg. no cell tower connections all day).

In work by Eagle and Pentland [4], 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 topic models. Unlike PCA, topic models are probabilistic, and thus have advantages with respect to clustering and ranking days. Our work also differs since we investigate proximity-driven routine discovery, in addition to location-driven. 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 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. The individuals selected had the most number of days with at least one W or H label. Of the people selected, six were business students and the others were Media Lab students of various levels (undergraduate and graduate). For the location experiments, we removed days which were entirely N (no data) labels since they contained no useful information. The resulting dataset is still massive and complex, amounting to 2856 days (over all people), and over 68 000 man-hours of very noisy data, as seen by Figure 1a.

For the proximity experiments, we used the same individuals and days as for the location experiments. Proximity entries were only considered if both proximate people were within the subset of 30 individuals considered. We did not consider days without any proximity entries since they contained no useful information. The resulting dataset amounts to 2236 days, and over 53 000 man-hours.

3.1 Location-Driven Routine Discovery

We apply our methodology to discover routines with $K = 30$ hidden topics (other values of K produced similar results). The results revealed routines of different types, with many topics characterized by $P(w|z)$ (top words given a topic) and $P(d|z) \propto P(z|d)$ (top days given a topic), following characteristic

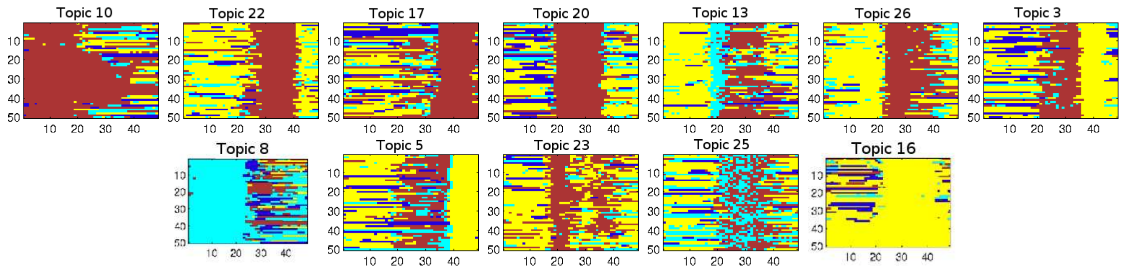


Figure 3: Visualization of some location-driven routines discovered, including: “going to W late/early”, “leaving W late/early”, “working constantly”, “working sporadically”, “H all day”.

trends. On a weekly level, some trends characteristic of weekends (topic 11) versus weekdays (topic 13) appeared, as illustrated in Figure 2, though other topics displayed weekend/weekday features as well. The top words for topic 13 contains H to O patterns co-occurring with O to W patterns in the mornings, corresponding to a “going to work” routine. Top words for topic 11, which are more characteristic of weekends, contain patterns of H or O labels in the afternoon and evenings, corresponding to “being at home” or “going out” in the afternoon and evenings.

Our method also discovered different types of routines for work days, visualized for top days for many topics in Figure 3. For example, documents in topics 10, 22, 17, and 20 reveal days for which people worked continuously without breaks, whereas topics 25 and 23 reveal fluctuations between W and O or H labels. Thus, it appears the method can differentiate “working constantly” versus “working sporadically” patterns. The routines of “going to work early” (topic 10) and “going to work late” (topic 17) are also differentiated. The “going home early” routines (topics 3, 20) have also been discovered, in contrast to “going home late” (topics 22, 17).

The method also found the routine of going “somewhere” (O location) between home and work in the mornings on certain days but not in the evenings (topic 13), perhaps indicating a class or event on certain days. Other days exhibited a similar trend in the evenings but not in the mornings (topic 5). Further, routines of “being out in mornings” (topic 8) and “being home all day” (topic 16) were discovered. See www.idiap.ch/˜kfarrahi/Demo/wc08.wmv.

3.2 Proximity-Driven Routine Discovery

We also apply our method on the proximity representation with $K = 30$ topics. While results corresponding to those of Figure 3 could be presented, for space reasons we present a different analysis that can be done with the learned topic model.

We rank the top words per topic, $P(w|z)$, and look at the top 99% of those words. Those top words are constructed of a timeslot (Figure 4a) and a pair of individuals (Figure 4b). It is clear that proximity patterns for certain timeslots are characteristic of various topics. For example, topic 14’s proximity routines occur mostly in timeslots 3 and 8 for a group of people. Many of the topics have top words with timeslots 4 and 5, corresponding to lunch time, revealing most group interaction occurs at this time. Topic 20 and 3 have strong components in earlier timeslots, indicating an interaction earlier in the day, whereas topics 2, 8, 25, and 27 have a stronger component in timeslot 6, corresponding to interactions later in the day. The visualization of proximate individuals per topic also reveals interesting patterns. Some topics (as can be seen in the columns in Figure 4b) are strongly correlated with the routines of a single person (topics 2, 8), pairs of people (topics 1, 30), or groups of people. The distribution of a person’s top words per topic (looking at rows of Figure 4b) reveals people with many interactions within the group, as opposed to people without many interactions. For example, individuals 7, 10, and 23 do not interact much within the group, whereas individuals 9, 26 and 30 have strong components over a few topics indicating many group interactions.

