Extracting Semantics of Individual Places from Movement Data by Analyzing Temporal Patterns of Visits

Gennady Andrienko[1,4], Natalia Andrienko[1,4], Georg Fuchs[1], Ana-Maria Olteanu Raimond[2], Juergen Symanzik[3], Cezary Ziemlicki[2]

[1] Fraunhofer Institute IAIS (Intelligent Analysis and Information Systems) and University of Bonn, Germany
[2] Orange Labs R&D, Paris, France
[3] Utah State University, Logan, USA
[4] City University, London, UK

ABSTRACT

Data reflecting movements of people, such as GPS or GSM tracks, can be a source of information about mobility behaviors and activities of people. Such information is required for various kinds of spatial planning in the public and business sectors. Movement data by themselves are semantically poor. Meaningful information can be derived by means of interactive visual analysis performed by a human expert; however, this is only possible for data about a small number of people. We suggest an approach that allows scaling to large datasets reflecting movements of numerous people. It includes extracting stops, clustering them for identifying personal places of interest (POIs), and creating temporal signatures of the POIs characterizing the temporal distribution of the stops with respect to the daily and weekly time cycles and the time line. The analyst can give meanings to selected POIs based on their temporal signatures (i.e., classify them as home, work, etc.), and then POIs with similar signatures can be classified automatically. We demonstrate the possibilities for interactive visual semantic analysis by example of GSM, GPS, and Twitter data. GPS data allow inferring richer semantic information, but temporal signatures alone may be insufficient for interpreting short stops. Twitter data are similar to GSM data but additionally contain message texts, which can help in place interpretation. We plan to develop an intelligent system that learns how to classify personal places and trips while a human analyst visually analyzes and semantically annotates selected subsets of movement data.

1 INTRODUCTION

Analysis of movement data is now a hot research topic in GI Science [10], spatial databases [8], data mining [7] and visual analytics [4]. There is a special interest to data about movements of people for their potential to enable insights into patterns of people's mobility and activities. Thus, for spatial planning applications, it is important to know the typical movement-related behaviors and needs of the people on a given territory in order to find suitable planning options and predict their possible effects, e.g., by means of simulation. A usual approach is asking a sample of the population to provide their diaries, i.e., reports about the daily activities and travels. The diaries are then generalized to a set of typical mobility and activity profiles, which can be used to generate synthetic populations for simulation of various "what if" scenarios. However, such population surveys are expensive and the resulting data are very limited in their temporal extent (usually each person reports about a single day or a few days).

Unlike personal diaries, movement data can be collected automatically by GPS trackers or by recording positions of mobile devices and, hence, are cheap and easy to acquire. However, the data consist of just geographic coordinates and time stamps and lack any semantic information. To make a reasonable alternative to diaries, movement data need to be semantically enriched by attaching meanings to the spatial positions and travels of the people. Semantic information can be derived by comparing the positions with locations of predefined places of interest (POI) [11]. For example, a daily trajectory that starts and ends at a hotel, visits one or more tourist attractions, and stops at a restaurant is classified as a "tourist's trajectory". This approach, however, does not uncover personal POIs such as home, work, child's school or kindergarten, and regularly visited grocery. In the area of human geography, mobile phone calls data are analyzed to find personal POIs and classify some of them as home or work places based on the frequency of the person's calls from each place, their average time of the day, and the standard deviation of the time of the day [1]. There is no attempt to classify other kinds of POIs.

Interactive visualizations allow a human analyst to discover and interpret personal POIs and mobility patterns from movement data of one person [2]; however, it is not feasible to do the same for many individuals as analyst's time is a very limited and costly resource. Our goal is to extend this visual analytics approach [2] to much larger datasets reflecting movements of many people. We apply a procedure that extracts stops, finds spatial clusters of the stops, which correspond to individual POIs, and creates "temporal signatures" characterizing the temporal distribution of person's presence in each POI. The signatures are then visualized to enable understanding of place meaning by an analyst. After the analyst assigns a subset of places to some category based on their temporal signatures, places with similar signatures can be automatically found in a database and assigned to the same category.

