Predicting User Locations and Trajectories

Eelco Herder, Patrick Siehndel and Ricardo Kawase

L3S Research Center, Leibniz University Hannover, Germany
{herder,siehndel,kawase}@L3S.de

Abstract. Location-based services usually recommend new locations based on the user's current location or a given destination. However, human mobility involves to a large extent routine behavior and visits to already visited locations. In this paper, we show how daily and weekly routines can be modeled with basic prediction techniques. We compare the methods based on their performance, entropy and correlation measures. Further, we discuss how location prediction for everyday activities can be used for personalization techniques, such as timely or delayed recommendations.

Keywords: GPS, Geolocation, Mobility Patterns, Personalization

1 Introduction

Location-based services suggest new locations that match the user's inferred interests and preferences, making use of content-based or collaborative recommendation techniques. In most cases, distance is used as the main criterion for inclusion in the recommendations. As argued by Mokbel et al. [14], location-based services usually only take the current location into account. However, apart from visiting new locations, users often visit places that they visited before [13]. These revisited places include home and work locations, but also less frequently visited places, such as specialty stores, hiking areas, friends and relatives.

Several studies confirmed the intuition that human mobility is highly predictable [9, 16], centered around a small number of base locations. This opens a wide range of opportunities for more intelligent recommendations and support of routine activities. Such recommendations may serve as reminders for activities or locations to be included in the user's schedule, and may be used to minimize traveling time between the destinations that a user is likely to visit.

In the literature, one can find only a few studies on common travel patterns, or on locations that are typically visited on certain hours during the week or during the weekend. Such insights are expected to be useful for selecting techniques for predicting a user's travel activity and likely destinations. In this paper, we analyze, visualize and discuss patterns found in a dataset of GPS trajectories. Further, we compare and analyze the performance of common prediction techniques that exploit the locations' popularity, recency, regularity, distance and connections with other locations.

The remainder of this paper is structured as follows. In the next section, we discuss background and related work. Then, we describe the dataset that we used, the preprocessing steps for identifying travel sequences, visited locations, and the likely purpose of locations. Subsequently, we show regularities in user travel activities, discuss the nature of different locations visited during weekdays and weekends, followed by a comparison of the performance of various common prediction techniques. We conclude the paper with a discussion of implications and opportunities for personalization and recommendation.

2 Background and Related Work

In this section, we discuss four strands of related work. First, we summarize the main insights from several studies on general and individual mobility patterns, followed by a number of studies that aim to predict next locations. Then, we continue with a brief discussion on the role of locations in popular social media services. We conclude with an overview of location-based services.

2.1 Human Mobility Patterns

González et al. [9] studied people movements, based on a sample of 100,000 randomly selected individuals, covering a six-month time period. The results show that human mobility patterns have a high degree of spatial and temporal regularity. Further, individuals typically return to a few highly frequented locations and most travel trajectories are rather short in terms of distance and travel time.

Song et al. [16] found that 93% of human mobility is predictable; how predictable an individual's movements is, depends on the *entropy* of his patterns. However, for predictability it did not make a difference whether an individual's life was constrained to a 10-km neighborhood or whether he travels hundreds of kilometers on a regular basis.

Zheng et al. [19] used GPS data for mining interesting locations and 'classical sequences', based on the number of visits and the individual visitors' location interests. The outcomes are reported to be useful for tourists, who can easily discover landmarks and popular routes.

2.2 Predicting Next Locations

Ashbrook [2] calculated the probability of transitions between locations, which were extracted from raw GPS data, using various orders of Markov models. The authors discussed the models qualitatively, without mentioning overall accuracy measures.

Krumm and Brush [13] used probabilistic schedules to predict at what times people would be at home or away. The predictive performance of the algorithms was shown to be significantly better than the participants' self-reports. Etter et al. [8] won the Next-Place Prediction task of Nokia Mobile Data Challenge,

making use of a wide range of predictors, including a Dynamic Bayesian Network that models the distributions of location transitions and popular locations on certain days and at certain times. Their models achieved a reasonable performance; still, the authors concluded with the open question whether 'unpredictability [is] mainly rooted in the users' personality' or 'a consequence of the data characteristics'.

