Mining User Behaviours: A Study of Check-in Patterns in Location Based Social Networks

Daniel Preoţiuc-Pietro, Trevor Cohn
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield, S1 4DP, United Kingdom
daniel@dcs.shef.ac.uk, t.cohn@dcs.shef.ac.uk

ABSTRACT

Understanding the patterns underlying human mobility is of an essential importance to applications like recommender systems. In this paper we investigate the behaviour of around 10,000 frequent users of Location Based Social Networks (LBSNs) making use of their full movement patterns. We analyse the metadata associated with the whereabouts of the users, with emphasis on the type of places and their evolution over time. We uncover patterns across different temporal scales for venue category usage. Then, focusing on individual users, we apply this knowledge in two tasks: 1) clustering users based on their behaviour and 2) predicting users' future movements. By this, we demonstrate both qualitatively and quantitatively that incorporating temporal regularities is beneficial for making better sense of user behaviour.

Keywords

Social networks, Location Based Social Networks, Foursquare, Mobility patterns, Clustering, User behaviour, User movement prediction, Data mining

Categories and Subject Descriptors

[**Information systems**]: [Information systems applications, Spatial-temporal systems, Location based services]

1. INTRODUCTION

Recently, online social networks (OSNs) have seen a rapid increase in usage which resulted in gaining access to large amounts of data. This presents us with the opportunity to study areas in which data collection and size was an issue to date. Location Based Social Networks (LBSNs), where the focus is on sharing the user's current location and activity, have became mainstream. The data that arises from the use of these services can be harnessed for different purposes, from providing spatial aware information to a better understanding of human mobility, the scope of the present research.

Consider the example of people that regularly use a LBSN (e.g. Foursquare `http://foursquare.com/`) to record all their day-to-day activities and movements, which we call trails, enriched by user generated or existing data. By having a collection of these users we can analyse individual and community behaviour in corroboration with geo-spatial, social and temporal factors. We can thus find out that, for example, people from one part of the globe use to go to bed earlier, or that they travel larger distances to relax on weekends during the summer rather than winter.

Our goal is to study the volatility and patterns of human movement in association with the type of venues at different time scales (e.g. time of day, day of week) and prove their effectiveness in applications. Although empirical evidence on LBSN data has illustrated some of these regularities, incorporating them into applications was limited [7] or not successful [10].

The main benefit of studying user trails from LBSNs is the richness of the generated data and the venue-oriented nature of it. Rather than only providing us with the geo-coordinates of users over time, this also includes metadata like the name of the venue, its type, comments about it or even photos. Most previous research into uncovering human mobility patterns was lacking semantics as they were mostly gathered by mobile phone coordinate tracking. Other experiments, although insightful, were limited in the number of users, their backgrounds and the variety of the metadata.

This paper makes 3 key contributions. First, we present a dataset gathered from the most popular LBSN, Foursquare, and consisting of around 10,000 full individual user trails gathered over a month. These users were chosen so as they are frequent users of the service that register all their important changes in locations over the course of a day. This method of data collection offers us trails of a wide variety of subjects in a cheap and easily extensible way. A descriptive analysis shows us that the dataset has all the proprieties found in previous human mobility studies on data from other sources [9]. We use this dataset to study how different types of venues are visited over the course of the day or the week. Through relevant graphs we show that regularities exist, most of which correspond to popular belief. The dataset and scripts are provided freely on the author's webpage.

Next, we contribute to two applications by focusing on the individual transition distributions. Markov Models have been shown [18] to yield good results when studying mobility

WebSci'13, May 2-4, 2013, Paris, France

ACM 978-1-4503-1889-1/13/05... $15.00.

patterns because they make use of the regularities in people's transitions between venues. First, we look at clustering users based on their behaviour. Using k-means clustering we can assign users to categories and represent their typical behaviour. Secondly, the problem of predicting the future transition of a user is studied. Markov Models have the downside that they can't represent regularities at different time scales. We thus create a simple frequency model based on the temporal patterns observed in the analysis section. Although the time length of the dataset is quite limited, we find that this method outperforms the Markov Models. This result is encouraging because it suggests that creating models and classifiers that use a better representation of time and of temporal regularities can lead to significant performance increases for this task.

The rest of the paper is structured in the following way. In Section 2 we review previous work on studying human mobility patterns as well as in analysing LBSN data. In Section 3 we present our novel method of collecting information about frequent users from LBSNs and present the dataset we have gathered. We then study the general characteristics of our dataset in Section 4 and relate them to previous work, verifying thus that our source of data follows the same underlying rules. Our analysis that includes location type information in studying mobility patterns is presented in Section 5 while work on transitions and their applications are presented in 6. We conclude and suggest future work in Section 7.

2. RELATED WORK

Studying the underlying patterns of human mobility has been the subject of study for many researchers over the years. The main bottleneck for this work was data availability. This is largely a consequence of anonymity and personal security.

Researchers have previously studied mobility patterns using proxies to movement such as U.S. banknote movement [2] and marine predators [16]. More recently, the development and availability of portable devices like mobile phones made tracking of peoples location easier. For example, [9] introduced a dataset consisting of 6 months of data from 100,000 users. This dataset contained the locations of the user's closest mobile phone tower every time they made a phone call. This allowed the researchers to have approximate data of each users location within a certain margin in time. Using this data, [17] presents an analysis of the predictability of human movements. Methods for predicting future transitions of users have been surveyed in [3], while newer methods for prediction in social networks are presented in [5, 8].

LBSNs are a relatively new type of OSNs in which users share their current location. One of the first studies on why and how people use these LBSNs is presented in [11]. An empirical study on LBSNs is conducted in [12] while a study on the socio-spatial proprieties relating to LBSNs is presented in [15].