We illustrate our ideas using three example datasets. The first dataset contains 609,241 time-stamped GSM positions (cell tower coordinates) of 67 persons in France for a period of 49 days. The data have been obtained by an active collection procedure: selected people were pinged every 7 minutes. The second dataset contains 81,389 GPS positions of a single person in the USA for a period of 351 days. We have also tested the approach on the Nokia mobile data challenge dataset consisting of 1,153,070 positions of 38 persons during 262 days. The data were provided by volunteers for use in research. The third dataset consists of 163,203 georeferenced tweets from the greater Seattle urban area in USA, which were posted by 2,607 local Twitter users during the two-month period from August 8th to October 8th, 2011.

http://geoanalytics.net/and; andrienko@geoanalytics.net

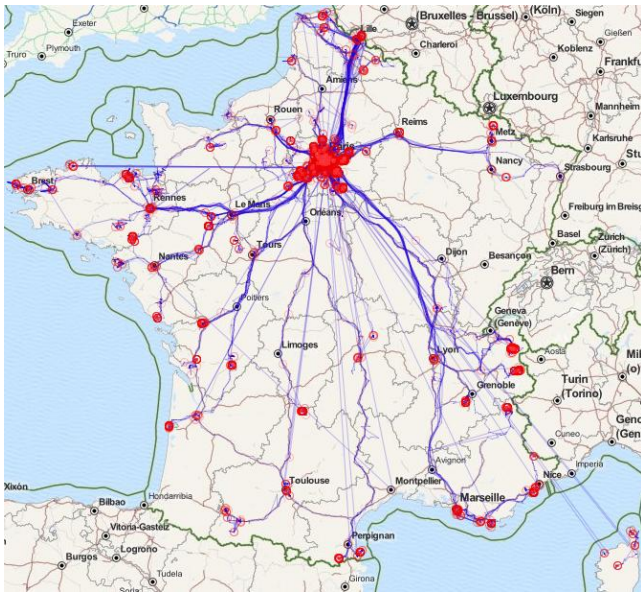

Figure 1. Overview of the GSM data set: trajectories and POIs

2 COMPUTATIONAL PROCESSING

To extract individual POIs, we adapt the procedure proposed in [5]. It starts with extracting each person's positions of stops with the user-specified minimum duration Δt_{min}. For this purpose, the diagonal of the bounding rectangle enclosing each recorded position and the following positions, if any, within the time window Δt_{min} is computed. If the diagonal length is below a user-specified distance threshold d_{max}, the position is treated as a stop position. The choice of Δt_{min} and d_{max} depends on the temporal and spatial resolution of the data. We take Δt_{min} = 30 minutes and d_{max} = 1km for the GSM data and Δt_{min} = 5 minutes and d_{max} = 100m for the GPS data. These values have been selected with the help of an interactive interface [6].

At the second step, clusters of each person's repeated stops are identified by means of density-based clustering [5], which is done separately for each person. For our GSM and GPS data examples, we use spatial thresholds of 1km and 100m and minimal numbers of neighbors 5 and 3, respectively. This results in 283 clusters for 67 persons in the GSM data set (Fig.1) and 19 clusters for a single person in the GPS data set. These clusters represent candidate individual POIs. To outline the POIs, spatial buffer zones around the clusters are built. In our examples, we build buffer zones with radii of 500m and 50m, respectively (one half of the clustering distance threshold).

At the third step, the following temporal aggregates are computed for each candidate POI:

- Total count of visits;
- Minimum/average/maximum/median durations of visits;
- Count of different days, weeks, months with visits;
- Count of different days of week;
- Time series of visit counts by hours of a day:
  - for all visits to the POI;
  - for the visits on working days;
  - for the visits on the weekend.
- Time series of visit counts by days of a week;
- Time series of visit counts by weeks (GSM data) and months (GPS data).

These aggregates represent "temporal signatures" of the POIs. We hypothesize that (1) the likely meanings of the POIs can be inferred from the temporal signatures and (2) similar temporal

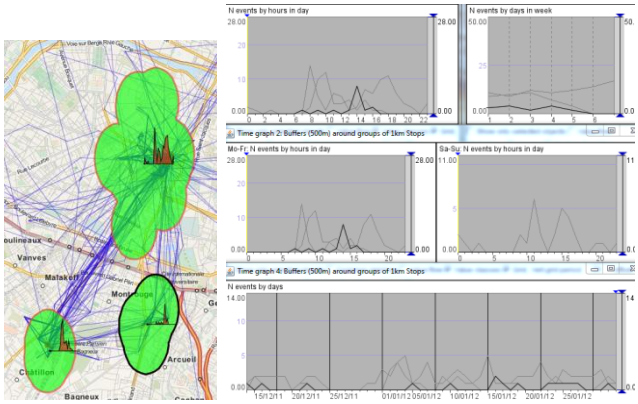

Figure 2. POIs and their temporal profiles for person A. Two gray curves on each graph correspond to the home and work places, the black curves – to the leisure place.

