International Cartographic Association
Association Cartographique Internationale

Semantic trajectory inference from geo-tagged tweets

Qunying Huang[a], *, Xinyi Liu[a]

[a] *Department of Geography, University of Wisconsin -Madison, qhuang46@wisc.edu, xliu636@wisc.edu*

* Corresponding author

Keywords: Social media, Data mining, Big data, VGI, Digital footprints, Human Mobility

Abstract:

Individual travel trajectories denote a series of places people visit along the time. These places (e.g., home, workspace, and park) reflect people's corresponding activities (e.g., dwelling, work, and entertainment), which are discussed as semantic knowledge and could be implicit under raw data (Yan et al. 2013, Cai et al. 2016). Traditional survey data directly describe people' activities at certain places, while costing tremendous labors and resources (Huang and Wong 2016). GPS data such as taxi logs record exact origin-destination pairs as well as people's stay time along the way, from which semantics can be easily inferred combining with geographical context data (Yan et al. 2013). Research has been done to understand the activity sequences indicated by either individual or collective spatiotemporal (ST) travel trajectories using those dense data. Different models are proposed for trajectory mining and activity inference, including location categorization, frequent region detection, and so on (Njoo et al. 2015). A typical method for matching a location or region with a known activity type is to detect stay points and stay intervals of trajectories and to find geographical context of these stay occurrences (Furtado et al. 2013, Njoo et al. 2015, Beber et al. 2016, Beber et al. 2017).

However, limited progress has been made to mine semantics of trajectory data collected from social media platforms. Specifically, detection of stay points and their intervals could be inaccurate using online trajectories because of data sparsity. Huang et al. (2014) define the notion of activity zone to detect activity types from digital footprints. In this method, individual travel trajectories first are aggregated using spatial clustering method such as density-based spatial clustering of applications with noise (DBSCAN). Then produced clusters are classified based on a regional land use map and Google Places application programming interface (API). Such land use data are only published at specific places, such as the state cartography office's website at University of Wisconsin-Madison. Researchers need to search for those data based on their study area. Moreover, while major land use maps can be searched for large areas such as the whole United States, detailed land use data for statewide or citywide areas are made in diverse standards, which adds extra work to classify activity zones consistently. Besides, Google Places API is a service that Google opened for developers and will return information about a place, given the place location (e.g., address or GPS coordinates), in the search request. However, API keys need to be generated before we can use these interfaces and each user can only make a limited number of free-charged requests every day (i.e., 1,000 requests per 24 hours period). In sum, previous methods to detect activity zone types using social media data are not sufficient and can hardly achieve effective data fusion. Comparing to the high cost of using officially published dataset, emerging Volunteered Geographic Information (VGI) data offer an alternative to infer the types of an individual's activities performed in each zone (i.e., cluster).

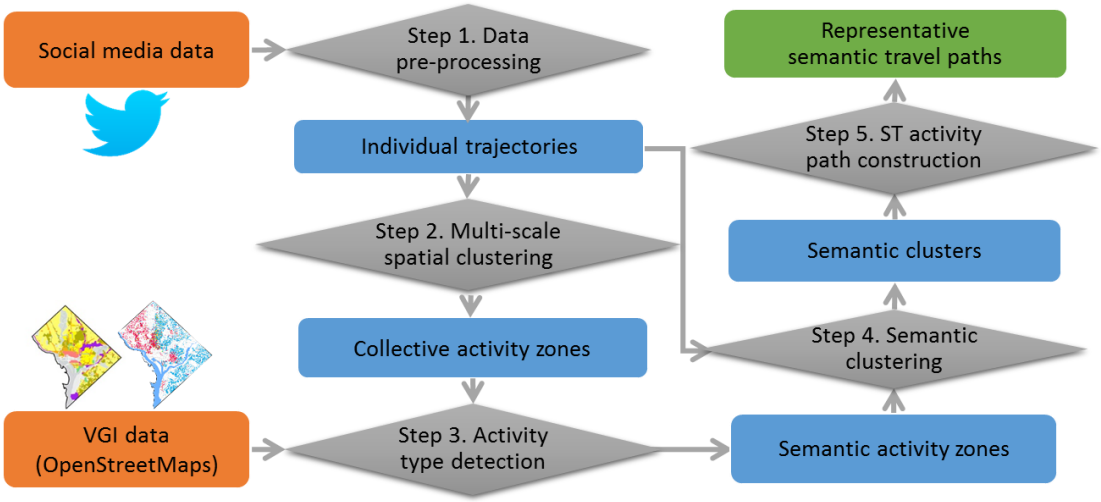

Figure 1. Workflow of semantic trajectory mining

Using geo-tagged tweets as an example, this research proposes a framework for mining social media data, detecting individual semantic travel trajectories, and individual representative daily travel trajectory paths by fusing with VGI data,

