Modeling of Human Movement Behavioral Knowledge from GPS Traces for Categorizing Mobile Users

Shreya Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur, India
shreya.cst@gmail.com

Soumya K Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur, India
skg@iitkgp.ac.in

ABSTRACT

Human movement analysis and categorization of mobile users based on their movement semantics are challenging tasks. Further, due to security and privacy issues, insufficient labelled or user-annotated data (or, ground-truth data) makes the user-classification from GPS traces more complex. In this work, we present a framework which models user movement patterns containing both spatio-temporal and semantic information, generates semantic stay-point taxonomy by analysing GPS traces of all users, summarizes individuals' GPS traces and clusters users based on the semantics of their movement patterns. To alleviate labelled data scarcity problem while user categorization in a particular region of interest (ROI), we propose a method to transfer knowledge derived from a set of GPS traces of a geographically distanced but similar type of ROI. An extensive set of experiments using real GPS trace dataset of *Kharagpur, India* and *Dartmouth, Hanover, USA* have been carried out to demonstrate the effectiveness of our proposed framework.

Keywords

Trajectory, GPS Data, GeoCoding, Geo-tagging, Categorization, Clustering, Spatial Transfer Learning

1. INTRODUCTION

The growing popularity of GPS embedded devices have motivated extensive research on analysing the voluminous GPS traces and various location-aware applications. Further, mobile users are capable to accumulate their own GPS logs conveniently (ex. Google Map Timeline) leading to generation of huge amount of location traces. It provides unprecedented opportunities to analyse and derive valuable knowledge of human mobility patterns, specifically, human interests and intentions which in turn facilitates varied location based services. The question is: "Can we map the knowledge of one known region to another unknown (target) region and use this knowledge to categorize the users in the target region?"

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054150

With the huge volume of collected GPS data, another interesting and obvious question arises that whether all the GPS points have same information or some of the GPS points carry more contextual information than rest. To this end, we aim to model and efficiently store user-movement traces without losing any significant information. However, instead of raw GPS log (time-stamped latitude, longitude data) human movement patterns are better understood when some contextual information, landmarks on the route, duration of stay points or activities performed at stay points are considered. Motivated by the potential merits, recently research trend is to extract semantic information or capturing inherent meaning of these huge volume of human movement data. This semantic enrichment of raw GPS log bridges the gap between collected GPS traces and various location based applications. We aim to capture behavioral differences in the movement patterns of the individuals and utilize the knowledge to cluster users having similar movement patterns. The problem becomes quite challenging for the unpredictable behavior of human mobility. For example, user X and user Y both are students, but may take different routes to university. To tackle this challenge, we propose Bayesian network for modelling user's movement pattern which captures probabilistic measures of the randomness of movement. Although advances in location-acquisition techniques have generated huge amount of GPS data, but unfortunately, scarcity of user-annotated or labelled data is still a major challenge in categorization of users. Therefore, we aim to learn movement patterns of a known region (source) and map the knowledge of the source region to another region (destination) of same domain (say, academic, commercial etc.). This problem of trajectory knowledge transfer has not been reported well in the existing literature.

To address the above mentioned challenges and issues, we propose a framework which involves (i) generating user trajectory from raw GPS log (ii) creating *User Trace Model* to represent individuals movement behavior, generating place knowledge base of the region from the GPS traces (iii) spatio-temporal movement pattern mining and similarity measurement (iv) transfer the human movement behavior knowledge to other geographical place.

The remainder of the paper is organized as follows: section 2 provides state-of-the-art works, section 3 introduces the proposed framework and gives a brief description of it. Section 4 depicts trajectory summarization, knowledge mining and classification and section 5 describes the method to transfer

the trajectory knowledge to different region. Finally, section 6 provides description of dataset and some experimental results and we conclude in section 7.

2. RELATED WORKS

The problem of human movement pattern analysis has been studied in recent years. In this section we brief some of the recent significant research works in three broad categories namely, *semantic trajectory processing*, *trajectory clustering and classification* and machine learning method *transfer learning* to map the knowledge from one region to another domain of interest.

One of the major challenges in any supervised learning technique is insufficient labelled or training dataset. Also, training and future data may not be in same data distribution or same feature space. To tackle this issue, researchers use a novel learning technique, Transfer Learning [14] to avoid the expensive data-labelling efforts. [20] provides novel learning scenario, heterogeneous transfer learning, where no correspondence between data instances of source and target domain is provided. Enabling knowledge transfer between domains with multimodal data has been carried out in [23],[13], [17]. [22] captures the multi-view nature of various real-life data set and transfers both model parameters and instances to target domain. Fang et. al [5] proposes a method called DISMUTE, which leverages discriminative feature selection for multi-view cross-domain learning and used in the application of object identification and image classification against. Most of the studies on transfer learning focuses on text document classifications [3], [2], classification of webpages, email-spam detection etc [23]. [13] introduces a novel problem of future health prediction from multimodal observation. To the best of our knowledge, there are only few attempts to resolve urban issues and challenges using transfer learning. Wei et. al [19] proposes FLORAL method to transfer knowledge from a city to another where data is insufficient using multimodal transfer learning. The paper learns semantically related dictionaries from multimodal data and predicts air-quality in three different cities.

