Extracting Urban Patterns from Location-based Social Networks

Laura Ferrari, Alberto Rosi, Marco Mamei, Franco Zambonelli
Dipartimento di Scienze e Metodi dell'Ingegneria
University of Modena and Reggio Emilia, Italy
name.surname@unimore.it

ABSTRACT

Social networks attract lots of new users every day and absorb from them information about events and facts happening in the real world. The exploitation of this information can help identifying mobility patterns that occur in an urban environment as well as produce services to take advantage of social commonalities between people. In this paper we set out to address the problem of extracting urban patterns from fragments of multiple and sparse people life traces, as they emerge from the participation to social networks. To investigate this challenging task, we analyzed 13 millions Twitter posts (3 GB) of data in New York. Then we test upon this data a probabilistic topic models approach to automatically extract urban patterns from location-based social network data. We find that the extracted patterns can identify hotspots in the city, and recognize a number of major crowd behaviors that recur over time and space in the urban scenario.

Categories and Subject Descriptors

G.3 [**Probability and statistics**]: Time series analysis; H.3.3 [**Information Search and Retrieval**]: Retrieval Models; I.5.2 [**Design Methodology**]: Pattern Analysis

General Terms

Algorithms, Human Factors, Measurements, Experimentation, Performance, Verification

Keywords

social dynamics, spatio-temporal data mining, information retrieval in location-based social networks, semantic meaning and knowledge discovery from location-related data

1. INTRODUCTION

Thanks to advancement in mobile technologies, PDAs and smart-phones are the most rapidly growing technologies in the world [8]. With the assent of the owner, such devices silently observe people going through their lives, record the places they visit, the information they share, the people with whom they interact. Finally, the information produced can be globally shared in social networks, under the shape of a Facebook post, a Twitter tweet, a Flickr geolocalized picture, etc.

ACM LBSN '11, November 1, 2011. Chicago, IL, USA
Copyright ©2011 ACM ISBN 978-1-4503-1033-8/11/11...$10.00

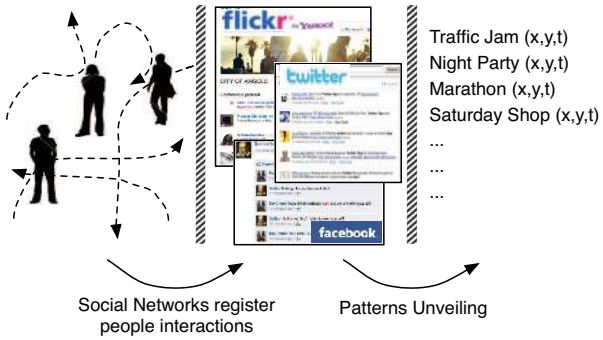

Figure 1: From people interactions to events

Social data represents a still under-explored treasure of information, especially when it is further enriched with the location dimension (e.g., Facebook Places, Geolocalized Twitter Post or Flickr Pictures). Beside sparse in the geographical space, incomplete in timespan and fragmented on millions of users, social data expresses a myriad of "life traces" describing, by social network conventions, the way in which people live and interact in their own city. Such vision is exemplified in Figure 1. On the left side, people produce traces during their own lives; such traces intersect when people participate, create or, are subjected to, events that are happening in the same time and place in the urban environment.

It is important to remark that if from one side social networks are effective in registering and capturing single user moods and comments, they are not conceived for performing analysis at a crowd level. Despite the sparsity of social networks data, in our view it is possible to extract from distinct traces higher level information such as people similarities, recurrent behaviors, or inference on upcoming events.

In this paper we set out to investigate the possibility of extracting urban patterns from location-based social networks data. In particular, we explore the extent that recurrent

patterns in an urban environment can be computed if we use only the spatio-temporal feature of social network data. To do so, we apply a topic model-based approach on a large dataset of Twitter posts. We want to show that our proposal, based on a probabilistic topic model approach, can successfully discover high-level patterns from location-based social networks. By applying our approach, we make two main contributions:

1. we show that a probabilistic topic model technique can be successfully applied not only to a "user-centric" scenario, but also to a "city-centric" ones. In fact, in previous state-of-the-art works, topic models have been applied to extract patterns and routines behaviors from a fairly large and accurate log of individual users behaviors [5, 6]. Here we propose a topic model based approach able to infer high-level patterns from data representing the life of a large set of users but in an extremely sparse and spotty way.

