BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Extracting user's significant places from location data

Bc. Alexandra Ligocká*

Abstract

How can Google or an iPhone guess the location of your home and work or places you visit? This work explores the process by which a user's home or work location can be determined based on raw GPS data collected from GPS-enabled devices. It investigates the feasibility and challenges involved in extracting this information from such data. Based on existing solutions and approaches, a framework was designed to extract the user's home and work locations and places they visit. The framework addresses key issues, including extracting points from GPS traces, identifying locations with a higher level of significance for users, extracting visited places and their semantic enrichment and interpretability for users. This paper further describes the steps involved in GPS data analysis for this task in more detail, existing approaches to solve given issues, and proposed solution as well as results obtained.

*xligoc03@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

With the rise of GPS-enabled devices such as smartphones, the volume of tracking data collected on users is also increasing. This data is utilized by various location-based services across various sectors, including social media, e-commerce, transportation, and healthcare, and can benefit users and service providers. However, raw location data often lack the semantic richness required to provide additional contextual information and deeper insights into user behaviour. Despite the growing quantity and precision of location data obtained from mobile devices, its semantic quality remains a concern. To address this issue, I have developed a framework that focuses on identifying a user's significant locations, including their home and work locations and places where they tend to spend more time.

The task of extracting personally interesting places involves various sub-tasks, with physical place extraction and semantic place recognition being the main ones [1]. Therefore, the primary challenges faced include identifying stops in raw GPS data and differentiating between different types of stops, such as a brief pause or more extended stay. Additionally, associating semantic meaning with these stops, such as identifying a place that belongs to a given stop and extracting places with higher levels of personal importance, such as home and work, pose additional challenges. Last but not least is the issue of the interpretability of results for end-users.

Different techniques have been proposed by the research community to tackle the major challenges in mining a user's significant location. Several methods have been suggested to detect stay points, including the distance-based approach, time-based approach, and density-based approach [2]. To address the issue of variability of stay points' spatial coordinates, most researchers incorporate aggregation of stay points into stay regions using clustering algorithms. Some of the commonly used clustering algorithms are centroid-based, such as K-means [3, 4], density-based CB-SMOT [5] and OPTICS [6] derived from DBSCAN or hierarchical clustering algorithms. Semantic enrichment is performed on the extracted stay regions to broaden the semantic context of locations. Approaches to enrich location with semantic context range from basic reverse geo-coding techniques [7] to more sophisticated approaches using point of interest (POI) databases [8]. Mapping a location to POI can be performed using predefined POI categories with a combination of spatial and temporal rules [8].

2. An overview of proposed framework

In the previous section, I briefly described key issues in mining users' significant locations. This section describes the proposed framework.

2.1 Input dataset

As a part of my work, I created a sample dataset containing GPS data - a series of points consisting of coordinates (latitude, longitude) and timestamps. I gathered the data using the Google Maps Platform and designed the framework to process raw GPS data. Thus, I eliminated extraneous data from the Google-exported dataset. This approach yields several benefits. For instance, it reduces the amount of data that needs to be processed and analyzed, which lowers the computational resources. Furthermore, it makes the system more widely applicable to a wider range of input datasets that contain only this essential information and reduces storage demands. Figure 1 shows the input dataset visualised using OpenStreetMap (OSM) background tiles.

2.2 Proposed solution

Based on an analysis of existing approaches and methods, I designed a framework for retrieving important user locations and their semantic labelling. This subsection describes the steps and algorithms involved in each step of data processing.

Stay-points detection is a key part of the whole system. In this part, stay points are extracted using a differential-based stay-point detection algorithm based on seeking the spatial region within a given radius where a user spent a given amount of time. The algorithm is based on the algorithm proposed in [9]. Both time and distance rules are used in order to detect stops where the user has stopped for an amount of time or wandered around the target place in a given distance. These two types of stops are shown in Figure 2. In addition to extracting stay points, the algorithm also calculates the estimated departure time. The process of **aggregation points into locations** involves taking a set of GPS stay-points and grouping them together based on their spatial proximity to form a single location. This approach is particularly useful when dealing with GPS data where multiple stay-points may correspond to the same physical location but have slightly different latitudes and longitudes due to GPS inaccuracies and other factors. I have used HDBSCAN clustering algorithm in this work because it works well with datasets of varying densities and complex geometries. Figure 3 demonstrates a zoomed-in result of clustering to illustrate the fact that multiple GPS points may correspond to the same physical location, such as a single building. **Semantic enrichment** refers to the process of augmenting GPS data points with additional contextual information, such as the name or type of the place. This work aims to find the user's home and work locations and other places the user visited. Thus semantic enrichment is divided into two stages. In both stages, OSM database is used to obtain additional information about places. To obtain correct address information, I incorporate reverse geo-coding techniques. **Home and work locations** are extracted using three conditions. First, the framework queries POI database to obtain the category for each building in a given cluster, then computes the proportions of each category; second, I define typical arrival/departure times; and third, the length of each stay is computed. Step of **mapping places to POI** involves querying POI database as well with a defined spatial rule based on maximum distance and POI selection based on the temporal domain rule defined by the availability of given POI. Putting all together, the result is visualised using OSM map tiles.

3. Results

Output from the framework includes a JSON file with places the user visited and a map of given places. Figure 6 shows obtained results for home and work locations. Sub-figures a) and b) show annotations of places with corresponding semantic meanings and addresses. Extracted places mapped to POI are shown in Figure 7 using icons to differentiate categories of POIs.

I validated results using a comparison with Google Maps, visualisation of this comparison is shown in Figure 8. According to the results, the framework correctly detected the user's home and work locations. Additionally, it extracted 128 places, of which 118 were correctly mapped to POI. The framework missed three places, resulting in an accuracy of 90%.

4. Conclusions

After experimenting with various algorithms and approaches, I have developed a solution that can accurately detect the user's home and work location while also identifying locations the user has visited with 90% accuracy. However, compared to Google Maps results, there are still some locations that my algorithm fails to detect, which presents an area for further improvement.

References

[1] Mingqi Lv, Ling Chen, Zhenxing Xu, Yinglong Li, and Gencai Chen. The discovery of personally semantic places based on trajectory data mining. *Neurocomputing*, 173:1142–1153, 2016.

[2] Rafael Pérez-Torres, César Torres-Huitzil, and Hiram Galeana-Zapién. Full on-device stay points detection in smartphones for location-based mobile applications. *Sensors*, 16(10):1693, Oct 2016.

[3] Xin Cao, Gao Cong, and Christian S. Jensen. Mining significant semantic locations from gps data. *Proceedings of the VLDB Endowment*, 3(1-2):1009–1020, 2010.

[4] Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[5] Andrey Tietbohl Palma, Vania Bogorny, Bart Kuijpers, and Luis Otavio Alvares. A clustering-based approach for discovering interesting places in trajectories. New York, NY, USA, 2008. Association for Computing Machinery.

[6] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. *Proceedings of the 18th international conference on World wide web*, 2009.

[7] Juhong Liu, O. Wolfson, and Huabei Yin. Extracting semantic location from outdoor positioning systems. *7th International Conference on Mobile Data Management (MDM'06)*, 2006.

[8] Barbara Furletti, Paolo Cintia, Chiara Renso, and Laura Spinsanti. Inferring human activities from gps tracks. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 2013.

[9] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. New York, NY, USA, 2008. Association for Computing Machinery.