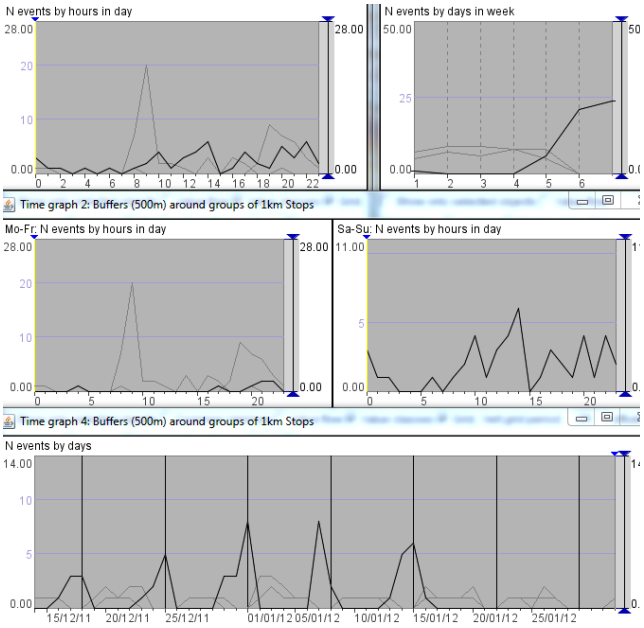

Figure 3. Temporal profiles of POIs of person B. Two gray curves on each graph correspond to the home and work places in Paris and the black curves to the weekend home in Lille.

signatures of POIs of different persons correspond to similar meanings of the POIs for these persons. We suggest a set of interactive visual tools that enable an analyst to explore the temporal signatures of different POIs and determine their probable meanings based on background knowledge of human behaviors. As the analyst gives a certain meaning to one POI, the same meaning can be automatically assigned to all POIs having similar temporal signatures according to a specially designed distance function. In this way, the method is scaled to large datasets.

3 VISUAL INTERPRETATION OF POIS IN GSM DATA

The interactive visual interface for POI exploration, interpretation, and semantic annotation includes two main components: a map and a set of time graphs. The map shows the spatial positions and extents of the POIs. One of the temporal aggregates can be represented by embedded diagrams at the POI locations. Additionally, the stop points and the trajectories can be drawn semi-transparently. The time graphs represent the computed time series.