2. we show that our approach can be successfully applied to the sparse data coming from location-based social networks. We tested our approach on a large (3 GB) collection of geo-localized Twitter posts, created over the city of New York from June 2010 to June 2011. We show that our approach can recognize and extract, in an unsupervised manner, recurrent behaviors and high-level patterns that recur over both the space and time dimensions. Such results can find a natural application in grouping together people with similar interests and usage of public spaces, as well as in tracking or predicting events that have a strong relation with the urban environment.

The remainder of this paper is organized as follows. In Section 2, we discuss related work in the area of pattern recognition of geo-localized data. In Section 3 we describe our model for adapting social mobility traces to topic models. In Section 4, we discuss, evaluate and validate results. In Section 5, we describe the possible applications that are enabled by the automatic extraction of urban patterns. Section 6 concludes and identifies areas for future research.

2. RELATED WORK

In recent years, many approaches have been proposed both for *(i)* extracting hotspots and *(ii)* identifying spatio-temporal patterns from geo-localized data.

Extracting hotspots.
The CitySense project (*http://www.citysense.com*) uses GPS and WiFi data to cluster people whereabouts and discover hotspots of activity in the city area. In a similar work based on extremely large anonymized mobility data coming from Telecom operators, authors were able to identify the most visited areas by tourists during the day and the typical time of the visit (see for example [3], [7]). Another group of works is based on photo-sharing sites which contain billions of publicly accessible images taken virtually every where on earth. These photos are annotated with various forms of information including geo-location that implicitly identifies the location of the user. Researchers have been able to analyze a global collection of geo-referenced photographs, and evaluate them on nearly 35 million images taken from Flickr with the goal of identifying hot spots (and also tourist routine behaviors) [9, 12, 14] . From a complementary perspective, the problem of finding boundaries for vague regions corresponding to human-centered areas of the city looking at Web query logs was also studied in [17, 19].

Identifying spatio-temporal patterns.
Several researches have been developed to identify recurrent patterns in mobility data. In particular, some works rely on the identification of frequent sequential patterns to analyze spatio-temporal datasets. For example, in [10] the presence of frequent sequential patterns are used to find recurrent mobility patterns. Eagle and Pentland [4], use Principal Component Analysis (PCA) to identify the main components structuring daily human behavior. The main components of the human activities, which are the top eigenvectors of the PCA decomposition are termed *eigenbehaviors*. Similarly, the work presented in [18] compares different data mining techniques to extract patterns from mobility data. In particular, they found Principal Component Analysis (PCA) and Independent Component Analysis (ICA) best suited to the task of identifying daily patterns.

A recent group of works is based on topic models, which are powerful tools initially developed to characterize text documents [2] and are at the base of our approach. In [5] authors propose the use of probabilistic topic models to capture human routines from cell tower connections. They were able to understand both individual behaviors and interactions. Similarly, in [16] authors propose a Latent Social Theme Dirichlet Allocation to automatically discover high-order temporal social patterns from very noisy and sparse location data. They apply the proposed framework to a real-world noisy dataset collected over 1.5 years, and show that useful and interesting patterns can be computed. In [6] topic models are applied on GPS signals obtained from the Google Latitude application. They used grid-based algorithms to extract significant locations that are either automatic and manually labeled. Locations are analyzed in a broader resolution than in [5, 16], involving places such as pub, cinema, disco, etc. Topic models are then used in a similar way as proposed in [5].

Our approach mainly deviates from the current state of the art works in two relevant aspects:

1. **Single-to-crowd analysis.** The focus of our analysis doesn't concern user-centric behavior (above works in literature extract routine patterns of individual users) but rather we mine for daily routines describing common patterns of the city as a whole. This kind of patterns extractions support the novel classes of applications shown in section 5.