The above-mentioned studies provide some insights on the predictability of individual mobility patterns. However, most of these studies did not investigate how this predictability depends on the temporal dynamics in human mobility. Biagioni and Krumm [4] provide some first insights, based on the assessment of the location traces of 30 volunteers. Making use of timeline visualizations, the volunteers indicated which days were most similar to each other. With edit-distance-based similarity measures, they managed to cluster similar days with up to 75% accuracy.

2.3 Location and Social Media

Apart from GPS data, a popular source for the analysis of human mobility is social media data. However, social media data is reported to be sparse: most Twitter users only mention a very generic home location and less than 1% of tweets contains metadata on the location where it stems from [6]. Similarly, data from Foursquare[1], a popular location-based social networking tool for mobile devices, is incomplete as well: Foursquare does not automatically track the locations of users and only registers the users' location when they 'check in' at some place. As argued by [11], Foursquare users typically do not 'check in' places that they consider uninteresting (e.g. home or work) or embarrassing (e.g. fast food restaurants).

2.4 Location-Based Services

Location-based information services are typically provided as recommendations [3] or as contextualized search results [18]. Several surveys show that restaurants and stores are the most popular locations that users search for, followed by local attractions and locations associated with leisure time [3, 18]. As noted before, these services usually provide suggestions for new locations, based on the user's preferences and current location. In a recent study, Amini et al. [1] showed the benefits of trajectory-aware suggestions that are based on the distance to the user's predicted destination instead of the user's current location.

3 Dataset and Tools Used

As a basis for our analysis, we used the GeoLife GPS Trajectory Dataset [2] [19], which contains a total of 17,621 trajectories from 178 users, mainly located in

[1] http://foursquare.com/

[2] http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/

Beijing. The dataset is complemented with the MSR GPS Privacy Dataset 2009 [3] [5], which contains 4,165 trajectories from 21 users, mainly located in and near Seattle, gathered in a 2-month period in 2009.

3.1 Preprocessing Steps

As we are interested in the start and end locations and the durations of the trajectories, we extracted the first and last entry of each trajectory in the dataset; this data was stored as a single entry in the database, representing a trajectory with a start point and an end point - the duration is the difference between the corresponding two time stamps.

Subsequently, the different longitudes and latitudes were merged into (numbered) locations, by comparing the distance of each new start point or end point with the person's previously stored locations. After experimentation with different thresholds (starting with 20 meter, which is reported to be the current precision of GPS [4]), we finally chose a fairly large threshold of 300 meter.

3.2 Estimation of Location Purposes

We estimated the likely purpose of the locations visited by the users by making use of the data provided by Foursquare, a location-based social networking website for mobile devices where users 'check in' at venues. The Foursquare API [5] provides access to all user-generated data, which allowed us to query for venues surrounding a given coordinate.

For each location, we identified venues up to 50 meters away from the location's coordinates. In total, we collected the name and categories from 21,167 venues, covering 4,487 unique locations in our dataset. Obviously, not all locations are associated with venues registered in Foursquare: particularly 'non-popular' sites, such as residential areas, have no nearby venues cataloged in Foursquare. For the GeoLife GPS Trajectory Dataset, we were able to find venues for 49% of the locations, while in the MSR GPS Privacy Dataset we found venues for more than 80% of the data.

4 Analysis of Patterns in Human Mobility

In this section, partially based on earlier work [10], we discuss patterns and regularities that we found in human mobility. First, we describe overall travel patterns on weekdays and during the weekend. Second, we exploit the category labels of the locations to identify their different purposes and to which ones users travel during different hours of the day.