A more in depth study on the properties of LBSNs with respect to mobility patterns was presented in [4]. Here, the general properties and dynamics of the users are presented and are shown to be similar to those from previous research having other sources of data. However, the dataset used was obtained by sampling (10%) of the entire stream of locations and taking the usage of all the users. While this is valid when analysing general proprieties, an in depth analysis at a user level is untenable.

The strong point of using LBSNs for studying mobility patterns is that it provides metadata of different modalities (text, photos, etc.) associated to every venue. This venue information has been exploited so far for different purposes like, for example, characterizing neighborhoods [6, 13] or for location and activity recommendation [21].

Temporal differences between venues have been highlighted with applications to automatic venue tagging [19, 20]. However, when integrated in an application that ranked locations based on a users check-in history, temporal features were shown to be irrelevant [10]. In contrast, [7] studied patterns using a limited set of users coming from a similar background and only 3 venue categories ('Home', 'Work', 'Other') with the aim of identifying 'typical' types of behaviours across users and for clustering. In this paper we suggest a better compromise for characterising venues that both keeps the structure of behaviour and deals with the problem of venue sparsity in a users' history.

3. DATASET

The dataset on which our study is undertaken is assembled from data of LBSNs. These are OSNs that are focused on sharing the current user's location and have at their focal point the venue, a user-specified real-world place with geo-spatial coordinates. In addition to these, venues have additional user-sourced metadata like a venue name, one or more categories (from a predefined list), tags, tips, comments, photos, etc. Users interact with the system by performing a *check-in* at one of these venues. These *check-ins* are performed using a GPS-enabled device and can only be registered when the device is near enough to the desired venue as a form of verification. Like in any other OSN, users of these services are connected with others in a network in which they are able to interact.

We first start by briefly presenting the source of our data, the most popular LBSN, Foursquare in Subsection 3.1 while we present our method of data collection and dataset in Section 3.2.

3.1 Foursquare

After it has seen a boom in activity and awareness immediately after its founding in late 2009, Foursquare, the most important LBSN, has now reached maturity. With this, a target audience that are regular and frequent users of the service has formed.

We collect data from Foursquare because besides its the largest LBSN, it also provides complex venue information of different modalities and its game-like aspect encourages users to check-in as often as possible and repeatedly in the same place. Foursquare recently announced that they have registered 20 million users that totaled 2 billion check-ins in the system's history[1]. The last official estimate of the number of daily check-ins was 3 million, we can thus conclude that on average, every user checks-in at any venue less than once a day. With this in mind, we conclude that in order to study the properties of individual user mobility patterns we must concentrate on users that use the system regularly.

The data that arises from the use of these applications

[1] `http://venturebeat.com/2012/04/16/foursquare-20m-users/`

No.users	No.check-ins	No.check-ins/user
9167	959,122	104.6 ± 49.4

Table 1: Dataset statistics

is subject to serious privacy restrictions and concerns. Due to this reason, the check-in data of a user is not directly available to the outside world. Users choose with what other users to connect ('friends') and only these have access to the recent locations of this user, but without being able to look into their entire history.

3.2 Gathering frequent users

In this section we will present our novel method of finding frequent users of the most popular location based social network, Foursquare, and collecting their activity. We define these *frequent users* as those who use Foursquare at least 3 times per day. We designed a novel method of extracting all the check-ins for some specific users using both the Twitter and the Foursquare APIs.

Our approach considers only users that choose to push their location sharing information on other public social networks, specifically Twitter. We note that this set of users may not be representative for the entire Foursquare user base or for the population in general, but it will still span multiple types of users that differ in behaviour, location and age group. We conduct the analysis in Section 4 in order to see if our dataset presents the same characteristics in terms of group behaviour as datasets from other sources.

Unlike previous studies, we have avoided bulk collection of check-ins from Twitter focusing on a user-centered approach. Instead, using the Twitter Streaming API, we identify a number of *frequent users* of Foursquare and then we use the Twitter Search API to collect all the check-ins for those users. The dataset collection interval is 31 August 2011 - 1 October 2011.

The dataset statistics are presented in Table 1. We should point out that the average number of check-ins/user/day is more than 10 times the number presented in other studies [10, 4]. In [4] the authors specify that 72% of the users have less than 100 check-ins in a 5 month interval. While this does not invalidate the results of their analysis, performing a study on individual user patterns is only possible using full trails of frequent users.

The dataset, together with all the scripts used for collecting the dataset are freely available online[2]. Due of privacy issues, we have anonymised our dataset by stripping user and venue ids. As services like Foursquare are are growing in popularity we can soon expect to obtain more data for a higher number of users over a longer time span.

4. DATASET PROPRIETIES

In this section we will study the proprieties of our data from LBSNs. We investigate the dataset by extracting information relating to the spatio-temporal patterns we observe. We compare our dataset statistics with previous findings in studying human mobility patterns [9] and show that our data confirms their results, even if we use a different data source.

[2]http://dcs.shef.ac.uk/~daniel/foursquare/

4.1 Time distribution

We start by first analysing the time distribution of check-ins. A plot of the number of check-ins for the entire dataset is presented in Figure 1.

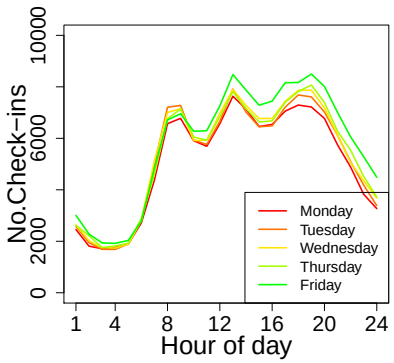


Figure 2: Weekly frequencies

As our data is collected during a month, we could expect to find regularities in the size and time of the total number of check-ins. We observe that, in general, there is a consistent weekly pattern of activity, with weekdays having each three peaks during the day and weekends with a smoother distribution and generally lower activity. Transitions between days can be observed by very low activity during nights as we would have expected. We also notice a slightly upward trend in the number of check-ins in weekdays as the week progresses.

