Visual Analytics Methodology for Scalable and Privacy-Respectful Discovery of Place Semantics from Episodic Mobility Data

Natalia Andrienko[1,2(✉)], Gennady Andrienko[1,2], Georg Fuchs[1], and Piotr Jankowski[3,4]

[1] Fraunhofer Institute IAIS, Sankt Augustin, Germany
{natalia.andrienko,gennady.andrienko,
georg.fuchs}@iais.fraunhofer.de
[2] City University London, London, UK
[3] San Diego State University, San Diego, USA
pjankows@mail.sdsu.edu
[4] Institute of Geoecology and Geoinformation,
Adam Mickiewicz University, Poznan, Poland

Abstract. People using mobile devices for making phone calls, accessing the internet, or posting georeferenced contents in social media create episodic digital traces of their presence in various places. Availability of personal traces over a long time period makes it possible to detect repeatedly visited places and identify them as home, work, place of social activities, etc. based on temporal patterns of the person's presence. Such analysis, however, can compromise personal privacy. We propose a visual analytics approach to semantic analysis of mobility data in which traces of a large number of people are processed simultaneously without accessing individual-level data. After extracting personal places and identifying their meanings in this privacy-respectful manner, the original georeferenced data are transformed to trajectories in an abstract semantic space. The semantically abstracted data can be further analyzed without the risk of re-identifying people based on the specific places they attend.

1 Introduction

The topic of this presentation, based on [1], is semantic modeling and semantic analysis of mobility data (trajectories of people). Currently, the main approach to attaching semantics to mobility data is comparing the locations of points from trajectories with locations of known places of interest (POI) [2]. This approach, however, cannot identify places having personal meanings, such as home and work. Identifying personal places is a challenging problem requiring scalable methods that can cope with numerous trajectories of numerous people while respecting their personal privacy [3, 4]. Our contribution consists of such an approach and a method for semantic abstraction of mobility data enabling further analyses without compromising personal privacy.

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 254–258, 2015.
DOI: 10.1007/978-3-319-23461-8_25

A special focus of the paper is episodic mobility data [5, 6], where large temporal and spatial gaps can exist between consecutive records, but the proposed approach also works for data with fine temporal resolution. It adheres to the visual analytics paradigm [7], which combines computational analysis methods, such as machine learning techniques, with interactive visual tools supporting human reasoning.

2 Problem Statement and Methodology Overview

The input **data** are episodic human mobility traces, such as records about the use of mobile phones or other mobile devices at various locations. Each record includes a person's (user's) identifier, time stamp, and location specification, which may be geographic coordinates or a reference to some spatial object with known coordinates, such as mobile network antennas. The **goal** is to obtain "semantic trajectories" [2], in which the geographic locations are substituted by semantic labels denoting the meanings of the visited places or types of activities performed there, e.g., 'home', 'work', 'eating', 'recreation', etc. The transformation needs to be done for a large set of individuals in such a manner that their geographic positions are hidden from the analyst. The resulting semantic trajectories are devoid of geographic positions and thus can be viewed and further analyzed without compromising individuals' location privacy.

For checking the plausibility of places to have this or that meaning, land use (LU) data are utilized. For example, when a set of places is going to be labelled as 'home', it is checked, based on LU data, whether most of them are in residential areas. A possible alternative is data about POI, such as public transport stops, schools, shops, and restaurants, which can be retrieved from geographic databases or obtained from map feature services, such as OpenStreetMap (www.openstreetmap.org). Having POI data, it is possible to derive counts of different POI types inside places or within a specified distance threshold. Before assigning some meaning to a set of places, the compatibility of this meaning with the POI types occurring in these places is checked.

The analytical workflow consists of the following steps:

1. Extract repeatedly visited personal and public places.
2. For each place, compute a time series of visits by hourly intervals within the weekly cycle, i.e., ignoring the specific dates.
3. Attach LU or POI attributes to the places.
4. For each target meaning ('home', 'work', 'eating', 'shopping', etc.):
   - 4.1. Derive relevant attributes (criteria) from the time series of place visits.
   - 4.2. Based on the attribute values, select candidate places for the target meaning.
   - 4.3. Validate the place selection with LU or POI data. Iteratively modify the selection for maximizing the proportion of relevant land uses or POI types.
   - 4.4. Assign the target meaning to the selected places. Exclude from further analysis the places that have already received meanings.
5. Replace the geographic positions in the input data with the semantic labels (meanings) of the places containing the positions.
6. Create a "semantic space", i.e., a spatial arrangement of the set of place meanings, and treat the transformed data as trajectories in this semantic space.
7. Apply movement analysis methods to the semantic space trajectories.

For place extraction, we have developed a special algorithm that groups position records by spatial proximity. To extract personal places, the positions of each individual are clustered separately; to extract public places, the positions of all people are clustered together. Places are defined by constructing boundaries (spatial convex hulls or buffers) around the clusters of positions. The tool works automatically. It takes input data from the database, processes the data, and puts the resulting place boundaries back in the database without showing them to the analyst.

The task of identifying place meanings requires utilization of a human analyst's background knowledge and cognitive capabilities. This task is supported by interactive visual techniques that only show data aggregated over either the whole set or groups of places and do not allow access to individual data. Multi-criteria ranking is used for identifying the most probable home and work places. The places with the best ranks are considered as candidates for receiving the target meaning. The fitness of the candidates is checked using the statistics of the associated LU or POI classes. The criteria weights can be interactively modified to maximize the proportion of relevant LU or POI classes and minimize the proportion of irrelevant classes. For target meanings other than 'home' and 'work', candidate places are selected through interactive filtering based on relevant temporal attributes and land use or POI information.

