Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 78 (2024) 215–221

Transportation Research Procedia

www.elsevier.com/locate/procedia

25th Euro Working Group on Transportation Meeting (EWGT 2023)

# Determining user specific semantics of locations extracted from trajectory data

Jens Golze*, Monika Sester

*Institute of Cartography and Geoinformatics, Leibniz University Hannover, Appelstraße 9A, Hannover 30167, Germany*

## Abstract

Knowledge about people's daily travel behavior is very relevant for transportation planning, but also for urban and regional planning in general. This information is typically collected through questionnaires or surveys. With the increasing availability of mobile devices capable of using Global Navigation Satellite Systems, it is possible to derive individual mobility behavior on a large scale and for a variety of different users. However, the challenge is to derive the relevant information from the mere GNSS trajectories; in this paper, the relevant information is semantic locations such as *home*, *work place* or *leisure places*. This paper presents an approach to first detect and cluster stop points as potential semantic locations of a user, which are then enriched with Points of Interest from OpenStreetMap and additional features, and finally a Viterbi optimization assigns the most probable semantics to these locations. Overall, this approach produces promising results for predicting user location semantics on a generalized level.

*Keywords:* location of interest; semantic place annotation; GPS data; clustering; Viterbi Optimization

* Corresponding author.
*E-mail address:* jens.golze@ikg.uni-hannover.de

10.1016/j.trpro.2024.02.028

## 1. Introduction

The increasing availability and integration of location-based services, especially in the context of smartphone applications, as well as the availability of mobile, reliable and fast Internet and connections to Global Navigation Satellite Systems (GNSS), offer seemingly unlimited possibilities for the collection of location-based information. The amount and type of this location data varies depending on the application that collects it. For example, it could be collected through social media platforms, which could be geo-tagged images or short text posts. In addition, data could be collected from check-ins at specific locations such as restaurants or others. This was a common practice, especially during the COVID-19 pandemic.

In addition, this location-based data can be used to extract movement patterns for various purposes, such as traffic management or route preference analysis. Global Positioning System (GPS) is one of the most popular known GNSS. GPS trajectory data provides a detailed view of the actual movement of individuals. Thus, GPS trajectory data, especially large trajectory collections, could provide information not only about where an individual has been moving, but also where the individual's places of interest or important places are located, such as their home place, work, shopping areas, but also general public places that they have visited.

In this paper, machine learning techniques and semantic point annotations are used to extract and enrich user locations of special interest with a set of descriptive features from three different categories. Furthermore, these features are used by a Viterbi optimization to assign a semantic label to the user's region of interest. The current state of the work in progress is presented below.

## 2. Related Work

In order to reveal the semantics of places visited by users during their daily mobility, GPS trajectory data is used. First, relevant places have to be identified among all trajectory points. Usually, so-called *stop points* or *interesting places* (Feuerhake et al. 2011) are candidates for such locations. The next question is which semantic annotation can be assigned to these extracted locations. This problem could be solved by using spatial proximity to a single or multiple Points of Interest (POI), or by using regional priors based on the general land use of an area to weight certain POI categories.

Sometimes semantic information is directly attached to user traces, so called *semantic trajectories*. The work of Ying et al. (2010) introduces a similarity measure for semantic trajectories in the context of a travel recommendation service. Semantic trajectories are used to analyze the semantic movement behavior of users.

Lin et al. (2018) identify dwell regions using the temporal domain of the data to extract mobility profiles in the semantic space. Furthermore, they introduce a user similarity measure for their comparison and further investigation.

Another approach is realized in the implementation of the *habit2vec* framework proposed by Cao et al. (2020). Using this framework, they encode the semantic and temporal information from the location data using representation learning to investigate the typical habits of the tracked users.

A different approach is presented by Andrienko and Andrienko (2018), who introduce visual interpretation of graph-based representations of states (location regions) overlaid with additional semantic information.

The work most similar to the presented approach is the work published by Lv et al. (2016), who use a hierarchical clustering approach to group extracted trajectory stop points. Furthermore, they use temporal and spatial features to describe the detected physical locations and assign them to a set of predefined semantic types.

In the work of Yang et al. (2018), they extract the user's home and work location by clustering stop points extracted from GPS trajectories, which are additionally annotated with temporal signatures. When considering the temporal domain, they assume a regular movement behavior for each user, e.g., typical for a full-time worker.