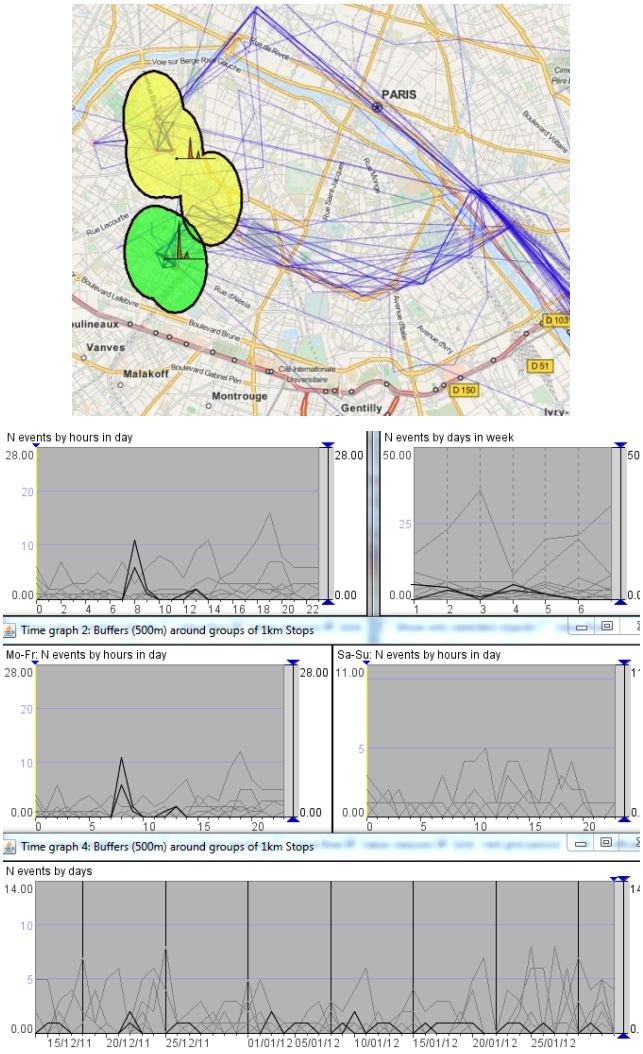

Figure 4. POIs ad their temporal profiles for person C. Highlighted are two probable places of work attended in different days.

Synchronous querying of all three information layers (trajectories, stops, and POIs) is supported by interactive filtering tools that link related datasets. Selecting a subset of trajectories also selects the stops extracted from these trajectories. Selecting a subset of stops also selects the POIs including these stops. The filters can also be turned to work in the opposite direction: from stops to trajectories and from POIs to stops. The filters affect all displays, in particular, the map and time graphs.

Figure 2 shows an example configuration on the screen after selecting a trajectory of one person (person A). Three areas on the map outline the major POIs and the diagrams show the temporal profiles of the stops in these areas by hours of a day. Five time graphs show the temporal signatures of the POIs with respect to hours of a day, days of a week, hours of a day for the working days and weekends, and different days. The vertical black lines on the time graph at the bottom correspond to Sundays.

By observing the temporal signatures of each POI, we can infer their likely meanings. The extended area on the north is visited almost every day, more often in the evening. It can be interpreted as the home place of A. The place on the southwest is, evidently, the working place: A stops there only on the working days, mostly in the morning. The POI highlighted in black is visited on some

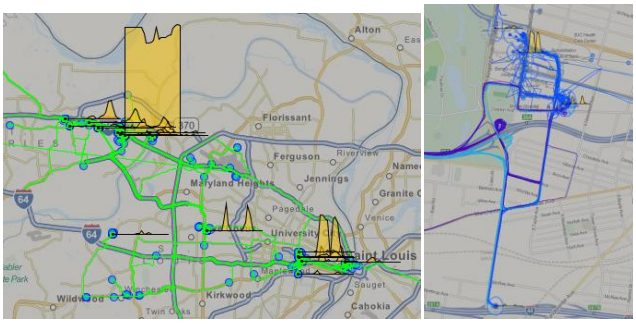

Figure 5. Trajectories from GPS data and extracted POIs.

working days (about 50%) in early afternoon. This may be either a lunch place or a leisure/sport activity location.

Figure 3 shows the temporal profiles of the POIs of person B, who has an easy-to-identify work POI and an area where the stops occur mostly in the evenings of the working days and rarely on weekends. Besides, there is a POI in Lille visited almost every weekend. Evidently, B has two home places, one for the workdays (close to the work place) and one for the weekends.

Figure 4 demonstrates a case when a person has, evidently, two work places. The one filled in light green is attended more often and the other (in yellow) in less than 20% of the working days. Both POIs are never visited on Wednesday.

Additionally to the POIs classified as home and work places, we could also find POIs of repeated evening activities (most likely, leisure) and destinations of occasional weekend trips, distant vacation trips (occurring within a holiday period and having a uniform hourly distribution of stops), and business trips (the hourly distributions of the stops is similar to places of work and home). It is quite easy to infer the transportation modes of the long-distance trips by matching the routes to the road networks and considering the average speeds.

The suitability of GSM data for semantic interpretation of mobility is somewhat limited by the nature of the data collection procedure. Small changes of position are not captured due to the low spatial resolution. Short stops cannot be captured, too.

4 Visual Interpretation of POIs in GPS Data

GPS tracks with fine spatial and temporal resolution enable more sophisticated semantic analysis. In particular, it is possible to find and interpret short stops. We have tested our approach on two GPS datasets describing long-term movements of 38 persons in Switzerland (Nokia mobile data challenge dataset) and one person in the USA. Unfortunately, the data access agreement prohibits publication of the results of the first experiment. We demonstrate the possibilities for analyzing GPS data on the second example.