Analysing in more detail, we plot in Figure 2 the check-in frequency per each weekday aggregated over the 4 fully observed weeks in our dataset. We notice the same daily period of check-ins that spikes at 3 times: at 9am when people start their day, at noon when most of the people have their lunch break and in the evenings when they leave the workplace or go out. Here we observe a trend: whilst check-ins in the morning and afternoon are similar, we notice an increase in activity in the evenings from Monday to Friday. This tells us, that as the week progresses and we get closer to the weekend, people tend to go out in the evenings.

4.2 Interevent times and distances

In this subsection we investigate the proprieties of consecutive check-ins of the same user. We study both the interevent times and distance. The distribution of the former is presented in Figure 3 while the later is presented in Figure 4, both presented in log-log space.

The distribution of interevent times indicate that most of the check-ins are performed within a 2 hour interval from the previous. This is because most check-ins in our dataset are performed at transport hubs, shops or food outlets where people don't spend much time. The frequencies decrease with increasing time intervals, with the exception of a small plateau at around 8-10 hours. This represents the usual length of the workday. The interevent distance fits a power laws of the form $x^{-1.56}$, with very few consecutive check-ins more than 100 km away.

4.3 Venue frequency distribution

We now study the distribution of check-ins/venue for users. This is for each user the distribution of frequencies for every

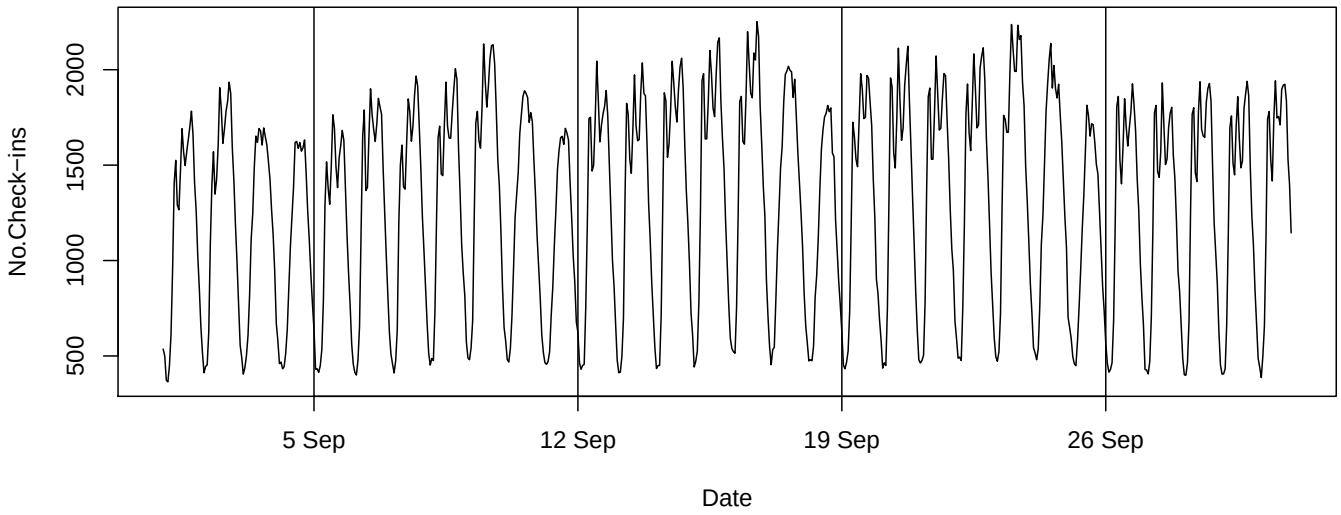


Figure 1: Time distribution of check-ins over a month. The start of the week is indicated by vertical lines.

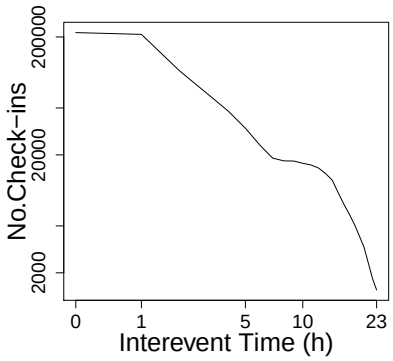


Figure 3: Distribution of interevent times

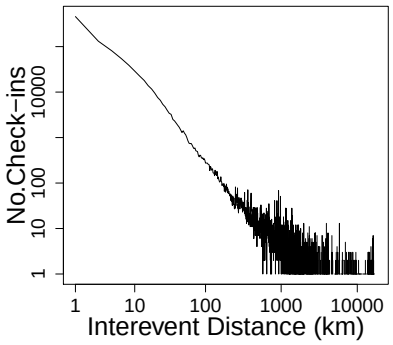


Figure 4: Distribution of interevent distances

venue that he has checked in. We aggregate all the distributions corresponding to every user in our dataset and show the results in log-log space in Figure 5.

We see that most users have a few places they visit very often (the peak at 10-20% of their total check-ins) and many places they seldom visit. It has to be noted that while it is likely that users would get 'bored' of performing check-ins in the same place they visit frequently, the game elements of Foursquare which offers mayorships keeps users motivated to continue checking in.

4.4 Returning frequency

Now, we examine the periodic patterns in human mobility by measuring the returning frequency at a venue. The returning frequency is the number of times a user returning to a place he visited h hours ago. This is presented for our dataset in Figure 6.

We notice spikes for daily intervals, which diminish rapidly for the future days. We also notice a strong weekly return probability. This behaviour is similar to the one reported in [9] who tracked mobile phone positioning. However, when comparing it to a similar figure from [4] that uses the same data source as us, we observe that the downward trend in the returning probability is diminished in their dataset. We can explain this by the fact that the authors used a sample of the underlying distribution. The returning probability gets 'spread' in future days, smoothing the distribution.