[3] `http://research.microsoft.com/en-us/um/people/jckrumm/gpsdata2009/index.html`

[4] `http://en.wikipedia.org/wiki/Global_Positioning_System`

[5] `http://developer.foursquare.com/docs`

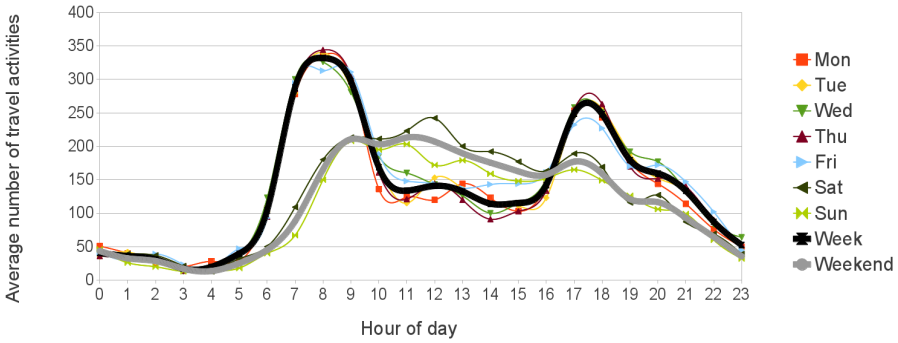

Fig. 1. Daily travel activity during the week and in weekends.

4.1 Overall Travel Activity

Our data shows similar characteristics as reported in [9]: trajectories follow a power-law distribution, with only a few locations that account for the majority of visits and a small number of trajectories that users follow most of the time.

In Figure 1, we visualized the number of trips that started at a specific hour on a certain day or group of days (week, weekend). The thick black line is the average of the five weekdays (Monday till Friday) and the thick gray line averages the weekend days (Saturday and Sunday).

Some strong regularities can be observed. On weekdays, the morning rush hour has a strong peak at 8am; the evening rush hour is more spread between 5pm and 9pm. Between both rush hours, traffic is moderate, with a small peak during lunchtime. During weekends, traffic starts somewhat later and remains relatively stable throughout the day, with a slight increase of traffic just before dinnertime. These differences can obviously be explained by the fact that most people work during the week and use the weekend for spare-time activities.

A further insight of the study was how the different locations are related with one another. For individual users, we visualized the locations and the trajectories between them using the graph visualization toolkit Gephi [6], see Figure 2. The graph layout is force-directed. In the figure, four frequently visited locations can be seen, of which location 0 and 1 are presumably the user's office and home locations; location 2 could be a shopping mall, and location 13 might be a (sport) club (see [10] for more details). A particular observation is that the long tail of other locations is typically only connected to one of these main locations, or shared by two locations (the cluster of small dots between home and office probably represents places that are visited on the commute between home and office). We verified this pattern with various other users with sufficient travel data and found similar graphs.

[6] `https://gephi.org/`

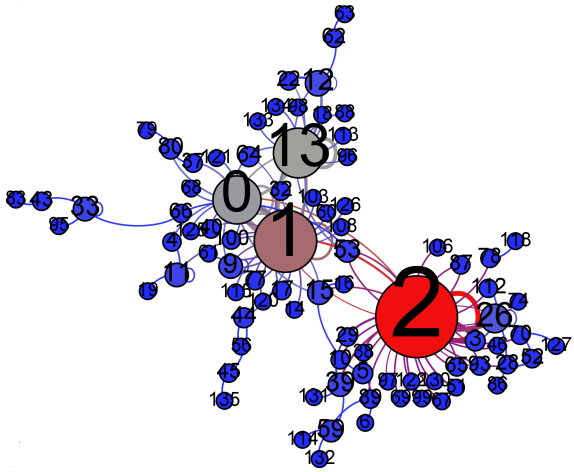

Fig. 2. Connections between locations of an exemplary user. The larger the node, the more often the user visited the corresponding location. The thicker the edge, the more often the user traveled between the two locations.

