THUMP: Semantic Analysis on Trajectory Traces to Explore Human Movement Patterns

Shreya Ghosh
Indian Institute of Technology, Kharagpur, India
shreya.cst@gmail.com

Soumya K. Ghosh
Indian Institute of Technology, Kharagpur, India
skg@iitkgp.ac.in

ABSTRACT

Exploring human movement pattern from raw GPS traces is an interesting and challenging task. This paper aims at analysing a large volume of GPS data in spatio-temporal context, clustering trajectories using geographic and semantic location information and identifying different categories of people. It tries to exploit the fact that human moves with an intent. The proposed framework yields encouraging results using a large scale GPS dataset of Microsoft GeoLife.

Keywords

Trajectory; GPS Data; GeoCoding; Geo-tagging; Categorization; Clustering

1. INTRODUCTION

With the proliferation of mobile phone users and GPS-enabled mobile devices, a large volume of GPS trajectory data is currently available. Incidentally, human movements are usually with some intent. However, these raw trajectory data collected from different sources are unstructured and it is a challenge to infer any meaningful pattern from there. The GPS data has been used for several location based services and applications, like trajectory matching [1], extracting classical travel sequences and top k interesting locations [2] etc.

In this work, we attempt to categorize people (or mobile users) based on their GPS traces. This involves (i) efficient handling of raw GPS traces, (ii) finding trajectories of individual users, (iii) spatio-temporal pattern matching of trajectory segments (iv) clustering and extracting human movement patterns and deduce semantic meaning correlated with the geographical regions. Most of the existing studies have focused on analysis and clustering of raw GPS trajectories without any semantic enrichment. In this work, we consider both spatio-temporal and semantic aspects of GPS traces.

The novelty of the work lies in (i) developing efficient storage mechanism for large volume of GPS traces, (ii) spatio-temporal clustering and extracting common patterns using semantic knowledge of the locations and (iii) categorization of users based on spatio-temporal and semantic features of the trajectories, (iv) finding the correlation/association between *user-user*, *user-place* and *place-place* automatically from the GPS traces.

Copyright is held by the author/owner(s).
WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2893188 .

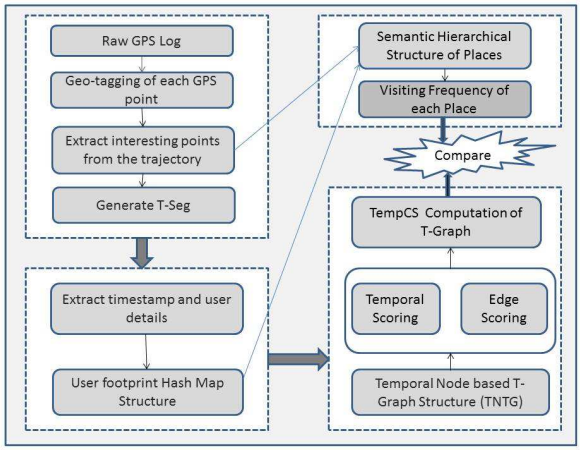

Figure 1: Architecture of THUMP framework

A motivating example may be considered to understand features of our system. For example, "Find some behavioural patterns between two sets of users." Here, in one of our case study, we consider set of users who visit gym or some health care center, and another set of users who visit hospital less frequently. Our experiment shows a correlation between these two set of users and computes 70% support value for this statement.

2. THUMP FRAMEWORK

Figure 1 shows the overall architecture of THUMP framework. Raw GPS data is taken as input to the system. We introduce *Geo-tagged GPS Log*, where each GPS point is associated with most appropriate land use information (e.g., University or Health Care Center) by reverse geo-coding and majority voting technique. Here, each GPS point p_i contains land use information $(p_i, place)$ along with *latitude*, *longitude* and *timestamp*. The complete trajectory of a user is segmented in multiple numbers of trajectory segments, *T_Seg*s, based on time spent on a particular place *(stay point)*. Stay points of each trajectory are identified from temporal and spatial distribution of the GPS traces.

Indexing Strategy to speed up Computation: Efficient storage of huge volume of trajectories is a big challenge. We have introduced UFHM *(User Footprint Hash Map Structure)* to index all geo-tagged information along with user details. $UFHM(H) = \{(B_1, L_1), ..., (B_n, L_n)\}$. H is a chain hash map structure and each bucket (B_i) of UFHM is associated with a list (L_i). $B_i =< Lat, Long, place >$: each tuple stores latitude, longitude along with land use information. Each list contains user details, geo-tagged places and timestamp of visiting the place. $L_i =< u_i, Geo_{tag}, t_i >$. A pairing function $H(l_x, l_y)$ is used as the hash function to map normalized latitude (l_x) and longitude (l_y) in *UFHM*.

$$H(l_x, l_y) = (l_x + l_y)(l_x + l_y + 1)/2 + l_y \quad (1)$$

UFHM structure provides efficient access to a particular geographic region and interesting land use information.