We also compute a *returning frequency ratio*. This measures the number of times a user returned to a venue after h hours divided by the times he was observed after h hours. Because we consider only days when users register more than 3 check-ins, we don't consider these days as observed and adjust the ratios accordingly. The distribution of returning frequency ratios is presented in Figure 7.

From this figure we observe that the highest returning ratio is the weekly one. High values are obtained for daily returning ratios as well. This is a interesting result on its own, showing that it is more likely that one visits the same place as a week ago than a place that one visited the day before. A reason for that may be the different activities one does on weekdays and weekends which affect the daily pattern. As a consequence, in the next chapters we will focus on studying and using these daily and weekly patterns.

5. EXPLORING VENUE INFORMATION

In this section we will focus on exploiting one of the characteristics of our dataset which consists of the metadata attached to the venue associated with the user's check-in. We specifically focus on the type of venues and analyse them from a temporal perspective, extracting general patterns. Observing such patterns can aid understanding human mo-

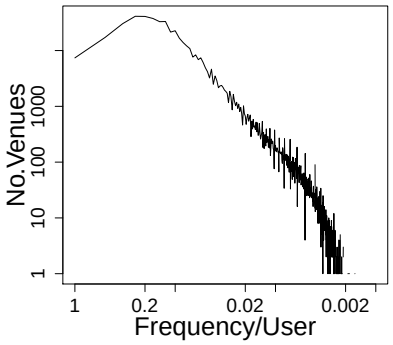


Figure 5: Venue Frequency Distribution

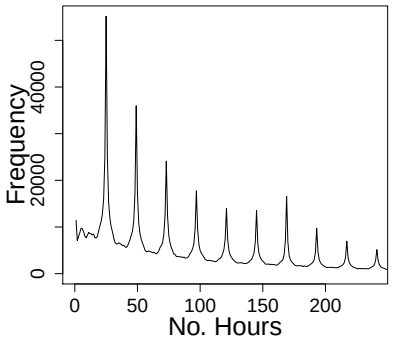


Figure 6: Returning frequency

bility with an immediate application in predicting human movements.

We highlight that although our paper focuses on studying the venue type information, our data is richer and we can perform regional analysis, use the text sentiment and explore social factors of the users. We will leave these for future research.

The metadata associated with each Foursquare venue includes the venue type. These venue types are organized in a hierarchy with 3 levels. In our study we only use the most general layer of the hierarchy which contains 9 venue types: 'Arts & Entertainment', 'Travel & Transport', 'Shop & Services', 'Food', 'Great Outdoors', 'Nightlife Spot', 'Residence', 'College & University' and 'Professional & Other Places'. The next layer of the hierarchy contains 259 venue types, which is too many for statistical and visualization purposes.

Each venue can have multiple categories. In the rest of the paper we use only the 'primary' category associated with every venue.

5.1 Check-ins and categories

First of all, we look at the total number of check-ins for each category over our entire dataset. This is presented in Figure 8. We observe that the categories with the highest number of check-ins are 'Food" and 'Shops & Services", 'Travel & Transport' and 'Professional & Other Places' have high percentages as well, which mostly indicate that people use the service when traveling to different places and when they arrive at their office. 'Residence' has a significant percentage of check-ins as well, indicating that many users from our dataset check-in at their homes, probably when they leave or arrive. 'Nightlife Spot' and 'Arts & Entertainment' have low percentages as we would expect from a dataset that

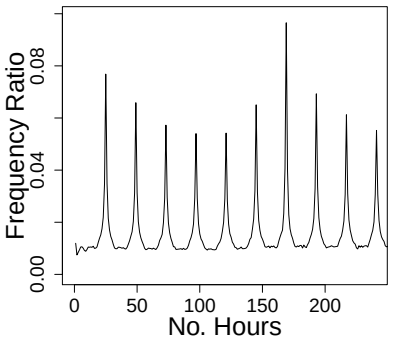


Figure 7: Returning frequency ratio

models regular day-to-day behaviour. When analysing LBSNs, there was a concern that many people use them only to show to their social network where they are going out. This appears not to be the case, which we ascribe to our filtering for frequent users who use the service more frequently and for a more general purpose.

Next, we examine the time distribution of the number of check-ins for each category. We present the distribution of check-ins on the entire month for the categories 'Professional & Other Places' and 'Nightlife Spot' in Figure 9 and Figure 10. Vertical lines indicate the beginning of the week and for Figure 10 the red lines indicate Friday and Saturday midnight. The low frequency sections indicate nights.

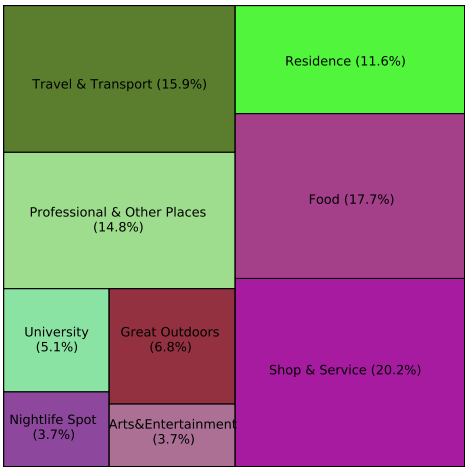


Figure 8: Check-in frequency per venue category

The results are similar to what we would have expected. The category 'Professional & Other Places' consists mostly in office buildings. Activity here increases during weekdays with a low activity in weekends. During the days of the week, check-in patterns are consistent, with a spike during the morning, when people arrive at work and with a smaller one after lunch, had the users went out for lunch and returned back to work. The only weekday in our dataset with lower activity than the others is the 5th of September. While this might look surprising at first, actually the day corresponds to a holiday ('Labour Day') in the U.S., where many of our users are based.