Xu et al. (2021) propose an approach using a Hidden Markov Model (HMM) to assign the most probable semantic annotation label to stop point clusters extracted from GPS trajectories.

Recently, Cheng et al. (2022) published their work on semantic location annotation based on POI and Area of Interest (AOI). In addition, they consider features such as *length of stay* or *visiting time* to account for the temporal domain. Furthermore, Li et al. (2023) present meta-graph, a life pattern clustering approach to identify user groups with similar life patterns. Within their approach, they detect different interesting places of users based on clustering and a rule-based system: home places, work places, night places, day places, and other significant places.

## 3. Methodology

The proposed methodology is divided into two parts, as shown in Figure 1. The first part performs a sub-selection of the input trajectories for a single user, while at a later stage each user could be processed in parallel. Then, stop points are extracted from the GPS trajectories using the Python framework MovingPandas (Graser, 2019). A minimum duration of five minutes is set as a constraint for the stop point detection to exclude stops with waiting times at intersections or pick-up/drop-off points.

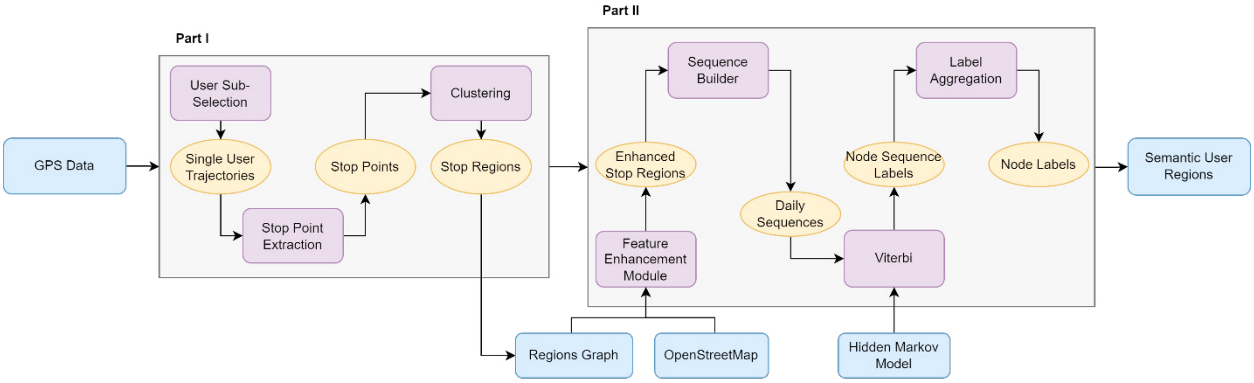

Fig. 1: Overview of the processing steps of the presented approach.

In order to obtain stop point regions, a variant of the popular Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering (hdbscan, McInnes et al. (2017)) is applied to the remaining stop points. The resulting stop clusters highlight the most interesting or characteristic stop regions for a given user. These are used as input for the second part of the presented approach. Furthermore, a graph-based representation of the stop regions is generated and also passed to the second part.

In Part II, the stop regions are enriched with different features of three feature categories: contextual, temporal, and network-based features (see Table 1). While other researchers have proposed subsets of the first two feature categories, using the dimensional aspects of the detected cluster regions (C4) and the network-based feature category is a novel approach.

The enrichment process is performed on two levels of entities. The first level is based on point entities, each individual stop point is annotated with recorded or additional data. The recorded data is mostly bound to the time domain. The additional data for the POI-related information are extracted from the OpenStreetMap (OSM) project OpenStreetMap contributors (2017) and aggregated into meaningful semantic POI groups at different levels of generalization (L-1 to L-3). Here, a higher level provides a deeper semantic distinction, e.g. L-1 represents the most generalized POI groups: *residential*, *transportation*, *public services*, *commercial* and *recreation*.

The second level is based on the detected clusters and deals with areal entities. Here, features are calculated from all stops belonging to the same stop cluster. These features are more general and describe the overall characteristics of the stop cluster. Each stop that falls into the same stop cluster receives the same feature values.

Given the user trajectories and the region graph representation, the daily sequences are generated with respect to the stop regions. In this way, each trajectory is transformed into a stop region sequence.

In order to solve the label assignment task, the Viterbi algorithm (Viterbi, 1967) is used as an optimizer. Here, the hidden states of the HMM correspond to the searched semantic labels of the stop regions and the observations correspond to the estimated features (see Table 1).