2. **Complete-to-sparse datasets.** Works in literature applied topic models to datasets that are very hard to obtain (mainly for privacy purposes), since they are built on complete traces over the 24h. Our approach adapts topic models to work on a large set of scattered mobility traces as they emerge from location-based social networks (e.g., Twitter, Foursquare, Brightkite);

most of the time such data locates a huge number of users only for a tiny fraction of the day.

3. A MODEL FOR REPRESENTING SOCIAL MOBILITY TRACES

In this section, we present the mechanisms at the core of our approach to extract patterns and routine behaviors from location-based social networks.

In extreme summary, our approach is based on three steps:

1. **Crowd Detection:** given a selected geographical area of analysis, we detect the most crowded areas at any given hour of our dataset. Crowd detection is based on social networks life traces, in our case we bank on the number of Twitter posts sent by users. Obviously such "measure" of crowd distribution could be replaced or combined with other location-based social networks, such as the number of picture taken in that area (considering the Flickr database), the number of Facebook posts sent from there, etc.

2. **Data preparation:** starting from the above results, we translate such data in a form that is suitable for topic models. In particular, we transformed each day in our dataset in a complete collection of *crowd-footprints* (as detailed in section 3.2) each indicating the time and the areas of the city that are most crowded. Such collection of *crowd-footprints* is the input data structure for urban patterns recognition.

3. **Urban patterns recognition:** on the above crowd aggregated data, we run a topic models algorithm to discover recurrent behaviors (i.e.,*topics*, as detailed in section 3.3) over space and time. As depicted in Section 5, such crowd behaviors could help in track or predicting events that have a strong relation with an urban environment.

In the following of this section we detail the above three steps.

3.1 Crowd Detection

The first step of our approach is to find what are the most crowded areas in the city at a given time slot. Our approach applies to a dataset of geolocalized posts from many users of a social network site spanning several days. The approach works as follow:

1. We divided each day into a $H = 12$ time-slots lasting 2 hours each. In general the greater the number of time-slots, the finer the routine behaviors being extracted. From our experiments we found that for our dataset 12 time-slots provide the most meaningful routine behaviors.

2. For each day d and time-slot h, we cluster location-based social networks data (in our case Twitter posts) with the aim of finding the most visited locations. In our experiments we used the EM algorithm [1] to perform this computation.

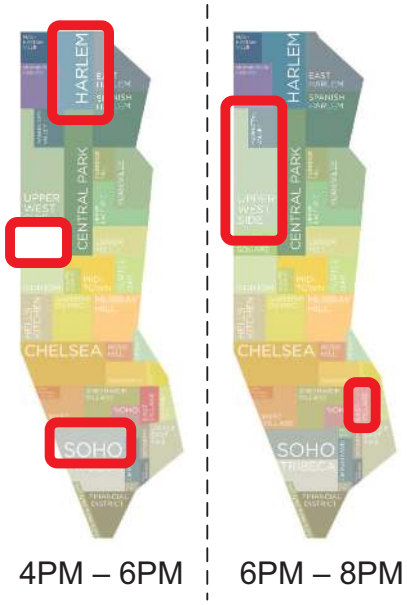

Figure 2: Crowd-footprint in a city. This figure shows the concentration of people in the city at different times.

3. We associate each cluster to the zip-code defined area it belongs to (see Fig. 2). Despite this step is not functional to topics extraction, it helps in characterizing an urban area in terms of the kind of activities being performed there.

This procedure identifies where and when the city is most crowded and how the crowd shifts across the city. This kind of detection focuses on city-centered information and disregards individual user behaviors.

3.2 Data preparation

In the second step, from the above identified footprints, we created for each day in our dataset a representation that is suitable for topic models.
Here we define a $crowd - footprint_x$ as a 2-tuple consisting of a time π_x and a place l_x where the $crowd - footprint_x$ is experienced: $crowd - footprint_x =< \pi_x, l_x >$.

More in details, we constructed each $crowd - footprint$ in the following way:

- starting time π_x is mapped into H time-intervals representing the hour of the day;
- a dictionary of Z zip codes is constructed and the nearest zip code of l_x is used.

This method of construction will theoretically yield a maximum number of *crowd-footprints* of $W = H * Z$ if all combinations of π_x and l_x are observed in the data.