The map in Figure 5 displays the trajectories of the person (in green) and the extracted POIs with the respective daily profiles of stop occurrences represented by diagrams. The highest diagram corresponds to the home place, and the diagram on the southeast with two high peaks and a drop between them to the work place, where the person stopped in the mornings and afternoons of work days (the days of POI attendance can be seen from time graphs as in Figures 2-4). There is a POI in the middle attended in the mornings and evenings of work days. The stop durations, which can be easily determined from GPS data, are around 4 minutes (maximum 9) in the mornings and 14 minutes on average (maximum 35) in the evenings. This POI can be interpreted as a place of workday activity (work or study) of a family member of the person under analysis. The person brings the family member to that place in the mornings and takes him/her in the evenings. The longer stop durations in the evenings indicate waiting time.

The area around the work place is enlarged in Figure 5 right. The work place can be distinguished from the POI southeast of it, where the stops occurred in the mornings and evenings. Judging from the duration of these stops (5 minutes on the average) and the relative times of their occurrence (before and after the stops in the work place), we can classify this POI as a parking place. The POI on the north was sometimes visited during the work time for up to two hours and the POI on the south was occasionally visited either before or after the work for 5-6 minutes. The former place is, very probably, related to person's work. The meaning of the latter place can be understood only after zooming in the map. We see that this is a place for garbage recycling, which could also be classified using a database of public POIs.

After extracting the POIs, the entire one-year trajectory of the person can be divided into smaller trajectories corresponding to the trips between the POIs. We have clustered these trajectories according to the similarity of the routes [2][12]. The colors of the trajectories in the map in Figure 5 right correspond to different clusters. These colors are also used in the space-time cubes [9] (STC) in Figure 6. To investigate the temporal distribution of person's trips and stops with respect to the weekly and daily cycles, we apply time transformations [3]. In Figure 6 top, the trajectories and stops are temporally aligned to the weekly cycle (Monday is at the bottom of the cube and Sunday at the top). We see repeated patterns of the trips in the five work days and quite different patterns in the weekend. In Figure 6 bottom, the trajectories and stops of the work days are temporally aligned to the daily cycle (morning is at the bottom and evening at the top). In the morning, the person usually drove from the home to the family member's place (these trajectories are colored in light green) and then to the work place (trajectories in cyan). In the evenings, the person drove from the work first to the family member's place (violet) and then to the home place (dark red). In yellow there are trips from the home to the shopping places located close to the home place (the respective daily diagrams are close to or overlap with the diagram of the home place in Figure 5 left). The shopping places were attended in the evenings of the work days. In Figure 6 top, we see that shopping trips also occurred on Saturday. On Thursdays, the person sometimes went from the work place to a POI east of it. From other displays we can see that this POI was visited at lunch time once per month on Thursday; the stop duration was about two hours. This indicates a place of some regular though infrequent activity.

Hence, by analyzing the GPS data, we could derive the same information as is typically provided in a personal diary. Moreover, since the data cover a long time period, we could extract and identify truly routine behaviors, infrequent regular behaviors, infrequent irregularly repeated behaviors, and occasional behaviors. This is much more that could be learned from a diary covering one or a few days.

5 Semantic Interpretation of POIs of Many Persons in Twitter Data

We use a subset of georeferenced Twitter data referring to the greater Seattle urban area in the USA and a time period of two months. We have constructed trajectories of the Twitter users from the geographical positions of the messages posted on the selected territory. We have discarded the trajectories of the individuals who were present on this territory for less than 10 days. The remaining 2,558 trajectories are regarded as belonging to residents of the selected area, who are likely to have repeatedly visited personal places.

We have extracted 4,245 personal places of these people by means of density-based clustering with the spatial distance threshold of 100 m and minimum 5 neighbours for a core point.