4.2 The Purposes of End Locations

Further information on typical activities that users engage in can be found by analyzing the purposes associated with typical locations on different times of the day. We extracted these purposes by manually aggregating Foursquare categories into 'purpose groups'. For example, the group 'Food' consists of different restaurant types, such as 'Japanese Restaurant' and 'Fast Food Restaurant'.

In Figure 3, the distribution of end-locations during the day is displayed, with separate graphs for weekdays and the weekend. The graphs are based on 6903 trajectories and corresponding end-locations on weekdays and 3199 end-locations during the weekend.

The distribution of the different category groups over the day shows some clear differences, which match common expectations and the observations from the previous subsection. On weekdays, locations related to transportation (e.g. train stations and bus stations) and work have peaks between 7am and 9am. During the remainder of the working day, travel activity remains low, with a slight peak during the lunch break. Shopping activities take place immediately after the lunch break or during the evening commute. At about 6pm, locations that are associated with food and sports are frequently visited; most people return home after having engaged in typical evening activities.

The distribution of peaks during the weekend is quite different from the weekday pattern. While weekday travel activities have peaks at the start and end of the day, weekend travel tends to be distributed throughout the day - we

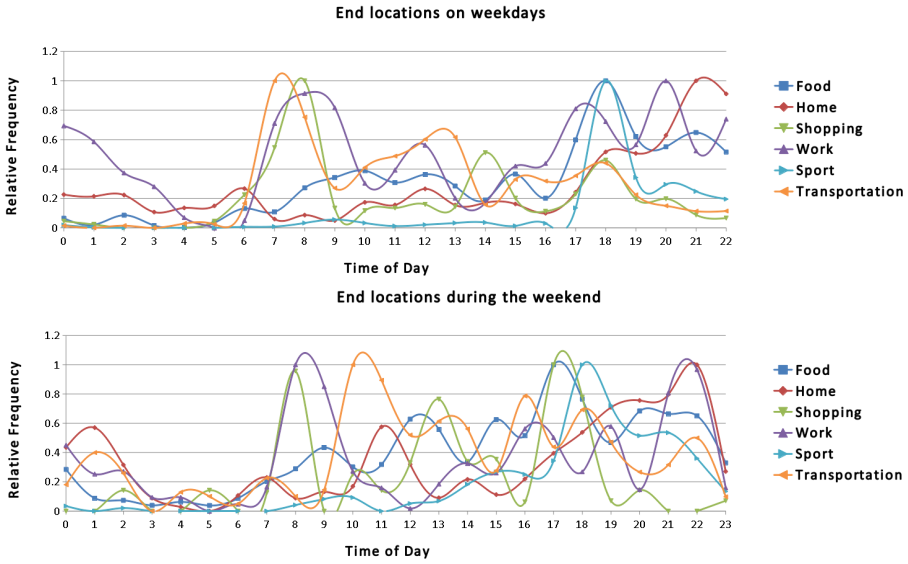

Fig. 3. Distribution over time for different groups of end-locations on weekdays and weekends.

observed the same effect in Figure 1. Shopping activities show the same peaks as on working days, but with a larger emphasis on the end of the day. Similarly, sport activities mainly take place during the evening hours - as on weekdays -, but the spread is wider and includes afternoon hours. Particularly interesting is the distribution of the home category: people often return home at 11am after their morning activities; people who engage in evening activities such as sports often return home at 10pm; the peak at 1am seems to indicate people who return home from a pub or a party - a phenomenon that is not observed on weekdays.

5 Predicting Future Locations

In this section, we use the routine travel patterns, as discussed in the previous section, as a basis for comparing five basic methods for predicting when a person will revisit a particular location. This problem has several similarities with predicting page revisits on the Web, where users also typically revisit only a couple of pages on a frequent basis, and less frequently revisited pages are often revisited together with other pages [15]. In a previous study [12], we compared various combinations of methods for predicting Web revisitation. In the context of this paper, we only consider basic methods and do not attempt to find optimal combinations of these methods - as has been done, among others, by Etter et al [8]; our purpose is to verify the performance of each method and to what extent these prediction methods are correlated.