In summary, each day is translated into a sequence of *crowd-footprints*. For example, a day is translated into the following sequence: $0-10001, 0-10004, 1-10008, 2-10010$,etc. thus representing that the crowd is located in zip 10001 and 10004 from 0am-2am, in zip 10008 from 2am-4am etc. The set of all days represents the input data structure for the next step.

3.3 Urban patterns recognition

In our methodology we adopted a probabilistic topic model technique [1] to identify the routine behaviors with which people move and cluster across the city. Topic models are powerful tools initially developed to characterize text documents, but can be extended to other collections of discrete data (e.g., mobility data). They are probabilistic generative models that can be used to explain multinomial observations by unsupervised learning. Formally, the entity termed word is the basic unit of discrete data defined to be an item from a vocabulary of size V. A document is a sequence of N words. A corpus is a collection of D documents. As above mentioned, in this paper a word is represented by a place and a time slot (i.e., $crowd-footprint$), while a document is a day of the city. There are K latent topics (i.e. routines) in the model, where K is defined by the user.

Among other topic modeling algorithms, in this paper we apply Latent Dirichlet Allocation (LDA) [2]. LDA has two main characteristics that make it suitable to our pattern discovery task. On the one hand, it is an unsupervised approach: it does not require to define classes (i.e. topics) *a priori* and it does not require difficult-to-be-acquired labeled data. On the other hand, topics represent meaningful probabilistic distributions over words and documents. This allows to analyze and understand the routine behavior they stand for. As shown in Section 2, LDA has been applied in this scenario only to "user-centric" data, where patterns and routine behaviors have been extracted from a fairly large and accurate log of individual users behaviors [5, 6]. One important contribution of this paper is to show an approach to apply LDA on data coming from location-based social networks, that represent the life of a large set of users but in an extremely sparse and spotty way.

More in detail, LDA is based on the Bayesian network depicted in Figure 3. A word w is the basic unit of data, representing $crowd-footprint$ at a given *time-label*. A set of N words defines a day of the city (i.e. a document). The city taken into consideration has a dataset consisting of D documents. Each day is viewed as a mixture of topics z, where topics are distributions over words (i.e., each topic can be represented by the list of words associated to the probability $p(w|z)$). For each day i, the probability of a word w_{ij} is given by $p(w_{ij}) = \sum_{t=1}^{K} p(w_{ij}|z_{ik})p(z_{ik})$, where K is the number of topics. $p(w_{ij}|z_{ik})$ and $p(z_{ik})$ are assumed to have multinomial distributions. Mixture parameters are assumed to have Dirichlet distributions with hyperparameters α and β respectively. LDA uses the EM-algorithms to learn the model parameters. In our implementation we use the library LingPipe (http://alias-i.com/lingpipe/) to perform these computations. Once the model is trained, Bayesian deduction allows to extract the topics best describing the routines of a given day (rank z on the basis of $p(d|z)$).

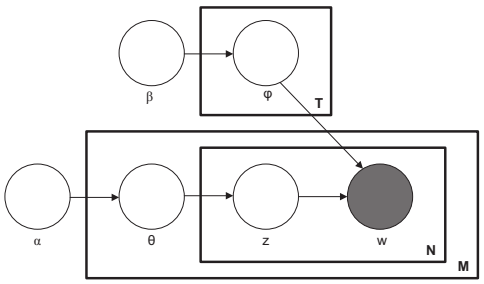

Figure 3: Plate notation representing the LDA model. α is the parameter of the uniform Dirichlet prior on the per-document topic distributions. β is the parameter of the uniform Dirichlet prior on the per-topic word distribution. θ_i is the topic distribution for document (day) i, ϕ_j is the topic distribution for word j, z_{ij} is the topic for the j-th word in document i, and w_{ij} is the specific word. The w_{ij} are the only observable variables.

4. EXPERIMENTS

In this section, we describe some experiments we conduced to evaluate the effectiveness of the proposed approach. First, we describe and motivate the adopted experiment setup, then we illustrate and discuss the obtained early results. Finally, we analyze the accuracy of the extracted topics.

4.1 Experiment Setup

The first aspect to consider to experiment with the proposed topic-based discovery is to select a dataset of geo-localized social network posts on which to ground the analysis. As the majority of social networks provide public API to access their data, the fundamental question is whether it is possible to have data critical mass to conduct meaningful analysis.