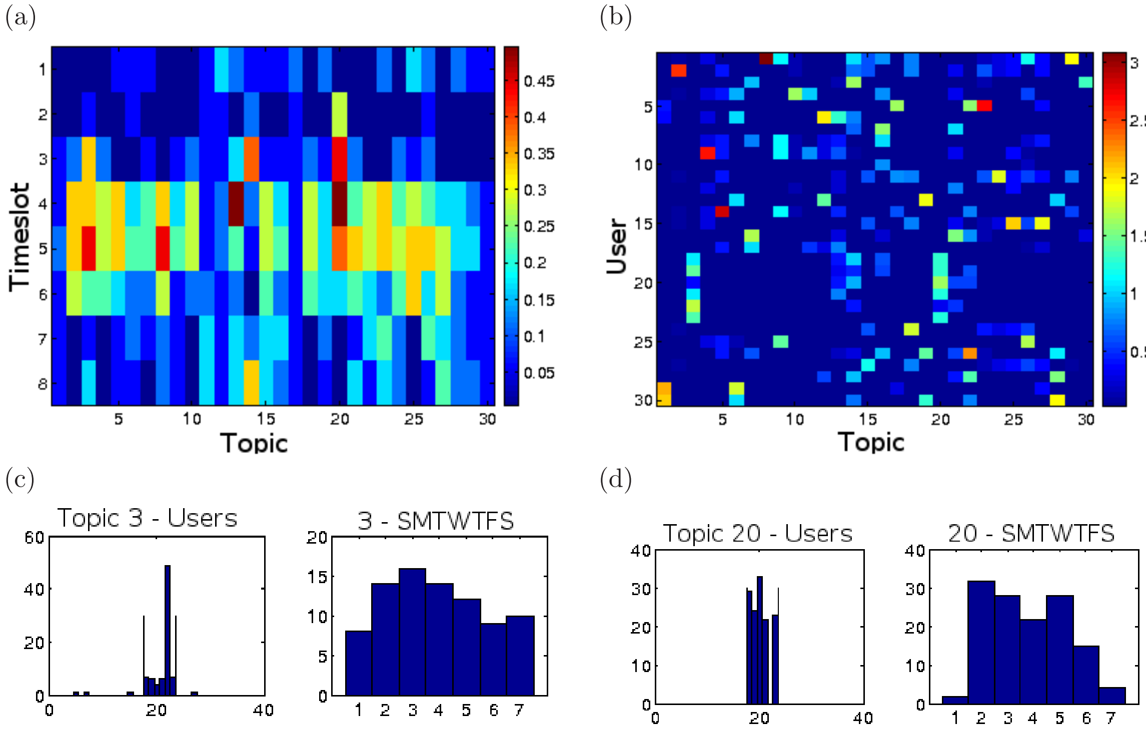


Figure 4: a) Timeslot component of top ranked words per topic, $P(w|z)$. The timeslots reveal most proximity interactions occur in slots 4 and 5, corresponding to lunch times. b) Users component of top ranked words per topic, $P(w|z)$. This visualization demonstrates topics characterizing interactions between pairs of people, groups of people, or a single person. c) Histogram of the individuals and days of the week for top ranked days of topic 3 ranked by $P(z|d)$. d) The same as c) but for topic 20. Users for topic 3 and 20 correspond to business students.

The method discovered two topics, 3 and 20, corresponding to routines of the group of business students, visualized for the top 20% days (ranked by $P(z|d)$) in Figure 4c and d. Individual 18 to 23 correspond to all of the business students from the 30 students considered in our study. Topic 20 (Figure 4d) is characteristic of all of the business students' interactions. The person whose proximity routines correspond closer to topic 3 contains routines which occur on weekends as well as weekdays, whereas the routines captured in topic 20 are characteristic of weekdays. Thus, our approach was able to automatically discover the group of business students, whose proximity routines differed from the other groups of interactions found, though this discovery would not be possible without prior knowledge of student types.

4 Conclusion

We have presented a framework for which location- and proximity-driven activity patterns are discovered automatically from 68 000+ hours of noisy, real-life cell phone data using topic models. Our method successfully discovered routines from both data types on a daily scale. Extensions of this work could include the discovery of routines on both larger and smaller time scales to determine what sorts of routines can be discovered in both unsupervised and supervised ways. Further, we hope to investigate new models in addition to the phone call data collection.

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

References

[1] N. Eagle and A. Pentland. "Reality mining: Sensing complex social systems," *Personal and Ubiquitous Computing 10(4)*, 255-268, 2006.

[2] D. Blei, A. Ng and M. Jordan. "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3*, 2003.

[3] T. Choudhury and A. Pentland. "Sensing and Modeling Human Networks using the Sociometer," *Proc. of ISWC*, 2003.

[4] N. Eagle and A. Pentland. "Eigenbehaviors: Identifying Structure in Routine," *Behavioral Ecology and Sociobiology (in submission)*, 2007.

[5] T. Hofmann. "Probabilistic Latent Semantic Analysis," *Proc. of Uncertainty in Artificial Intelligence*, Stockholm, 1999.

[6] F. Monay and D. Gatica-Perez. "Modeling Semantic Aspects for Cross-Media Image Retrieval," *IEEE Trans. on PAMI*, 2007.