5.1 Prediction Methods

For predicting the next location a user will visit we applied five basic, commonly used methods. These methods are:

- *Top-N locations*: Take the top-N most popular locations and use this for predicting the next location (baseline).
- *Last-N locations*: This method uses the last N visited locations as a prediction for the next location - this approach is commonly applied for revisitation support in Web browsers.
- *Hour top-N locations*: Top-N endpoints that are most popular at a particular time of day (on a hourly basis).
- *Top-N closest locations*: The N locations that are closest to the user's current location. This approach is often used in location-based services.
- *Simple Markov Model*: This model calculates, based on previous travels, the probability that a user will travel to some location starting from the current location.

5.2 Evaluation Measures

As we are interested in predicting locations that people will revisit as part of their routine patterns, we apply the above-mentioned methods to each user individually. We only considered the 57 participants with more than 100 trajectories in their travel logs. We 'replayed' the users' travel activities and used each above-mentioned method for predicting the next location in the log.

As evaluation measures, we use the success rates $S@1$ and $S@5$, which indicate whether a next location is part of the set of predicted locations. For applications such as pro-active scheduling it is important that the next location achieves the first rank, but for many other applications it is sufficient if the next location is included in a small set of recommendations. The reported values are averages between users.

In order to verify to what extent the location predictions cover the whole set of frequently and less frequently visited locations, we also report the Shannon entropy for the location predictions and the actually visited locations.

5.3 Results

The prediction methods were applied to all end-locations for each individual user, covering both weekdays and the weekend. As weekend mobility follows different patterns than weekdays, we repeated the experiment with separate models for weekdays and weekends. The differences with the all-week models are discussed at the end of this section.

Success Rates Table 1 shows the success rates for the prediction methods. As expected, the $S@1$ rates are relatively low, except for the Markov model, which performs in line with the results reported by Etter et al [8]. We focus on the

Table 1. Prediction results for all methods

	Top-N	Last-N	Hour	Distance	Markov
$S@1$	0.286	0.204	0.467	0.275	**0.626**
$S@5$	0.612	0.546	0.829	0.49	**0.931**

performance in terms of $S@5$, which - as discussed earlier - is often sufficient for recommendation purposes. The baseline method, Top-N, which always predicts the most frequently visited locations, has a moderate performance with $S@5$ of 61%. This confirms the importance to include the long tail of less frequently visited locations. The Last-N method, which predicts that the next location will be a recently visited location, has an even lower performance - which shows that recency plays only a moderate role in location revisitation.

The worst performing method is the distance-based approach, which predicts that users will revisit a location that is close to the current location. In less than 50% of the cases, this prediction is correct. This may come as a surprise, as most location-based services consider distance as an important factor for recommendations [14].

The hour-based method performs significantly better than the previous methods ($S@5$ about 83%). This indicates that location revisitation highly depends on the time of day, an effect that we have observed in the previous section. The simple Markov model achieves the best performance. If just one single location is predicted, the prediction is correct in 62% of the cases; a list of five locations contains the actual end-location in 93% of the cases.

Table 2. Entropy of locations for predictions and visited locations

	Top-N	Last-N	Hour	Distance	Markov	Actual
Top-1	0	4.142	1.803	4.165	2.852	4.139
Top-5	2.322	4.635	4.201	4.896	3.157	-

Entropy and Revisit Rate Apart from the success rate of predicted locations, it is also important to take the variety and coverage of the predictions into account. For this, we employed the Shannon entropy measure [7]. Low entropy measures for a prediction method indicate that they often suggest the same (most popular or most visited) locations. As can be seen in Table 2, this is - not surprisingly - the case for the Top-N method. The hour-based method reaches a higher entropy and probably for this reason a higher success rate than Top-N. The Last-N and Distance-based methods reach the highest entropy values, but rather low success rates. Apparently, the variety in visited locations is not successfully captured by these methods. The Markov model has reasonable