Among several candidates, Flickr and Twitter are those systems offering best access modality to a large number of geo-localized data. However, in our experience even Flickr and Twitter, despite the terabytes of data they produce daily, provide enough geo-localized data only on few selected cities. Accordingly, In this paper we used a dataset of mobility traces collected with Twitter over the city of New York (NY, USA) to test the effectiveness of the proposed approach. In particular, the dataset consists of all geo-referenced Twitter over a period of 1 year centered over Manhattan.

Despite many social sources haven't yet reached a mass of data enabling for robust information extraction, our strong belief is that in the next future social networks will be subjected to explosive adoption rates, rapidly reaching the threshold for being exploited in future pervasive applications and services.

Before staring the experiment, we run some clean-up on the data. Since we are interested on mobility patterns of people, we removed all those Twitter users that post always from the same coordinates, since they are likely to be services in which the geo-tags are associated to the service's premises and not to a real moving user. After this, we were left with 13 millions tweet with an average of 30.000 tweets per day.

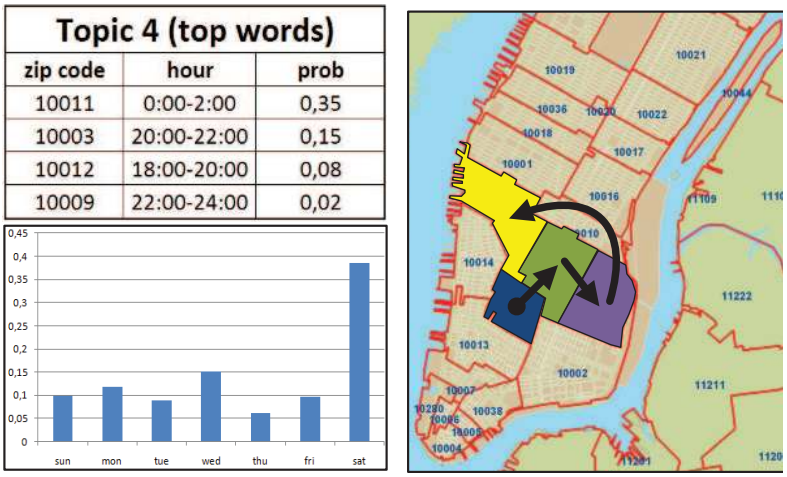

Figure 4: Routine behavior in Manhattan.

4.2 Results and Discussion

We applied the proposed mechanism to Twitter data with the aim of identifying hotspots in the city life, or rather crowd behaviors that recur over time and space in an urban scenario.

In particular, for these experiments we instantiated the LDA model setting $K = 30$ topics. In general the greater the number of topics, the finer the routine behaviors being extracted. The estimation of the optimal number of topics is an active research challenge and some mechanisms have been proposed to guide this choice [2, 5].

In a first experiment, we try to understand if the extracted patterns are meaningful and coherent with the expected city life. The LDA model successfully revealed some trends characteristic of the city. In Fig. 4 and Fig. 5, we illustrate two exemplary topics.

Topic 4 represents a typical weekend activity centered on Saturdays. It shows an hotspot centered around Greenwich village, East village and Nolita districts in Manhattan, characterizing the nightlife. The topic is represented by the table at the top left of the figure. It comprises several zip codes associated with a time period and a probability of that area begin one of the most crowded areas in the city. The topic is also represented in the map on the right of the figure, where zip codes are highlighted and arrows are drawn to illustrate how the concentration of people (i.e., crowd-footprint) shifts from an area to the other. Finally, at the bottom-left of the figure, we present the distribution of the topic over the days of the week. From this, we can see that this is a typical Saturdays' pattern.

Given the lack of accurate ground-truth information, and the fact that topics cluster data in classes that are not defined a priori, it is difficult to provide sound measures on the accuracy of the obtained results. However, the fact that the identified district from the hotspot are well-known areas of the city of night-life entertainment, partially validate the obtained results.