Tab. 1: Overview of the three feature categories and their respective features with description.

| | Category | Feature | Level | Description / Definition |
|---|---|---|---|---|
| C1 | Contextual | Closest POI | Point | Closest POI category of a stop point |
| C2 | Contextual | Center POI | Cluster | Closest POI category of a stop region center point |
| C3 | Contextual | Dominant POI | Cluster | Dominant POI category of a stop region based on all contained stop point |
| C4 | Contextual | Inverse Density | Cluster | Inverse of the cluster density (points per km²), describes the compactness of a cluster |
| T1 | Temporal | Arrival time | Point | Arrival time of a stop point at the stop location |
| T2 | Temporal | Stay duration | Point | Stay duration of a stop point |
| T3 | Temporal | avg. stay duration | Cluster | Average stay duration of all stop points in a stop region |
| T4 | Temporal | Weekday | Point | Weekday of a stop point |
| T5 | Temporal | Dominant weekday | Cluster | Dominant weekday of all stop points of a stop region |
| T6 | Temporal | Dominant period | Cluster | Dominant arrival time in a stop region |
| N1 | Network | Amount of start points | Cluster | Percentage of start points in a stop region |
| N2 | Network | Amount of stops points | Cluster | Percentage of stop points in a stop region |
| N3 | Network | Amount of ends points | Cluster | Percentage of end points in a stop region |
| N4 | Network | Degree centrality | Cluster | Node importance base on the number of connections of a respective node |
| N5 | Network | Eigenvector centrality | Cluster | Influence of a node over the whole network |
| N6 | Network | Closeness centrality | Cluster | Network score based on the closeness of a node w.r.t. to all nodes in the network |

Each stop that falls into the same stop cluster receives the same feature values.

The transition and emission probability matrices of the HMM are populated by observations and common knowledge. The probability for a hidden state $state_x$ of the semantic label $x$ is estimated as the multiplication of the observation probabilities $P_i$ with respect to each feature feature$_\mathrm{i}$, according to the following Equation (1):

$$state_x = \prod_i^n P_i(state_x \mid feature_i) \tag{1}$$

According to this equation, each possible state $x$ provides the state that has the highest probability based on the feature input of the stop region sequence and the HMM. Due to the fact that the stop regions are shared by several daily sequences, the assigned semantic labels need to be aggregated to finally determine a semantic label for each stop region. So far, a naive approach has been chosen for the aggregation: majority voting. In this way, the most frequent semantic label is assigned to the stop region.

## 4. Preliminary Results

The presented approach is applied to a smaller test data set collected by a single person over a period of 16 months in the city of Hannover (Germany). An anonymized version of the used trajectory data set is published by Zourlidou et al. 2022. The ground-truth semantics were kindly provided by the authors and are thus known for the most frequently visited locations.

Preliminary results on this data set show that the obtained clusters are highly dependent on the processing steps of the approach in Part I. Here, the stop point detection and the subsequent stop point clustering influence the number and distribution of the resulting user stop regions. Therefore, an empirical adaptation of the stop detection and clustering parameters is applied. The resulting stop regions are considered as the most dominant ones of a user and can therefore be interpreted as locations of special interest for the respective user.

In Part II, the HMM models the above characteristics of the three feature categories. In addition, it is assumed that the model represents a full-time worker, which is consistent with the data set used. Although only the most generalized representation of the POI groups (L-1) is currently included in the processing pipeline, the results already provide a good insight into the potential outcome of the presented approach (see Table 2 and Figure 2).

Tab. 2: Resulting semantic labels of the detected stop regions based on the majority voting compared with the available ground-truth labels.

| Stop Region ID | Samples | Unique Pred. Labels | Ratio of Pred. Label | Pred. Semantic Label | Ground-truth Semantic Label |
|---|---|---|---|---|---|
| 0 | 83 | 1 | 1.0 | leisure time | leisure time |
| 1 | 68 | 2 | 0.72 | leisure time | leisure time |
| 2 | 611 | 2 | 0.99 | home | home |
| 3 | 50 | 2 | 0.52 | work | leisure time |
| 4 | 41 | 1 | 1.0 | leisure time | leisure time |
| 5 | 82 | 1 | 1.0 | work | work |
| 6 | 26 | 1 | 1.0 | work | work |
| 7 | 45 | 1 | 1.0 | leisure time | leisure time |
| 8 | 20 | 1 | 1.0 | leisure time | leisure time |

Comparing the obtained semantic labels of the stop regions with the available ground-truth semantics shows that most of the semantic labels are correctly identified (see Table 2). A slight inaccuracy is present for the semantic labels *work* and *leisure time*. It can be assumed that the majority voting process for the final semantic label of a stop region introduces the inaccuracy. An increasing number of unique predicted labels for a single stop region will lead to a decreasing label ratio and thus to a more uncertain final semantic label.