[7] `http://en.wikipedia.org/wiki/Entropy_(information_theory)`

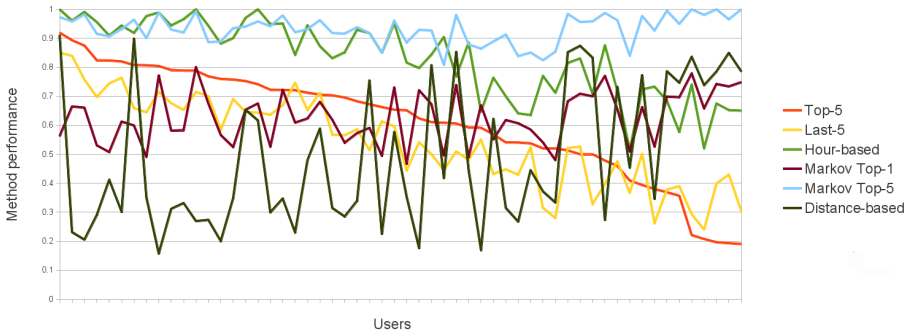

Fig. 4. Performance of the prediction methods for individual users, ordered by the top-5 performance.

entropy values; the success rates indicate that the entropy within the Markov model represents actual user behavior.

There are no significant correlations between the method performance measures and the number of trips, or end locations, of a user. This indicates that the method performance does not depend on the data size.

Another measure on user mobility that may impact the method performance, is the extent to which a user revisits locations. Similar to [17] we define the *revisit rate* as the ratio between the number of end locations and the number of trips of a user. The average revisit rate for the participants in the analysis is 74% ($\sigma = .11$), with a minimum of 46% and a maximum of 91%. Indeed, there are significant correlations with $R > .5$ between the revisit rate and the Top-N, Last-N and hour-based methods. By contrast, the revisit rate corresponds negatively with the performance of the Markov and distance-based methods.

Correlations Figure 4 shows a significant interaction between the Top-N, Last-N and hour-based prediction methods ($R > .84, P < .01$), which confirms that all three methods capture the most popular locations (as shown by [7], the last-n locations often contain top-n locations). Of the three, the hour-based method performs best in capturing the user's behavior. The Markov-based and distance-based methods are positively correlated ($R = .46, p < .01$), which indicates that there is a tendency to select next locations based on the distance, but that other factors (such as locations that are often visited together) play a role as well.

Given the differences in mobility patterns on weekdays and during the weekend - usually due to the absence of commuting on Saturday and Sunday - we repeated our experiments with separate models for weekdays and weekend. All prediction techniques performed more or less similar to the all-week versions. The $S@5$ values for Top-N and Last-N were about 7% higher during weekdays and about 7% lower during the weekend, which confirms our observation that weekend patterns are less stable than weekday patterns.

6 Discussion and Conclusions

In this paper, we analyzed human mobility patterns based on GPS data of 199 people. In line with [9], we observed that human mobility patterns contain strong regularities: people typically spend most of their time at and between a small number of locations. In addition, we found that these popular locations and most-followed trajectories (e.g. the daily commute) also serve as starting points for visits to several other locations that form the long tail of a person's whereabouts.

We also found that most people have a relatively regular schedule for traveling from one location to another (e.g. commuting on weekdays, fixed weekend activities). The purpose of the end-locations, as derived from keywords associated with the locations, also depends on the time of day.

The mobility patterns that we observed can be modeled with different basic methods for revisitation prediction. The comparison of several basic methods showed that a simple Markov model has the best performance, followed by the hour of the day. Note that the Markov model only needs location identifiers without geographic coordinates, which makes it a suitable technique for privacy-preserving location-based personalization. By contrast, distance between locations, even though widely used a main criterion for current location-based services, seems to be a less important factor. The entropy and correlation measures of these methods provide indications on how they can be combined in more complex models.