From a similar analysis, Topic 22 represents one possible week day in which the focus of activity is scattered all-over Manhattan ranging from Gramercy, Chelsea, Soho and East Harlem. Many other topics, involving other districts, describe other possible configurations of weekday activities. This is in-line and compatible with the mixed structure of Manhattan districts, but missing groundtruth data it is difficult to validate results further.

It is worth emphasizing that the represented topics do not correspond to the actual movement of people. People in an area might be different from the the people in another area. Topics represent only where there is a concentration of people and how this concentration moves across the city. The identity of individual is totally disregarded. This is very important in our opinion in that this kind of topic analysis completely preserves individual privacy and only illustrates aggregated information. The extensive study on protecting privacy in location-based social networks (e.g., [15, 20]) further motivates our topic models-based approach.

In summary, this first experiment illustrates that the LDA model applied to Twitter data successfully reveals different types of patterns, by assigning characteristic trends to various topics with a probability measure (p(w|z) and p(d|z)). In addition, based on such method, we are able to answer to several interesting questions such as "Are there specific patterns occurring on weekends versus weekdays?" or "How do the topics characterize the days in the dataset?". The above results illustrate also one of the key advantages of LDA compared to other clustering mechanisms (e.g., k-means). While most other clustering algorithms group together days that are similar for the whole 24 hours, LDA can cluster days that are similar only in a given time interval. For example, LDA can cluster the days for a specific night-life activity, even if those days have very different signatures in the morning. Other clustering mechanisms would not be able to identify that cluster since they consider whole days only [5].

Topic 22 (top words)		
zip code	hour	prob
10016	6:00-8:00	0.14
10036	10:00-12:00	0.11
10013	8:00-10:00	0.02
10128	12:00-14:00	0.01

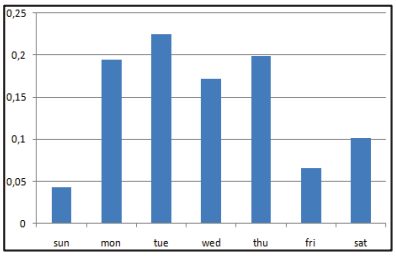

Figure 5: Routine behavior in Manhattan.

Day: Thursday

Topic	Time Slot	Zip Code
10	10-12	10001
	12-14	10027
	14-16	10003
3	16-18	10001
	18-20	10003
	20-22	10014

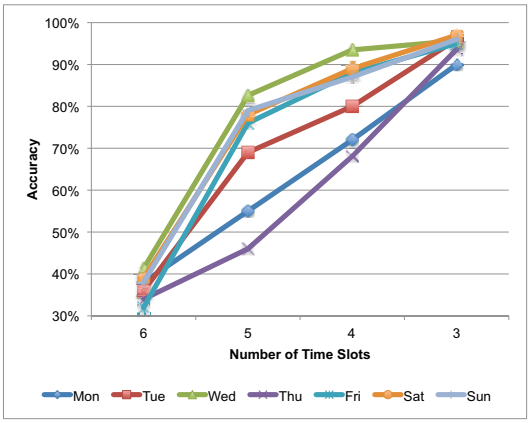

Figure 6: Topic Accuracy

4.3 Topics Accuracy

The purpose of this section is to perform a topic accuracy analysis, i.e., we investigate the ability for topics to recognize the event, or the pattern that they actually describe.

In particular, we employ our system to test the assumption that each day of the week, from Monday to Sunday, differs from the others for a peculiar distribution of crowd dislocations, around city areas and hours of the day. To perform such test, we divided our 12 months dataset in a 9 months training set and a 3 months testing set.

Starting from the training set, we run our system to provide, for each day of the week, the list of topics that better describe them. From these topics, we compose the sequence of visited areas that most likely (from a probabilistic point of view) describes each day. To better clarify this operation, we show in the left part of Figure 6 an example of a typical "Thursday" pattern. The figure illustrates that a typical "Thursday" can be described by topics 10 and 3. For each time slot, topics indicate which is the area (i.e., zip code) people are likely to get together. For example, the figure shows that in a typical "Thursday" from 10 am to 12 pm, a number of people is clustered in zip code 10001. Then from 12 pm to 2 pm, people cluster in zip code 10027, and so on.