3 Feasibility Studies

The feasibility of the approach is demonstrated using two case studies: one with simulated tracks from the VAST Challenge 2014 [8], for which ground truth is available, and the other with real traces built from georeferenced tweets posted during one year within the metropolitan area encompassing San Diego (USA) and the surrounding communities. The datasets contain positions of 35 personal cars and 4,286 Twitter users, respectively. We extracted 202 personal and 41 public places from the VAST Challenge data and 38,225 personal and 9,301 public places from the San Diego data.

By applying our methodology to the VAST Challenge data, we were able to identify the meanings of 170 personal places (84%) and 40 public places out of 41 (97.5%). The results match the available ground truth information. For the San Diego test case, we managed to attach semantic labels to 65% of the personal places and 55.3% of the public places. We were able to identify the probable home places of 3,873 persons (90.4%) and the probable work or study places for 2,171 persons (50.7%). For 1,950 persons (45.5%), it was possible to find both home and work places. The largest class of personal places is 'shopping' (4,695 places). Other large classes include 'eating' (2,194), 'social life' (1,497), which includes places with many visits in the evening and night hours and on the weekend, and 'transport' (1,315). No ground truth is available for checking these results; however, further analysis of the semantically abstracted data (semantic trajectories) corroborates the plausibility of place meaning assignments. Fig. 1 shows an example of a possible avenue to further analysis.

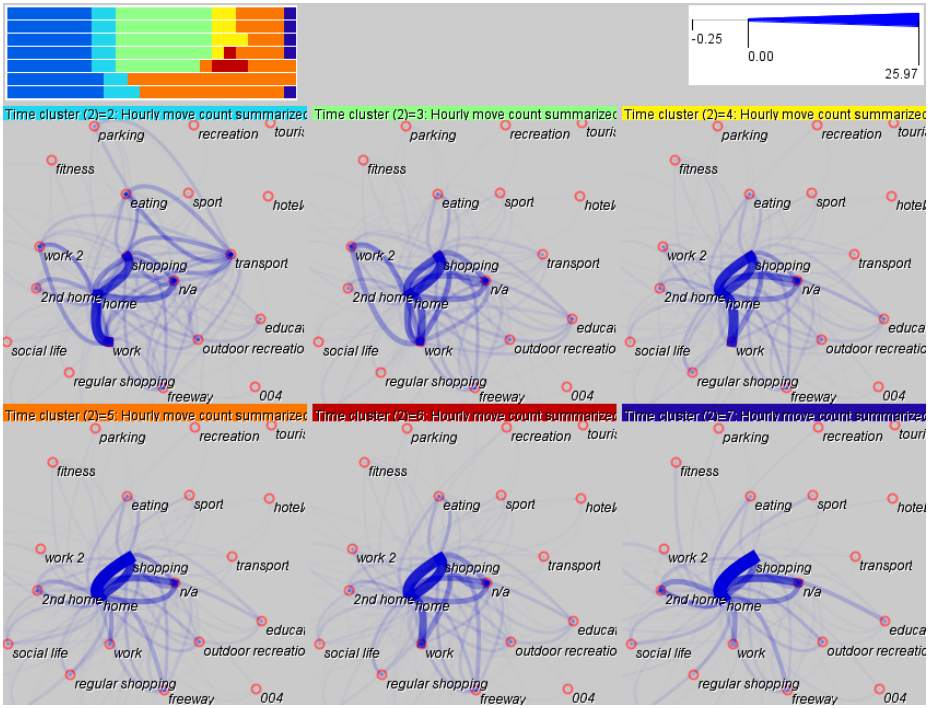

Fig. 1. The background of the maps is a "semantic space", i.e., a 2D spatial arrangement of the set of semantic labels of places. The San Diego data transformed to semantic trajectories have been aggregated into flows between the semantic space locations by hourly time intervals of the weekly cycle. The intervals have been clustered based on the similarity of the sets of the flows; the cluster membership is represented by colors in the calendar view on the top left (each cluster has a specific color). The maps show the flows summarized by the time clusters.

References

1. Andrienko, N., Andrienko, G., Fuchs, G., Jankowski, P.: Scalable and Privacy-respectful Interactive Discovery of Place Semantics from Human Mobility Traces. Information Visualization (2015). doi:10.1177/1473871615581216, Appendix: http://geoanalytics.net/and/papers/placeSemantics/
2. Parent, C., Spaccapietra, S., Renso, C., et al.: Semantic Trajectories Modeling and Analysis. ACM Computing Surveys **45**(4), article 42 (2013)
3. Giannotti, F., Pedreschi, D. (eds.): Mobility, Data Mining and Privacy - Geographic Knowledge Discovery. Springer, Berlin (2008)
4. Cuellar, J., Ochoa, M., Rios, R.: Indistinguishable regions in geographic privacy. In: Ossowski, S., Lecca, P. (eds.) Proc. 27th Annual ACM Symposium Applied Computing (SAC 2012), pp. 1463–1469. ACM, March 26–30, 2012
5. Andrienko, G., Andrienko, N., Stange, H., Liebig, T., Hecker, D.: Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. Künstliche Intelligenz **26**(3), 241–251 (2012)

6. Andrienko G., Andrienko N., Bak P., Keim D., Wrobel, S.: Visual Analytics of Movement. Springer (2013)
7. Keim, D.A., Kohlhammer, J., Ellis, G., Mansman, F. (eds.): Mastering the Information Age - Solving Problems with Visual Analytics, Eurographics (2010)
8. VAST Challenge 2014: Mini-Challenge 2. http://www.vacommunity.org/VAST+Challenge+2014%3A+Mini-Challenge+2