For the category 'Nightlife Spot', we observe a different pattern. While generally, activity is low during the day and spikes in the evenings, we also notice a trend. Activity is lowest on Sundays and Monday and starts growing until

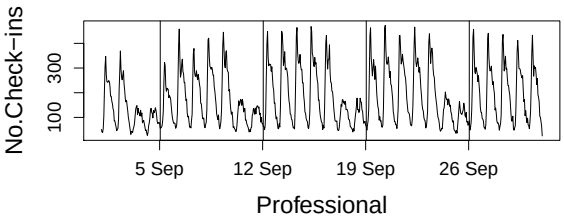


Figure 9: Check-in diagram for venue type Professional

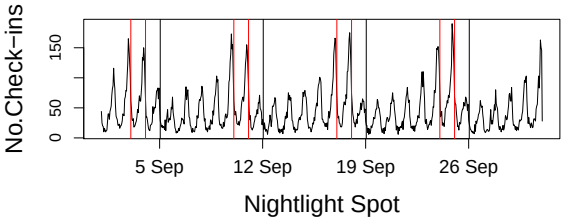


Figure 10: Check-in diagram for venue type Nightlife

peaking on Friday and Saturday nights. This shows us that during weekdays, people tend to go out more in the evenings as the week progresses, with the weekend nights reaching a high.

We highlight that despite the number of check-ins being relatively low (at most 500 check-ins/hour for a category) the results present significant consistency in periods and trends over the entire month. This gives us another indication of the quality of our dataset.

5.2 Daily patterns

In this section we will examine the frequency of check-ins to venues in different categories when looking at a specific days of the week. We chose to analyse activity on Saturdays as well as activity on all the weekdays combined as they show similar patterns for most of the categories. We present the check-in frequencies for each venue type in Figure 11a and Figure 11b.

From the graphs we observe interesting patterns. Check-ins into 'Professional & Other Places' are very frequent on weekdays and especially during the morning, when people arrive at work. This influx is preceded by an increase in check-ins to venues from the 'Travel & Transport' category. As we would expect, 'Colleges & University' follow the same trend as the 'Professional & Other Places' only at a smaller scale both on weekdays and Saturdays. The 'Residence' category has the most check-ins in the morning (people wake up and check-in to their house) and in the afternoon (after 6pm) when people arrive home from work. For this category, the distribution is smoother on Saturdays but follows the same pattern. The 'Shops & Services' category has a consistent pattern as well, with increasing activity after lunch and decreasing as the night approaches. On Saturdays, 'Shop & Service' is the most important category during the day. As the night approaches, 'Nightlife Spot' become more popular, as during the day their frequency was the lowest.

While these graphs show us the number of check-ins at a category in a point in time, we are interested also in how long do the people actually stay at the venue they checked-in. We would expect that the venue type will influence how much time people actually spend there before moving to a new destination.

We thus assume that between two consecutive check-ins the user is located at his last registered location. By this, we would hope to eliminate the check-ins with high frequency and low interevent times and focus more on where the users are at a certain point in time. If many check-ins are registered in the same hour interval, we consider for that hour the category of the first check-in and for the future hours with no check-ins the category of the last registered check-in. The results of our analysis are presented for weekdays in Figure 11c and for Saturdays in Figure 11d.

We notice a number of dissimilarities from the previous graphs. First of all, all the distributions are smoother, with fewer abrupt changes. For the 'Professional & Other Places' category, we notice a burst at the hours of the morning and a slow decay afterwards. Correlating with the results for the 'Residence' category, we observe that some of the people don't check-in back when they get home. This is indeed a problem with our dataset, as most users check-in their homes in the mornings but don't check-in when they arrive from their activities. For 'Residence' we observe that the proportion of people that stay home on Saturdays is larger than the one during weekdays, as we would have expected. Also, the 'Travel & Transport' category, which had registered a high number of check-ins has now a diminished importance. We would expect that users spend much less time at these types of venues than they do at others, like 'Professional & Other Places' or 'Residence'.

5.3 Interevent times

As we have discovered from the previous section, we expect that people are likely to spend different amounts of times depending on the type of location they are in at the moment. To confirm this hypothesis we analyse interevent times based on the venue category. A plot that shows the interevent times for each of the categories is presented in Figure 12.

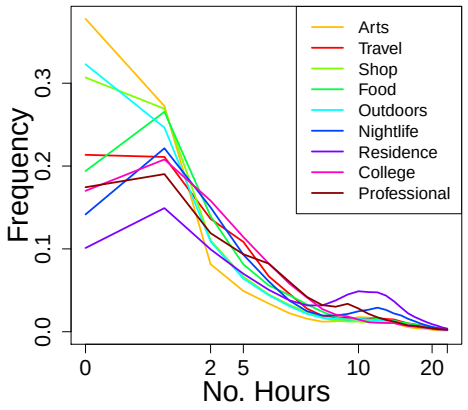
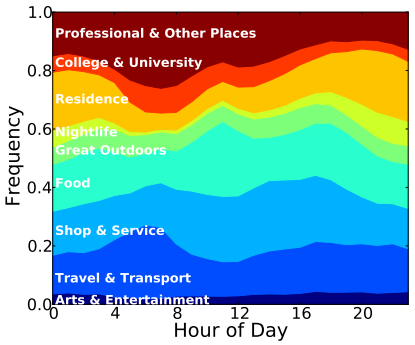
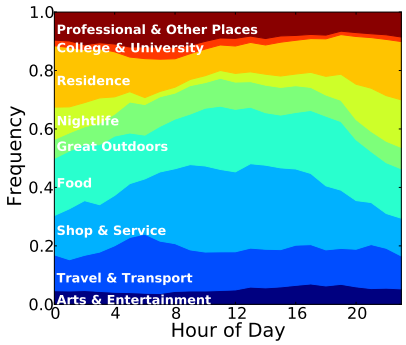