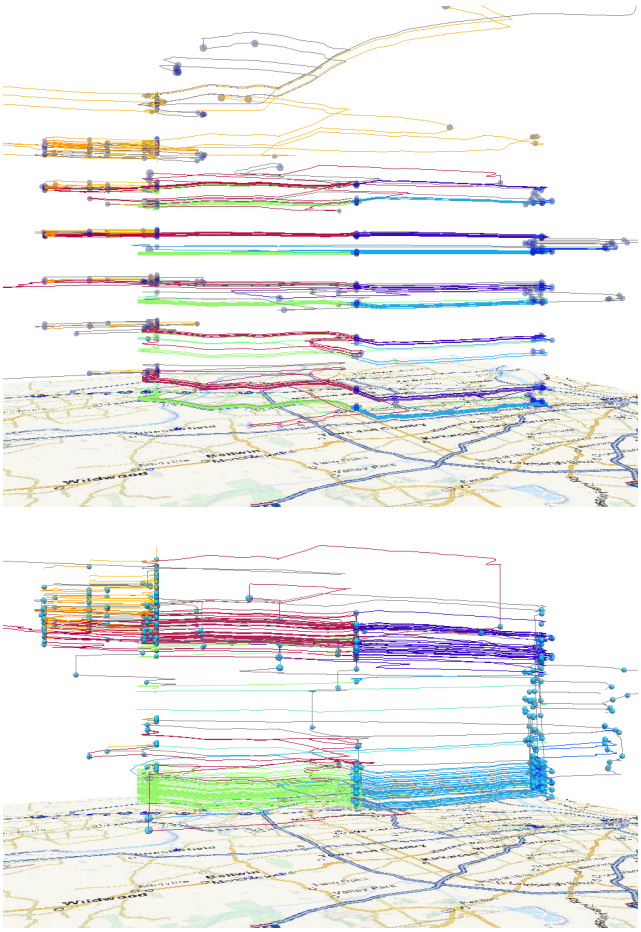

Figure 6. Person's trips (line) and stops (balls) in space-time cubes. Top: the distribution over a week. Bottom: the distribution of the work day trips and stops over a day.

Based on the time series of place visits by days, we have computed for each place the number of different days it was visited. There are 1,488 personal places (35.1% of all) that were visited in 10 or more different days. In our experiment on place interpretation, we focus on these places.

The Twitter data are episodic movement data, where the time gaps between the position records may be quite large. For some time intervals by which the data are aggregated there may be no records; hence, the places where a person was present in these intervals are unknown. Therefore, the values in the time series need to be considered as the lower bounds of the actual presence counts. A value decrease in a place time series does not necessarily mean that the person moved somewhere else. The person could stay in the place but not post new messages from this place in the next one or more time intervals. This needs to be taken into account in interpreting the time series.

Figure 7 gives an example of temporal signatures of personal places of one person. Four time graphs show the time series of place visits by hours on the work days (A), Saturdays (B), and Sundays (C) and the time series of visits by days (D). The line coloured in blue demonstrates a typical time series for a work place: the person was present there only on work days from hour 5 till hour 14. There are two lines coloured in red. The one reaching higher values may correspond to the home place of the person, since the person is present there in the afternoons and evenings of the work days and at any times in the weekend. The second red-

coloured line has a similar temporal distribution, but the values are very low; hence, this is not likely to be a home place. The two lines coloured in orange may correspond to shopping places: they are visited in the afternoons and evenings of the work days and in different times of the day on the weekend.

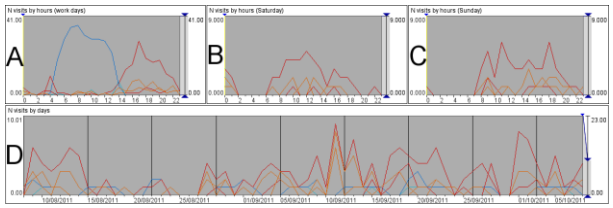

Figure 7. The time graphs show temporal signatures of personal places of one Twitter user. A, B, C: time series of place visits by hours of the day for the work days (A), Saturdays (B), and Sundays (C); D: time series of place visits by days.

To be able to analyze a large set of places without considering the temporal signatures of each place one by one, two approaches are possible: similarity analysis of the time series and clustering of the time series. For both cases, a suitable distance function for assessing the similarity of two time series is needed. In this case, the Euclidean distance may not work well enough. It is reasonable to apply a more sophisticated distance function that can perform moderate transformations of the time series: shifting, stretching, shrinking, and scaling. This is needed, in particular, to account for different working times and, consequently, for different times of coming home and other activities taking place on work days. For our illustrations, we estimate the similarity between time series by means of the Euclidean distance; however, before computing the distances, we apply temporal smoothing and then transform the absolute values to normalized deviations from the means.