Most location-based services focus on the recommendation of new locations, usually based on the user's current location. Individual daily and weekly patterns provide a basis for supporting everyday activities involving already visited locations. Particularly the observation that most locations can be connected to one 'base location' or one trajectory - can be exploited in various ways, varying from recommendations to navigate to regular stops on the way back home to targeted advertisements at the moment that a user embarks on a Saturday-morning shopping trip.

References

1. Amini, S., Brush, A., Krumm, J., Teevan, J., Karlson, A.: Trajectory-aware mobile search. In: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems. pp. 2561–2564. ACM (2012)
2. Ashbrook, D., Starner, T.: Using gps to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing 7, 275–286 (2003)
3. Bellotti, V., Begole, B., Chi, E., Ducheneaut, N., Fang, J., Isaacs, E., King, T., Newman, M., Partridge, K., Price, B., et al.: Activity-based serendipitous recommendations with the magitti mobile leisure guide. In: Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems. pp. 1157–1166. ACM (2008)
4. Biagioni, J., Krumm, J.: Days of our lives: Assessing day similarity from location traces. In: User Modeling, Adaptation, and Personalization. pp. 89–101. Springer (2013)

5. Brush, A.B., Krumm, J., Scott, J.: Exploring end user preferences for location obfuscation, location-based services, and the value of location. In: Proceedings of the 12th ACM international conference on Ubiquitous computing. pp. 95–104. Ubicomp '10, ACM, New York, NY, USA (2010)
6. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 759–768. CIKM '10, ACM, New York, NY, USA (2010)
7. Cockburn, A., McKenzie, B.J.: What do web users do? an empirical analysis of web use. Int. J. Hum.-Comput. Stud. 54(6), 903–922 (2001)
8. Etter, V., Kafsi, M., Kazemi, E., Grossglauser, M., Thiran, P.: Where to go from here? mobility prediction from instantaneous information. Pervasive and Mobile Computing 9(6), 784–797 (2013)
9. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature 453(7196), 779–782 (June 2008)
10. Herder, E., Siehndel, P.: Daily and weekly patterns in human mobility. In: AUM 2012, Workshop on Augmented User Modeling. Extended Proceedings of UMAP 2012 (2012)
11. Joseph, K., Tan, C.H., Carley, K.M.: Beyond "local", "categories" and "friends": clustering foursquare users with latent "topics". In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. pp. 919–926. UbiComp '12, ACM, New York, NY, USA (2012)
12. Kawase, R., Papadakis, G., Herder, E., Nejdl, W.: Beyond the usual suspects: context-aware revisitation support. In: Hypertext. pp. 27–36 (2011)
13. Krumm, J., Brush, A.J.B.: Learning time-based presence probabilities. In: Proceedings of the 9th international conference on Pervasive computing. pp. 79–96. Pervasive'11, Springer-Verlag, Berlin, Heidelberg (2011)
14. Mokbel, Mohamed, F., Bao, J., Eldawy, A., Levandoski, J.J., Sarwat, M.: Personalization, socialization, and recommendations in location-based services 2.0. In: Proceedings of the PersDB 2001 Workshop (2011)
15. Obendorf, H., Weinreich, H., Herder, E., Mayer, M.: Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In: CHI. pp. 597–606 (2007)
16. Song, C., Qu, Z., Blumm, N., Barabsi, A.L.: Limits of predictability in human mobility. Science 327(5968), 1018–1021 (2010)
17. Tauscher, L., Greenberg, S.: How people revisit web pages: empirical findings and implications for the design of history systems. Int. J. Hum.-Comput. Stud. 47(1), 97–137 (1997)
18. Teevan, J., Karlson, A., Amini, S., Brush, A.J.B., Krumm, J.: Understanding the importance of location, time, and people in mobile local search behavior. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. pp. 77–80. MobileHCI '11, ACM, New York, NY, USA (2011)
19. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining interesting locations and travel sequences from gps trajectories. In: Proceedings of the 18th international conference on World wide web. pp. 791–800. WWW '09, ACM, New York, NY, USA (2009)