Ongoing research trend is devoted to overcome the semantic gap between raw GPS log collected from mobile devices and actual personal activity performed in that particular location. Most of the recent studies append contextual information along with the time-stamped latitude, longitude information for enhancing semantic richness of the trajectories. [15] provides a systematic overview of the recent trends of capturing semantics of the movement patterns rather than analysing raw GPS traces. Semantic trajectory is defined based on the application's requirements, namely appending transportation mode, [24] street information, point of interests [6], city map etc. [16] presents a novel idea to map a syntactic trajectory to a semantic trajectory. Based on the movement pattern discovery and human behavior inference, it formalizes a semantic-enriched knowledge discovery process. Trajectory segmentation is another pre-processing step of trajectory data mining. In several work, trajectory is shown as sequence of stop and moves [27],[26]. Yu et. al [26] presents a modeling human location history and mining correlations between different locations. [9] presents a framework to detect semantic places automatically from GPS trajectories. The paper presents a Bayes classifier to categorize trajectory stop-points into predefined category of places. There are various applications (e.g, next location prediction [21], travel route recommendation, hot-spot detection) and methods of trajectory clustering and pattern mining. Recent studies [10], [27] propose various similarity measurements and novel methods to forecast next move of an user and travel sequences. Various spatio-temporal clustering [27], [26], [6] based on the application have been introduced in the last decade. [27] introduces HITS-based model to mine users' interesting location and clustering the movement patterns. Ghosh et. al [6] proposes spatio-temporal clustering TempCS, variant of LCS to extract the common movement patterns of users.

To the best of our knowledge, only a very few studies [6], [25], [8] have focused on mobile-user categorization or finding similar users. [4] has presented an innovative problem to catch pick pocket suspects from large scale transit records. We have proposed an end-to-end framework to model human movement behavioral knowledge and categorize mobile-users. The novelty of the present work is to transfer the knowledge of human movement pattern to another region and finding similar user movements.

3. PROPOSED FRAMEWORK

In this section, we present architecture of our proposed framework. We explain the terms *Semantic-Trajectory*, *User-trajectory Segment*, $SSP_{Taxonomy}$ and *User-Trace Summary*.

Basically, a GPS log is a collection of time stamped GPS points P= $\{p_1, p_2, ... , p_n\}$. Each GPS point $p_i \in P$ contains latitude (p_i, Lat), longitude $(p_i, Lngt)$ and timestamp (p_i, t_i) when the point is captured. [8]

1. **User Trajectory Segment**
   User Trajectory Segment is a triple of $< S[], W[], Traj\_Win[] >$ We represent S[] as list of stay-points, s= $< lat, lon, Geo_{tagg} >$ where within $d > D_{thres}$ distance user stays $T > t_{thres}$ time and Geo_{tagg} is present within r radius of the point.
   Here, Geo_{tagg} or ***Geotagged Stay Point*** is introduced where each GPS stay point is associated with the nearest land use information. Each GPS point p_i contains $(p_i, place)$ along with latitude, longitude and timestamp information

   W[], list of waiting points, $w =< lat, lon >$ is similar to stay points, but no land use is present within r radius of the point.

   $Traj\_Win = \{S_1, (x_1, y_1), (x_2, y_2), ..., S_2)\}$ where, S_1 and S_2 are two consecutive stay-points of the trace.

2. **Semantic Stay Point Taxonomy or ($SSP_{Taxonomy}$)**
   $SSP_{Taxonomy} =< N, N_c, W >$; where, N represents place type at different height of the Taxonomy and N_c associated code of the node-place, W denotes aggregated footprints of the users in that particular node. $SSP_{Taxonomy}$ represents the type of stay points of users (ex, university, food-joints etc) in a hierarchical manner.