## 5. Outlook

While the preliminary results are already promising, there are several aspects that need to be addressed in the future work. Currently, the presented approach considers only a simplified user type model, full-time worker, while it can be extended for different user type models, such as part-time worker.

Furthermore, the unique semantic labels (*home*, *work*, *leisure* and *unknown*) in the HMM are very few and provide only a more generalized understanding of the person under investigation. This will be changed in the next step, along with increasing the level of detail for the POI groups to L-2.

For the general determination of the transition matrices, the Baum-Welch algorithm (Welch, 2003) could be used, if ground-truth annotations are widely available with respect to the considered user type model. In this way, no expert knowledge would be required to manually set up the matrices needed in the HMM. In addition, the labeling process could be implemented globally via the Viterbi optimization, so that a specific stop region in one sequence would also receive the same semantic label in another sequence. This leads to a conditional optimization with the possibility to include further constraints, such as restricting each user to a single *home* region.

In addition, the approach will be applied to a larger multi-user trajectory data set to test the applicability of the full-time worker user type model and to investigate differences in the user base.

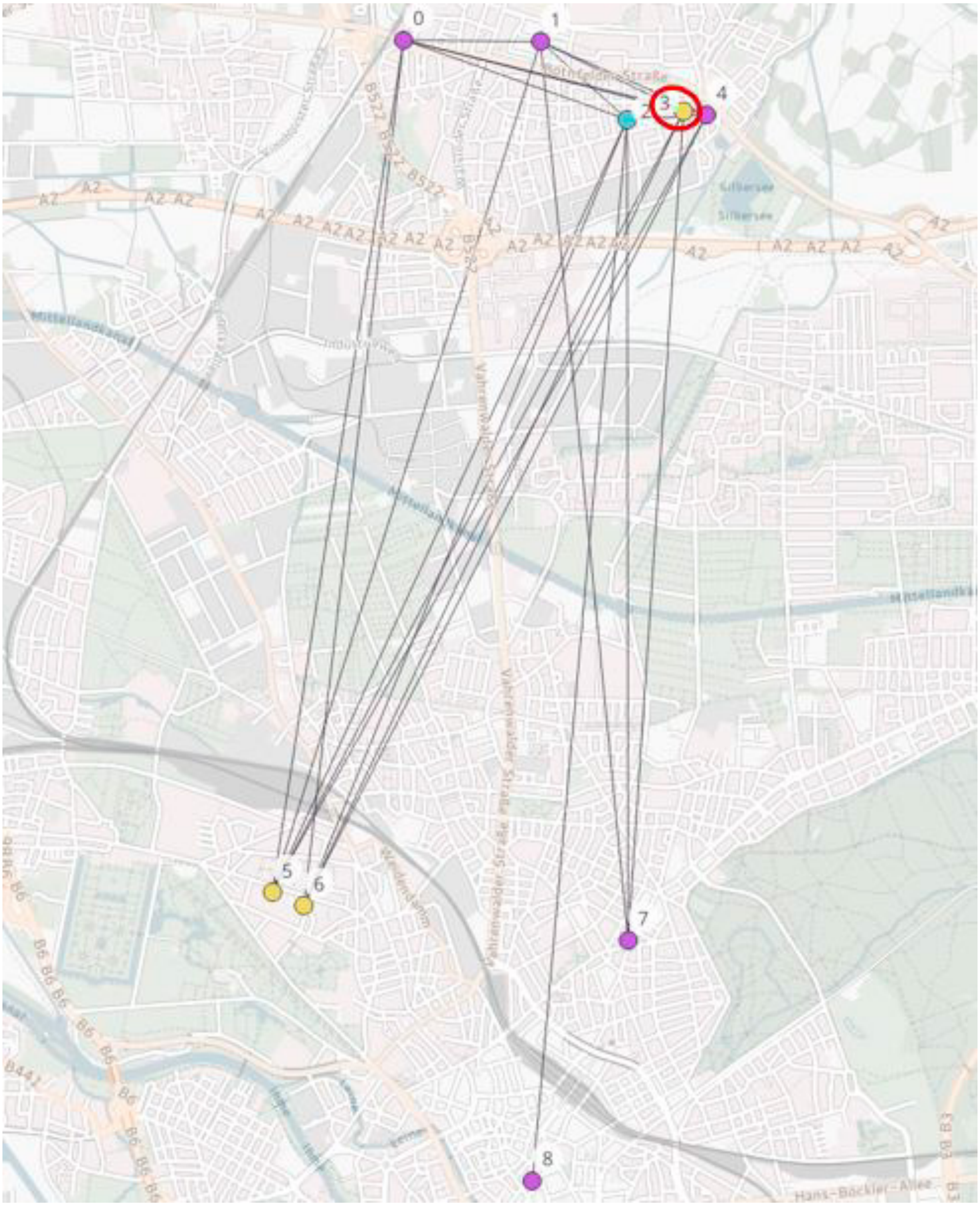

Fig. 2: User stop regions identified by ID and graph-based connections (black). Predicted region labels are color-coded: purple - leisure time, yellow - work and home - cyan.

## References

Andrienko, N., Andrienko, G., 2018. State transition graphs for semantic analysis of movement behaviours. Information Visualization 17, 41–65. doi:10.1177/1473871617692841.

Cao, H., Xu, F., Sankaranarayanan, J., Li, Y., Samet, H., 2020. Habit2vec: Trajectory semantic embedding for living pattern recognition in population. IEEE Transactions on Mobile Computing 19, 1096–1108. doi:10.1109/TMC.2019.2902403.

Cheng, J., Zhang, X., Luo, P., Huang, J., Huang, J., 2022. An unsupervised approach for semantic place annotation of trajectories based on the prior probability. Information Sciences 607, 1311–1327. doi:10.1016/j.ins.2022.06.034.