Then, we tried to verify whether the extracted topics can actually describe a given day in general. We take the testing set into consideration. We compare each day of the week with the associated topic extracted before. For example, for each Thursday, we take the Thursday topic into account, and we verified whether the movements patterns of that particular day are in line with those described by the topic (i.e., whether the concentration of people in that Thursday, for a given time slot, happens to be in the zip code indicated by the Thursday topic).

In particular, we verified whether the particular day and

the associated topic are in line considering 3 time slots (the specific day and the topic have to be in line for at least 6 hours - 3 slots of 2 hours each), up to 6 time slots (the specific day and the topic have to be in line for at least 12 hours - 6 slots of 2 hours each).

Results are illustrated in Figure 6. For a given number of time slots the average accuracy of the topic description of that day is reported. For example, an accuracy of 40% for 6 time slots on Mondays means that, only 40% of Mondays in the testing set is described by the Monday topic for 6 time slots (12 hours - 6 slots of 2 hours each).

It is rather natural to see that the average accuracy is inverse proportional to the number of time slots being considered. The more the extent of the day we are tying to describe the less accurate we are because of the inherent variability in each individual day. Nevertheless, the graph shows a fairly good stability in the patterns being discovered as we are able to describe any given day over 3 time slots with accuracy greater than 90%.

As a side consideration, we should say that the limited employment of geolocalization over social networks posts (only the 3% of generated data is geolocalized) is still limiting the real effectiveness of our system. In fact, despite obtained daily patterns were all different, frequently they share each other multiple common sequences of "time slot - zip code". That general likeness between days of the week, in the case of matching based on few time-slots, leads our system on an higher level of false positive results (around 40% for matching based on 3 time slots).

5. APPLICATIONS

The presented topic-extraction mechanism and, more in general, the study of human-generated patterns could find a natural application in discovering social commonalities among people, as well as to track and predict events that have a strong correlation with a urban environment in terms of space and time.

Marketing and advertisement are natural domains for this kind of technologies [11]. Services built on top of the proposed topic extraction mechanism could group together people with similar interests and usage patterns of public spaces. For example, they could provide customized recommendations on *where to go*, and *what to see*, to people going to visit tourist places for the first time. In this context, patterns of visit could be inferred from past crowd experiences, and once a new user has been labeled with a tourist pattern (e.g. she has visited in sequence a list of popular places described by a topic), an applications could suggest her the next location to be visited, or a restaurant to have dinner [13, 21].

From a complementary perspective, once a sequence of events (or in general human activities) have been recognized as a characteristic topic for the city, we can use this information to make predictions about future events. In particular, once early events indicate the initial fulfillment of a topic, we can predict the next topic evolution. On this basis, two types of social inference could be performed:

- *Direct prediction.* This kind of prediction is related to the fact that a topic is about to happening. In this context, typical examples concern the prediction of urban dynamics as traffic occurrences and crowd displacement. A congestion on some side roads in most of the case will lead to choke up the contiguous main highways. Once a "traffic intensification" pattern is delineated, police could be dispatched in advance to cover the places that the topic predicts as the next ones for traffic diffusion. Similarly, on the happening of mass events (e.g., concerts, fairs, sport events, etc.), anomalous crowd behaviors could be caught and corrected by arranging the flow of people to avoid possible upcoming dangerous situations.
- *Indirect prediction.* This kind of prediction is related to the fact that an expected topic is not happening. This is about detection anomalous and rare event, for example in the case of disaster response. Usually a natural or human driven disaster modifies daily routines bringing people to different and anomalous behaviors. These behaviors (a growing number of phone calls, different car routes, traffic redirection, etc.) differ from typical daily patterns and are easily recognizable. If an emergency would take the ICT networks to the point of collapse, variations in patterns will result even more evident.

These application domains further motivate the proposed topic-extraction mechanism. In addition, since our mechanism can be applied to several location-based social network data, applications tailoring specific networks could be realized. For example, topics coming from photo-sharing sites will support more naturally tourist-oriented applications, while topics coming from Twitter our Foursquare are more in line with understanding the everyday activity of the city.