For similarity analysis, the analyst selects a place with previously assigned semantic interpretation and uses the distance function to compute the distances between the time series of this place and those of all other places. Then the analyst applies interactive dynamic filtering by the distances and looks at the time graphs to select a subset of sufficiently similar time series. The places these time series belong to can be given the same interpretation as the exemplar place. An illustration is given in Figure 8. The time graphs in the upper row show the selected time series similar to the work place time series from Figure 7. The time series belong to 154 different places, not counting the exemplar place. In the lower row, the selected time series are similar to the home place time series from Figure 7. These time series belong to only 5 different places.

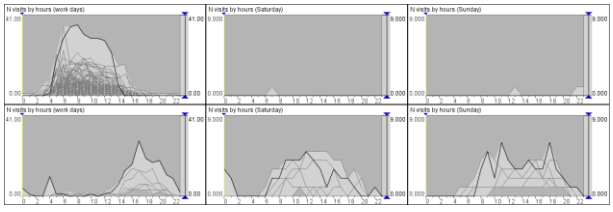

Figure 8. By means of similarity analysis, places with temporal signatures similar to selected ones have been found. Upper row: likely work places; lower row: likely home places.

Figure 9 demonstrates selected results of clustering. The cluster presented in the upper row (197 members) consists mostly of time series characteristic for work places. The cluster in the middle row can be interpreted as a cluster of likely home places (74 members). The cluster in the lower row, probably, includes places of stops of public transport (134 places), where people mostly appear in early morning hours of the work days. It should be noted, however, that the clusters are not very "clean". Thus, the "work" cluster includes also a few time series with relatively high presence in the evenings. The time series in the other clusters we have obtained are more difficult to interpret, i.e., the meanings of the place for the individuals cannot be guessed from the shapes of the time series.

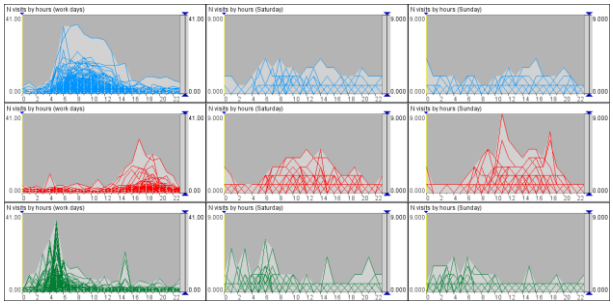

Figure 9. Temporal signatures of personal places have been clustered by similarity. The images show three selected clusters.

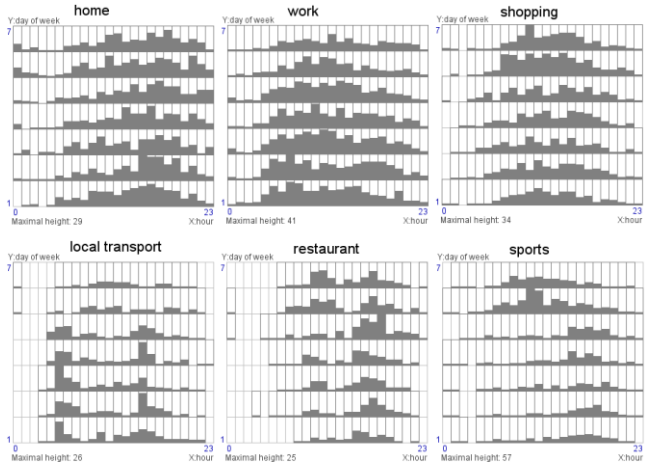

Figure 10. Temporal distributions of occurrences of selected topics in tweets. Rows: days of the week; columns: hours of the day.

Our experiments showed that deriving semantic interpretation from only temporal and statistical information may be possible for a small proportion of personal places. For the remaining places, it may be necessary to look at their relative geographical positions with respect to the other personal places and the relative times of visiting with respect to the times of visiting the other personal places. For example, a place visited on the way from home to work may be a place of child's school or kindergarten. Such more sophisticated analyses are not yet supported by visual and computational techniques.