Spatio-Temporal Clustering and Extracting Common pattern: To analyse movement pattern, TNTG *(Temporal Node based T-Graph Structure)* is generated for each user's trajectory path. $TNTG = \{(V, E)|1 < v_i < |V|, 1 < e_i < |E|\}$, where each node $v_i \in |V|$ denotes stay point of the trajectory segment. We have represented v_i as **Temporal Node** (TN). $TN =< node_{id}, user, time >$. Each TN stores footprints of users who visited that particular node, a unique node id and stay duration. Directed edges from one (TN) to another defines the transition from one place to another place based on time series of GPS traces. To cluster trajectory segments, location (geographical and land use) information, time duration in a stay-point, speed of movement, transportation mode and direction of the trajectories are considered. Each TN and edge of TNTG are weighted based on the above features. To measure similarity among trajectory segments of different users, minimum stay duration in a common node is computed along with other mentioned features, namely geo-tagged information of the stay points (e.g., university, restaurant), transportation mode (e.g., cycle, car), timestamp of the visits etc. Based on the similarity measurement the users' trajectories are clustered.

For finding common patterns among a set of trajectories an extension of LCS (*Longest Common Sub-sequence*) problem, namely *Temporal Common Sub-sequence* (TempCS) is introduced. $TempCS(X_i, Y_j)$ finds common sub-sequence among trajectories X and Y with i and j stay points respectively as depicted in (2).

$$TempCS(X_i, Y_j) = \begin{cases} 0 & if(i == 0) \\ & or(j == 0) \\ TempCS(x_{i-1}, y_{j-1}) & if((x_i == y_j) \\ +Min(XT_{Score_i}, YT_{Score_j}) & and(x_{i+1}) \neq (y_{j+1})) \\ TempCS(X_{i-1}, Y_{j-1}) + C\times & if((x_i == y_j) \\ Min(XT_{Score_i}, YT_{Score_j}) & and(x_{i+1}) == (y_{j+1}) \\ MAX(TempCS(X_{i-1}, Y_j), & \\ TempCS(X_i, Y_{j-1})) & if(x_i \neq y_j) \end{cases} \quad (2)$$

TempCS for a cluster of users trajectory returns a set of TN, average stay duration in each TN and directed edges which represents common trajectory path followed by all users in the particular cluster.

Categorization and Correlation: Each common trajectory path of clusters is analysed to categorize users based on the movement patterns. In our experiment (using GPS dataset of Microsoft GeoLife [2]), users' footprints on frequently visited places, namely student dormitory, professor office, laboratory etc. are analysed. Users are categorized into four broad categories, namely $Student$, $Professor$, $Staff$ and $Guest$ according to the frequency of visiting places, timestamp of the visit and few assumptions, like students visit library, laboratory, student-cafe more frequently than professor/staff etc.

Given a dataset of *user-user* or *user-place* or *place-place* pair, association/correlation can be automatically determined using THUMP framework.

3. EVALUATION OF THUMP

In this section, we evaluate the effectiveness of the THUMP framework using the real data set of GeoLife Trajectory [2] with GPS traces of 182 users in a period of over five years around Beijing, China and over 17,621 trajectories. Few representative results are presented here.

1. Computational speed-up: Using *UFHM*, search time reduction is captured in a dataset of 258 places/ locations.

Type of Query	Search Time (UFHM)	Linear Scan (Naive)
Point	6.3s	18.5s
Range	8.6s	41.2s

Table 1: Computational efficiency of THUMP

2. Association/Correlation: Association/Correlations between GPS traces of users are determined from *TNTG* and *TempCS*. One of the generated association rules is shown below:

$$Users(Visits\ Health - care\ center\ frequently) \rightarrow Users(Less\ Hospital\ Visits) \quad [Support: 70\%, Confidence: 80\%] \quad (3)$$

4. CONCLUSION

This paper presents a novel framework for analysing the mobile trajectory traces in spatio-temporal context, model human movement patterns using semantic aspects and categorize people based on these patterns. It also helps in extracting interesting association rules from the GPS traces.

5. REFERENCES

[1] S. Shang, R. Ding, K. Zheng, C. S. Jensen, P. Kalnis, and X. Zhou. Personalized trajectory matching in spatial networks. *In The VLDB Journal, The International Journal on Very Large Data Bases*, 23(3):449–468, 2014.

[2] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. *In Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.