Feuerhake, U., Kuntzsch, C., Sester, M., 2011. Finding interesting places and characteristic patterns in spatio-temporal trajectories, in: 8th LBS symposium.

Graser, A., 2019. Movingpandas: Efficient structures for movement data in python. GI _Forum 1, 54–68. doi:10.1553/giscience2019_01_s54.

Li, W., Zhang, H., Chen, J., Li, P., Yao, Y., Shi, X., Shibasaki, M., Kobayashi, H.H., Song, X., Shibasaki, R., 2023. Metagraph-based life pattern clustering with big human mobility data. IEEE Transactions on Big Data 9, 227–240. doi:10.1109/TBDATA.2022.3155752.

Lin, Z., Zeng, Q., Duan, H., Lu, F., 2018. Finding similar users from gps data based on assignment problem, in: Ben-Othman, J., Yu, H., Unger, H., Arai, M. (Eds.), Proceedings of the 4th International Conference on Communication and Information Processing, ACM, New York, NY, USA. pp. 283–288. doi:10.1145/3290420.3290470.

Lv, M., Chen, L., Xu, Z., Li, Y., Chen, G., 2016. The discovery of personally semantic places based on trajectory data mining. Neurocomputing 173, 1142–1153. doi:10.1016/j.neucom.2015.08.071.

McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. Journal of Open Source Software 2, 205. doi:10.21105/joss.00205.

OpenStreetMap contributors, 2017. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Viterbi, A., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13, 260–269. doi:10.1109/TIT.1967.1054010.

Welch, L.R., 2003. Hidden markov models and the baum-welch algorithm. IEEE Information Theory Society Newsletter 53.

Xu, H., Ye, Z., Jiao, M., 2021. A poi classification semantic annotation algorithm based on hmm, in: Wu, F., Liu, J., Chen, Y. (Eds.), International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2021), SPIE. p. 93. doi:10.1117/12. 2631298.

Yang, M., Cheng, C., Chen, B., 2018. Mining individual similarity by assessing interactions with personally significant places from gps trajectories. ISPRS International Journal of Geo-Information 7, 126. doi:10.3390/ijgi7030126.

Ying, J.J.C., Lu, E.H.C., Lee, W.C., Weng, T.C., Tseng, V.S., 2010. Mining User Similarity from Semantic Trajectories. ACM Conferences, ACM, New York, NY. doi:10.1145/1867699.

Zourlidou, S., Golze, J., Sester, M., 2022. GPS Trajectory Dataset of the Region of Hannover, Germany. doi:10.25835/9BIDQXVL.