We have also investigated whether the texts of the tweets can help in interpreting the meanings of the places from which they are posted. We have used a manually created ad-hoc mini-vocabulary for recognizing the topics of tweets based on the occurrences of particular words. For instance, the topic "local transport" is recognized from occurrences of any of the words "bus", "metro", "tram", and so on. We then explored the temporal distributions of the topic occurrences over days of the week and

times of the day (Figure 10) to check whether the temporal patterns correspond to expected temporal patterns of people's activities. The temporal distributions of some topics, including "home", "work", "education", "shopping", and others, do not exhibit prominent patterns and do not correspond sufficiently well to the expected times of the corresponding activities. However, "local transport", "restaurant", "sports", and some other topics have quite clear patterns that match the expectations regarding the corresponding activity times. Hence, for some types of places, occurrences of particular topics in tweets potentially provide evidence for inferring the place meaning. However, we could not yet find a good method for combining evidences coming from the analysis of temporal signatures and from the contents of the texts.

6 Discussion and conclusion

Our examples have demonstrated that semantic information about personal mobility behaviors can be extracted from movement data by means of interactive visual analysis. As noted in section 2, a human analyst can classify selected POIs according to their temporal signatures, and other POIs with similar signatures can be then classified automatically. When movement data have sufficiently high temporal resolution, as in GPS tracks or GSM data obtained through active collection, this approach can work well for home, work, and shopping places. However, even for high-resolution data, the temporal signatures in terms of daily and weekly stop distributions may be insufficient for interpreting repeated short stops. In our experiments, we also paid attention to the relative temporal and spatial positions of the stops with respect to the stops in already known places (e.g. home and work) and the variation of the stop durations. Our current research topic is how to support translating such human inferences into classification rules that could be then applied automatically. We envisage an intelligent system that learns how to classify personal places and trips while a human analyst visually analyzes and semantically annotates selected portions of movement data.

Episodic movement data, such as standard GSM data (call data records) and georeferenced tweets, pose much greater challenges than temporally regular data. Personal places can be found and interpreted in such data only when persons frequently make calls or send messages from these places. Still, the number of active users represented in a dataset may be high enough for extracting large numbers of mobility and activity patterns potentially valuable for different applications.

References

[1] R.Ahas, S.Silm, O.Järv, E.Saluveer, M.Tiru. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. Journal of Urban Technology, 17(1), pp.3-27, April 2010

[2] G.Andrienko, N.Andrienko, S.Wrobel. Visual Analytics Tools for Analysis of Movement Data. ACM SIGKDD Explorations, 9(2), pp. 38-46, Dec. 2007

[3] G.Andrienko, N.Andrienko. Poster: Dynamic Time Transformation for Interpreting Clusters of Trajectories with Space-Time Cube. IEEE VAST 2010, pp. 213-214, Oct. 2010

[4] G.Andrienko, N.Andrienko, P.Bak, D.Keim, S.Kisilevich, S.Wrobel. A conceptual framework and taxonomy of techniques for analyzing movement. Journal of Visual Languages and Computing, 22(3), pp.213-223, 2011

[5] G.Andrienko, N.Andrienko, C.Hurter, S.Rinzivillo, S.Wrobel. From Movement Tracks through Events to Places: Extracting and Characterizing Significant Places from Mobility Data. IEEE Visual Analytics Science and Technology (VAST 2011), Proceedings, IEEE Computer Society Press, pp.161-170, 2011

[6] G.Andrienko, N.Andrienko, M.Heurich. An Event-Based Conceptual Model for Context-Aware Movement Analysis. International Journal on Geographical Information Science, 25 (9), pp.1347-1370, 2011

[7] F. Giannotti, D. Pedreschi. Mobility, data mining, and privacy: geographic knowledge discovery. Springer, 2008.

[8] R. Gueting, M. Schneider. Moving objects databases. Morgan Kaufmann, 2005.

[9] M.-J.Kraak. The space-time cube revisited from a geovisualization perspective. In: Proc. 21st International Cartographic Conference, Durban, South-Africa, pp. 1988-1995, Aug. 2003

[10] P. Laube. Progress in movement pattern analysis. In Behaviour Monitoring and Interpretation – BMI - Smart Environments, pp. 43–71, IOS Press, 2009.

[11] C.Parent, S.Spaccapietra, C.Renso et al. Semantic Trajectories Modeling and Analysis. ACM Computing Surveys, 45(4), 2013

[12] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, G. Andrienko. Visually driven analysis of movement data by progressive clustering. Information Visualization 7 (3-4), pp. 225-239, 2008

[13] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, D. Thom. Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics. *Computers in Science and Engineering*, 15(3):72-82, 2013.