specifically OpenStreetMap (OSM) datasets. First, inactive users and abnormal users (e.g., users representing a company with account being shared by many employees) are removed through data pre-processing (Step 1 in Figure 1). Next, a multi-scale spatial clustering method is developed to aggregate online trajectories captured through geo-tagged tweets of a group of users into collective spatial hot-spots (i.e., activity zones; Step 2). By integrating multiple OSM datasets the activity type (e.g., dwelling, service, transportation and work) of each collective zone then can be identified (Step 3). Each geo-tagged tweet of an individual, represented as a ST point, is then attached with a collective activity zone that either includes or overlaps a buffer zone of the ST point. Herein, the buffer zone is generated by using the point as the centroid and a predefined threshold as the radius. Given an individual's ST points with semantics (i.e., activity type information) derived from the attached collective activity zone, a semantic activity clustering method is then developed to detect daily representative activity clusters of the individual (Step 4). Finally, individual representative daily semantic travel trajectory paths (i.e., semantic travel trajectory, defined as chronological travel activity sequences) are constructed between every two subsequent activity clusters (Step 5). Experiments with the historic geo-tagged tweets collected within Madison, Wisconsin reveal that: 1) The proposed method can detect most significant activity zones with accurate zone types identified (Figure 2); and 2) The semantic activity clustering method based on the derived activity zones can aggregate individual travel trajectories into activity clusters more efficiently comparing to DBSCAN and varying DBSCAN (VDBSCAN).

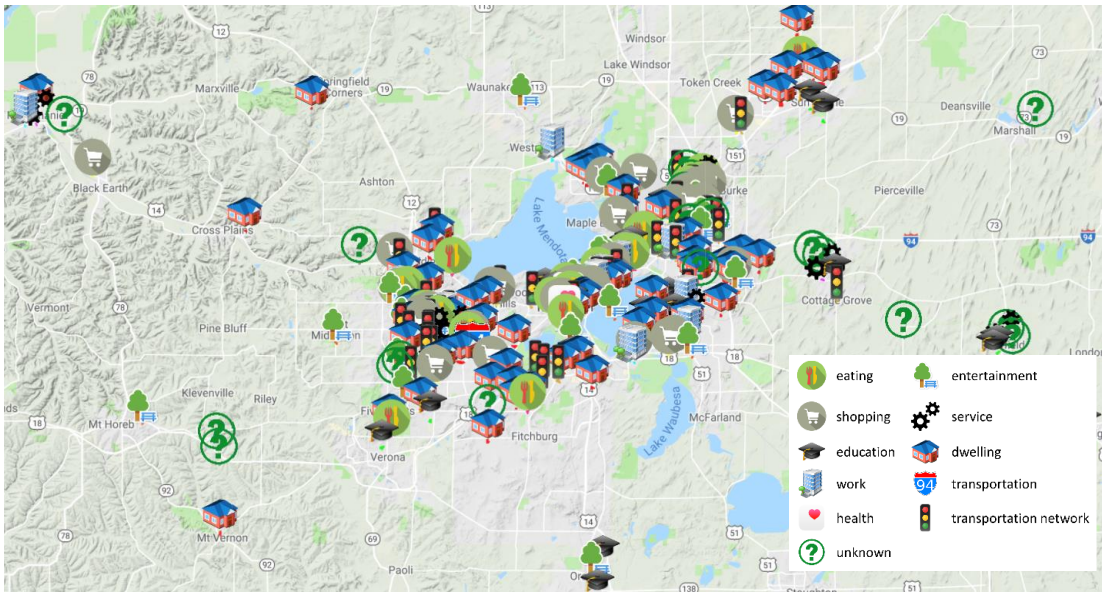

Figure 2. Detection of semantic activity zones at Madison, Wisconsin

Reference

1. Ahlers, D., 2013. Assessment of the accuracy of GeoNames gazetteer data. Proceedings of *the 7th Workshop on Geographic Information Retrieval*. Orlando, Florida: ACM, 74-81.
2. Beber, M. A., Ferrero, C. A., Fileto, R., & Bogorny, V. (2016). Towards activity recognition in moving object trajectories from Twitter data. In *GeoInfo* (pp. 68-79).
3. Beber, M.A., et al. 2017. Individual and Group Activity Recognition in Moving Object Trajectories. *JIDM*, 8, 50-66.
4. Huang Q., Cao G., Wang C., 2014. From Where Do Tweets Originate? - A GIS Approach for User Location Inference. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN '14)*, ACM SIGSPATIAL 2014, Nov 6-9, Dallas, TX.
5. Huang, Q. and Wong, D. W. S. 2016. Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, 30(9), 1873-1898.
6. Cai, G., Lee, K., & Lee, I. (2016, January). Mining semantic sequential patterns from geo-tagged photos. In *2016 49th Hawaii International Conference on System Sciences (HICSS)* (pp. 2187-2196). IEEE.
7. Fonte, C. C., et al. 2015. Usability of VGI for validation of land cover maps. International Journal of Geographical Information Science, 29(7), 1269-1291.
8. Njoo, G. S., Ruan, X. W., Hsu, K. W., & Peng, W. C. (2015, October). A fusion-based approach for user activities recognition on smart phones. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on* (pp. 1-10). IEEE.
9. Urtado, A. S., Fileto, R. and Renso, C. 2013. Assessing the Attractiveness of Places with Movement Data. Journal of Information and Data Management, 4(2).
10. Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., & Aberer, K. (2013). Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(3), 49.