Figure 12: Interevent time for each category

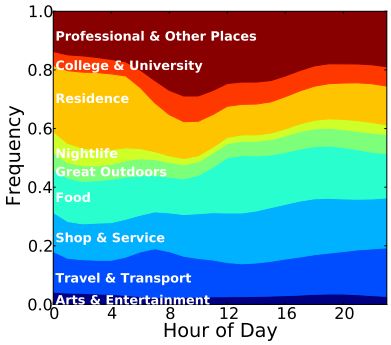
The figure shows us that indeed, the interevent times have different distributions based on the venue category. For example, in the 'Transport & Travel' category, people spend very little time, with almost 40% registering at another venue in less than an hour. This distribution further decays as the number of hours increases. In contrast, for the 'Residences' category, the distribution is much smoother, with relatively high frequencies even for interevent times of



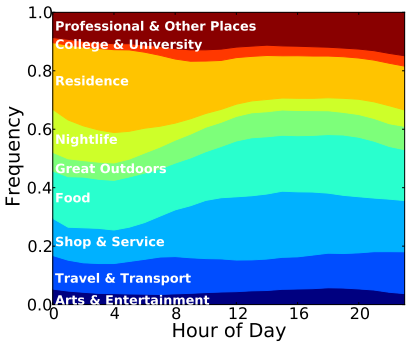
(a) Weekday check-in distribution



(b) Saturday check-in distribution



(c) Weekday location distribution



(d) Saturday location distribution

Figure 11: Location and check-in distributions per category

12 hours. For the 'Professional & Other Places' category, the distribution is smoother, but not as much as the one for 'Residences'. Even though we might have expected the time to peak at 8-9 hours (the length of a usual workday), we must keep in mind that most people go out during the lunch break and thus they get checked-in at another venues, splitting the interevent time. Also, some 'Professional & Other Places' check-ins correspond to people visiting other offices or institutions.

6. ANALYSING HUMAN MOVEMENT

Our study has focused so far on general periodicity and trends of location categories over time. We now switch to a user-centered view by analysing individual mobility trails. This enables us to make use of the patterns over venue categories in applications.

We first look at transitions between venue categories. It was previously shown that Markov models of human behaviour perform well when predicting future locations [18]. This means that there is an underlying structure in transitions. We will then present two tasks performed at the user level. The first is clustering users based on their transition distributions. By using a k-means clustering approach, we are able to cluster users to groups and illustrate their centroid. These centroids or 'typical behaviours' can be then interpreted and assigned to different human categories (e.g. student, stay-at-home). Finally, we use this information in domain-independent methods for location prediction. We will analyse and compare these with methods that explicitly take into account information about the periodicity of check-ins at different timescales (e.g. day of week, time of day).

6.1 Transitions

We present a heatmap of the transition probabilities from one venue category to another, aggregated over all the users in the dataset, in Figure 13. The transitions take in account the order, with the source being presented on the vertical axis and the destination on the horizontal axis.

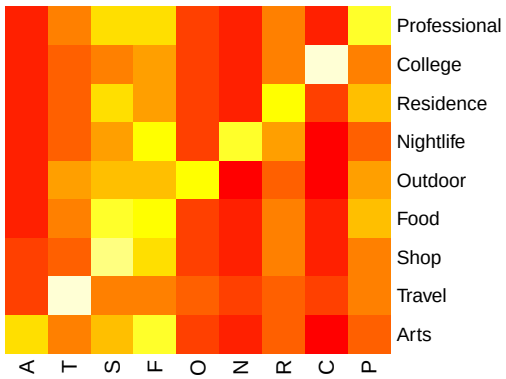


Figure 13: Transition diagram between categories (lighter represents higher values)

At first we notice the high frequency on the principal diagonal. This means, for most of the categories, that if a user is at a certain type of location, it is very likely for him to

transition to a place belonging to the same category. This behaviour is most predominant for the 'Travel & Transport' and 'College & University' categories. This is expected, as people often visit many transport locations before reaching their destination and, if they are at a University building, it's very likely that they'll visit another building from the campus next.

We also notice the lack of symmetry in the heatmap. This highlights that there are some transitions that are more likely to happen in one direction rather than the other. For example, it's very likely for people to go to eat (category 'Food') after an artistic event (category 'Arts & Entertainment'), while the opposite is less common. Other examples include the high probability for people to visit shops or to eat after work and in the lunch break or to go home after going out in the evenings. On the other hand, we also observe some reciprocal ties, like the one between food and shopping activities, with all transitions with high values.

6.2 Behavioural clustering of users

By analysing the transition matrix for the entire dataset we discover general patterns of user behaviour. We now focus on the mobility patterns of individual users to identify different types of users.

For each user, we build its transition matrix of counts for the venue categories. Then, we cluster the users into groups based on this matrix. For this purpose, we use k-means clustering [1] using the Frobenius norm to compute the distance between matrices. We first scale the matrices so that rows and columns sum to unit length. This way, the algorithm is not affected by different magnitudes in the data for each user. We have ran the algorithm with the number of clusters K empirically set to 8. We show the centroids in Figure 14.

Analysing the centroid of the clusters we observe the different types of users. Many users are of a 'generic' type that goes to offices and work often, but also visits shops and transport locations (Cluster 4). The rest of the activities are spread out over all categories. Besides this generic user, we identify clusters of specific categories of people. Cluster number 3 for example consists of persons who check-in regularly to work and not often to other venues *i.e.* 'workaholics' or 'businessmen'. We can associate cluster number 5 with 'stay-at-home' persons that check-in most frequently at residences and then at food and shop locations. Cluster number 8 represents students, with the rest of venues besides 'College' with significantly lower transition probabilities. We highlight that we obtain these general patterns of behaviour without using any supervision or world knowledge.

6.3 Predicting future user movements

Although [7] attempts to empirically show some temporal patterns, integrating them into prediction models has not been very successful [10]. This can be because the historical information was considered in conjunction with individual venues [8]. Very sparse representations are obtained and, as we collect more user data, the number of observed venues increases over time leading to a decrease in prediction performance. As also highlighted in [10], we must find a proxy for these venues that can be used for prediction. We use the category type information as this proxy, adding some semantics to user transitions. The advantage of using category information is that it is fixed to a limited number (i.e. 9), alleviating sparsity concerns, but still captures individual preferences and patterns.