In conclusion, for the study of human-generated patterns via topic-extraction mechanisms we envision two promising areas of applicability. The first sees the use of such mechanisms as a standalone system for urban management, with the purpose of performing the above described social analysis and studies over an urban environment. The latter concerns the integration of topic-extraction mechanisms in future pervasive services where the view on urban facts and events, and crowd behaviors, will enrich pervasive applications with a further level of context awareness (i.e., "social-awareness").

6. CONCLUSIONS

In this paper, we presented an approach for the automatic extraction of urban patterns and recurrent behaviors from location-based social networks. In particular, we adopt a probabilistic topic model, Latent Dirichlet Allocation (LDA) to identify the routine behaviors with which people move and cluster across the city. Results illustrates that meaningful patterns about city routines can be discovered.

Beside further investigate the proposed application scenarios, and further refine our approach, our future work will proceed in two main directions. On the one hand we will try to apply the proposed approach to other kind of data from

social networks. This is important to better evaluate the generality of the proposed approach. In fact cross-validation across multiple data sources can validate the extracted topics also without ground truth information. On the other hand, we will compare the proposed LDA approach with other data mining techniques to extract patterns from geo-localized data. The goal of this study is to determine if combining some different mechanisms could lead to better results and precisions in the recognition of patterns.

Acknowledgements: Work supported by the SAPERE (*Self-Aware Pervasive Service Ecosystems*) project (EU FP7-FET, Contract No. 256874).

7. REFERENCES

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022, 2003.

[3] F. Calabrese, J. Reades, and C. Ratti. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *IEEE Pervasive Computing*, 9(1):78–84, 2010.

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

[5] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology*, 2(1), 2011.

[6] L. Ferrari and M. Mamei. Discovering daily routines from google latitude with topic models. In *IEEE International Workshop on Context Modeling and Reasoning*, 2011.

[7] F. Girardin, J. Blat, F. Calabrese, F. D. Fiore, and C. Ratti. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing*, 7(4):36–43, 2008.

[8] R. Heeks. Ict4d 2.0: The nextphase of applying ict for international development. *IEEE Computer*, 41(6):26–33, 2008.

[9] L. Hollenstein and R. Purves. Exploring place through user-generated content: using flickr to describe city cores. In *JOSIS*, number 1, pages 21–48, 2010.

[10] Y. Huang, L. Zhang, and P. Zhang. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):433–448, 2008.

[11] J. Krumm. Ubiquitous advertising: The killer application for the 21st century. *IEEE Pervasive Computing*, 10(1):66–73, 2011.

[12] D. Leung and S. Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco (CA),USA, 2010.

[13] K. W.-T. Leung, D. L. Lee, and W.-C. Lee. Personalized web search with location preferences. In *IEEE International Conference On Data Engineering*, Long Beach (CA),USA, 2010.

[14] M. Mamei, A. Rosi, and F. Zambonelli. Automatic analysis of geotagged photos for intelligent tourist services. In *IEEE Intelligent Environment*, Kuala Lumpur, Malaysia, 2010.

[15] M. Mano and Y. Ishikawa. Anonymizing user location and profile information for privacy-aware mobile services. In *Proceedings of 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, San Jose (CA),USA, 2010.

[16] D. Phung, B. Adams, and S. Venkatesh. Computable social patterns from sparse sensor data. In *First international workshop on Location and the Web*, 2008.

[17] M. Sharifzadeh, C. Shahabi, and C. A. Knoblock. Learning approximate thematic maps from labeled geospatial data. In *Proceedings of International Workshop on Next Generation Geospatial Information*, Boston (MA),USA, 2003.

[18] S. Sigg, S. Haseloff, and K. David. An alignment approach for context prediction tasks in ubicomp environments. *IEEE Pervasive Computing*, 9(4):90–97, 2010.

[19] F. Twaroch, C. Jones, and A. Abdelmoty. Acquisition of a vernacular gazetteer from web sources. In *LocWeb*, Beijing, China, 2008.

[20] C. R. Vicente, D. Freni, C. Bettini, and C. S. Jensen. Location-related privacy in geo-social networks. *IEEE Internet Computing*, 15(3):20–27, 2011, to appear.

[21] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of 19th International World Wide Web Conference*, Raleigh (NC),USA, 2010.