We will use domain-independent methods for prediction. These methods take into account for prediction only the previous history of transitions of a user. The most common class of methods of this type is the Order-K Markov model, where we use as context the last K transitions of the user and find the most likely location where the user will transition next based on this context. This method allows us to learn simple transition rules for every user and predict the future location accordingly (e.g. after work a person will go to a food vendor, after a restaurant and a bar the person will likely go home or to another bar).

We want to integrate our knowledge about existing temporal patterns related to venue types. We have observed there some very strong patters for returning probabilities for every user. The strongest were the daily pattern, where a user visits the same category type as the day before at the same time, and the weekly pattern, where a user transitions to the same venue category as the same day and time as one week before. We want to build these explicitly into a simple model in order to assess their predictive power.

Although other domain-independent methods exist, previous studies [18] have shown that the improvement over the simple Markov based methods is marginal and, as our focus is on testing if incorporating the temporal returning patterns is beneficial, we chose not to implement them.

The methods that we test are the following:

- **Most Frequent Category** We assign to each testing instance the category that was most frequent in the users' history. Note that this is equivalent to an Order-0 Markov predictor.

- **Order 1 Markov** We assign to the testing instance the most frequent category based only on the last visited venue and its category. In case the transition was not observed in the users' history, we backoff to the Most Frequent Category.

- **Order 2 Markov** We assign to the testing instance the most frequent category based on the last 2 visited venues. In case the transitions were not observed in the users history, we backoff to the Order 1 Markov model.

  The probability of the next check-in c_{n+1} at location l with an Order K Markov model is:

$$P_k(c_{n+1} = l|H) = P(c_{n+1} = l|c_{n-k+1}, ..., c_n) = \frac{|\{c_r|c_r \in H, c_r = l, c_{r-j} = c_{n-j+1}\}|}{|\{c_r|c_r \in H, c_{r-j} = c_{n-j+1}\}|} \quad (1)$$

  where $H = \{c_1...c_n\}$ is the history of the previous check-ins.

- **Most Frequent Hour** We assign to the testing instance the category that was most visited by the user in the same hour of the day in it's history. This model assigns the probability of the next check-in c_{n+1} at location l at time h as the probability of the location l occurring at time h in the previous check-in history.

$$P_{MFH}(c_{n+1} = l|H, t_{n+1} = h) = \frac{|\{c|c_r \in H, c_r = l, t_r = h\}|}{|\{c_r|c_r \in H, t_r = h\}|} \quad (2)$$

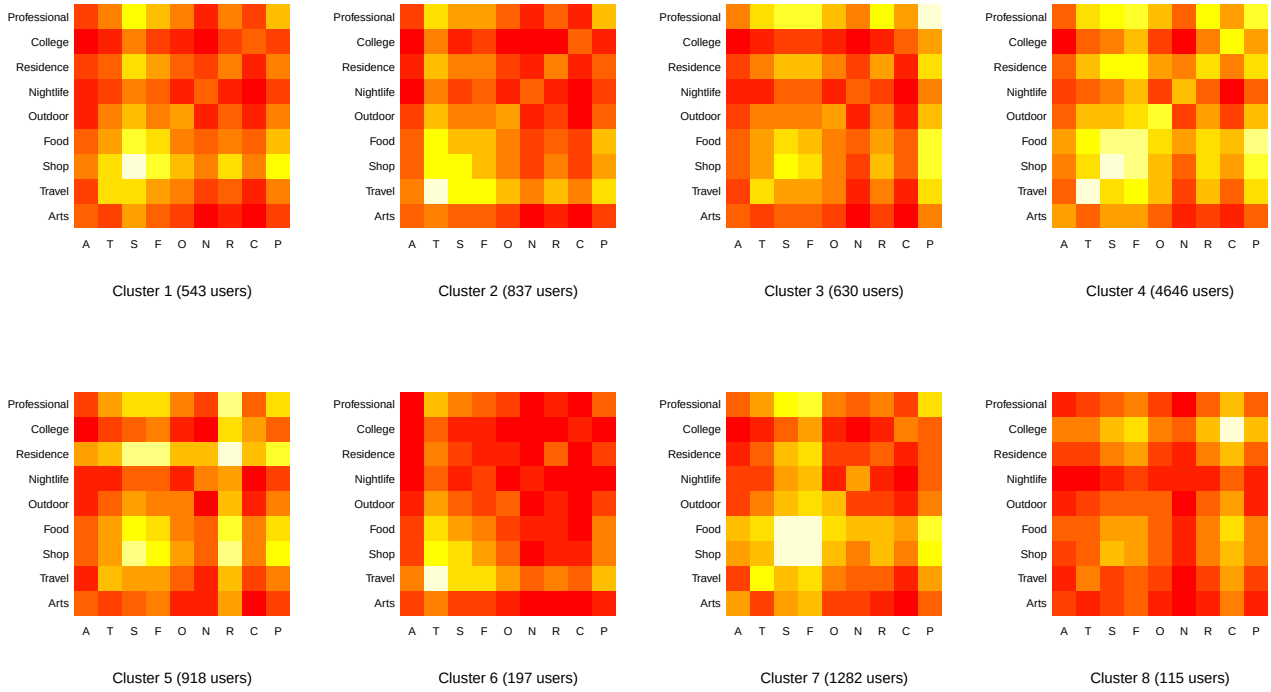


Figure 14: Centroids of k-means clustering on user's venue category transition matrices

Method	Accuracy
Random Baseline	11.11%
Most Frequent Category (Markov-0)	35.21%
Markov-1 (with backoff to Markov-0)	36.13%
Markov-2 (with backoff to Markov-1)	34.21%
Most Frequent Hour	38.92%
Most Frequent Day of Week and Hour	40.65%

Table 2: Accuracy of different methods

- **Most Frequent Day of Week and Hour** We assign to the testing instance the category that was most visited by the user in the same hour of the same day of the week in the training set. The probability of the model is similar to that of MFH, but conditioning in addition on the day of week.

For this task we use the following testing scenario. We train our models on the first three weeks (21 days) of data and test on the remaining days (9 days). After we obtain an average prediction percentage for each user, we then macro average the results and report this as the performance of the method.

The results are shown in Table 2. We observe that the results match our intuitions. We notice that the MFC baseline is quite high when compared to the other models. Markov-1 with backoff performs slightly better which confirms that some information is gained by looking at the previous step of a transition. This improvement is however not that significant, indicating that user movement is governed also by other factors and patterns. Markov-2 performs slightly worst than the baseline. This confirms the previous results of [18].

Incorporating explicitly the periodicity of user behaviour shows to give the best results. We observe that the Day of Week pattern is strongest and obtains the best results, even if the history for a user is very restricted, basically to only 3 weeks of previously observed data. Actually, for almost half of the cases, we have used the backoff, which shows that the Day of Week and Hour of Day prediction is very effective when applicable. This method improves on all the Order-K Markov methods, showing that periodicity is a factor that has to be taken in account when studying human mobility patterns.

In all models, the overall macro-average is under 50%. The task of predicting future movements in the case of users which are not restricted to an age group, profession or geographical area is shown to be much harder then when using subjects that have the same characteristics such as lab students [7]. Filtering first on this behaviour by using a method such as the one presented in the previous subsection can also aid our predictive models. Some users have very different behaviours (e.g. holidays) in the testing period. For example over 8% of users have under 10% accuracy for the Most Frequent Category baseline, meaning that these users have checked-in to their most frequent venue category from the training period less than random in the test set.

The results show us the need for prediction algorithms that take into account temporal periodicities over different time intervals. With a better representation of time, they should be able to go beyond the sequence assumption for prediction and quickly switch between types of user behaviour. This will be subject of future work.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have studied human mobility patterns combining the temporal information about the whereabouts of users with information on the types of places they visit. We have discovered interesting associations (*e.g.* very strong weekly patterns, increase of activity as the week progresses), most of which relate to our real world expectations.

We have shown that our data, derived from LBSNs, is in tune with previous findings in studying of mobility patterns using other data sources, such as mobile phone usage. Our data provides the added benefit of having semantically annotated venues and represents general behaviour of a large number of users from diverse backgrounds.

Adopting a user-centered view, we have analysed the transitions of users and studied two different applications. First, we shown how we can cluster users based on their behaviour. Then, we demonstrated by employing domain-independent predictors of human mobility that by folding observed periodicities into a simple model we can better predict human mobility, even when provided with a restricted user history.

Future work will look more into the problem of location prediction using models that can better incorporate temporal patters across different timescales as well as using continuous time rather than a discrete representation, such as Gaussian Processes [14]. Also, we will look more into using other types of information associated with the venues, such as comments or photos.

Acknowledgement

This research was funded by the Trendminer project, EU FP7-ICT Programme, grant agreement no.287863.

8. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[2] Brockmann, L. Hufnagel, and T. Geisel. The Scaling Laws of Human Travel. *Nature*, 439(7075):462–465, Jan. 2006.

[3] C. Cheng, R. Jain, and E. van den Berg. Location Prediction Algorithms for Mobile Wireless Systems. pages 245–263, 2003.

[4] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM 2011.

[5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and Mobility: User Movement in Location-Based Social Networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2011, pages 1082–1090.

[6] J. Cranshaw and T. Yano. Seeing a Home away from the Home: Distilling Proto-neighborhoods from Incidental Data with Latent Topic Modeling. In *Proceedings of the Workshop on Computational Social Science and the Wisdom of Crowds, NIPS 2010*.

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

[8] H. Gao, J. Tang, and H. Liu. Exploring Social-Historical Ties on Location-Based Social Networks. In *Proceedings of the Sixth International Conference on Weblogs and Social Media*, ICWSM 2012.

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding Individual Human Mobility Patterns. *Nature*, 453(7196):779–782, June 2008.

[10] D. Lian and X. Xie. Learning Location Naming from User Check-in Histories. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 112–121, 2011.

[11] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the Mayor of my House: Examining why People use Foursquare - a Social-driven Location Sharing Application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2011.

[12] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM 2011.

[13] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *The Social Mobile Web*, 2011.

[14] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[15] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM 2011.

[16] D. W. Sims, E. J. Southall, N. E. Humphries, G. C. Hays, C. J. A. Bradshaw, J. W. Pitchford, A. James, M. Z. Ahmed, A. S. Brierley, M. A. Hindell, D. Morritt, M. K. Musyl, D. Righton, E. L. C. Shepard, V. J. Wearmouth, R. P. Wilson, M. J. Witt, and J. D. Metcalfe. Scaling Laws of Marine Predator Search Behaviour. *Nature*, 451(7182):1098–1102, Feb. 2008.

[17] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, 2010.

[18] L. Song, D. Kotz, R. Jain, and X. He. Evaluating Location Predictors with Extensive Wi-Fi Mobility Data. *SIGMOBILE Mob. Comput. Commun. Rev.*, 7(4):64–65, Oct. 2003.

[19] M. Ye, K. Janowicz, C. Mülligann, and W.-C. Lee. What you are is when you are: The Temporal Dimension of Feature Types in Location-Based Social Networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 102–111, 2011.

[20] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the Semantic Annotation of Places in Location-Based Social Networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2011, pages 520–528.

[21] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative Location and Activity Recommendations with GPS History Data. In *Proceedings of the 19th International Conference on the World Wide Web*, WWW 2010, pages 1029–1038.