3. **User-Trace Summary(UTS)**
   We define *User-Trace Summary* as probabilistic directed graph $G = (V, E)$, where each node represents

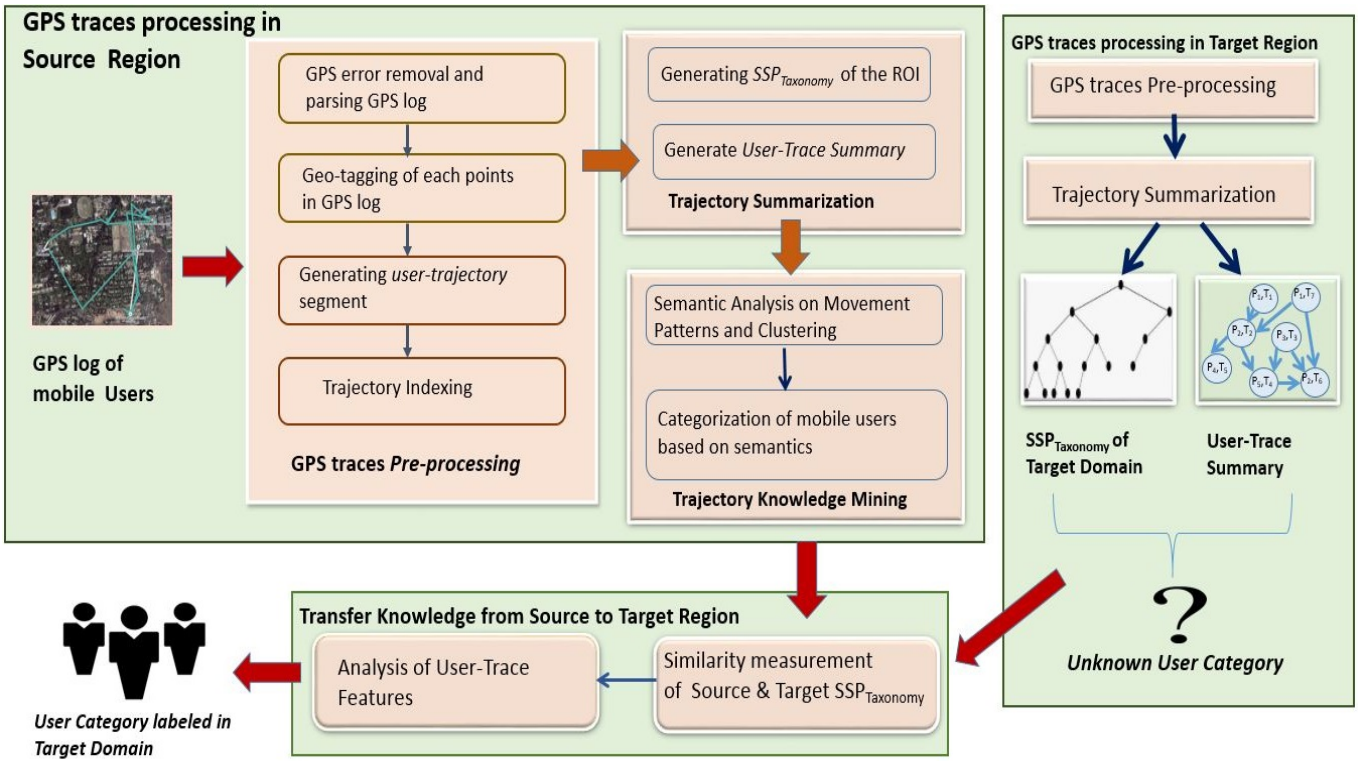


Figure 1: Architecture of the proposed framework

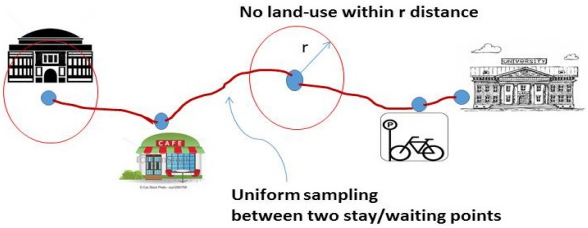


Figure 2: Snapshot of User-Trajectory Segment

a random variable, consisting the visited place of the individual along with the timestamp and time spent at the point.

We propose modelling of *User-Trace Summary* of individuals in three different categories, namely *Weekday UTS*, *Weekend UTS* and *Anomaly UTS*. Here, Weekday UTS and Weekend UTS represents user movement patterns in weekdays and weekends respectively. Anomaly UTS represents movement patterns which are not in the defined ROI or deviates from the regular movement pattern. Generation of UTS is explained in section 4.

4. **Semantic-Trajectory Trace**
Semantic-Trajectory Trace (or STT) is defined as

i) Each of the staypoint of the trajectory segment is geo-tagged or ii) For a particular ROI R_1 and corresponding $SSP_{Taxonomy}$, STT is a User-Trace summary which represents signature movement pattern of an user-category in R_1.

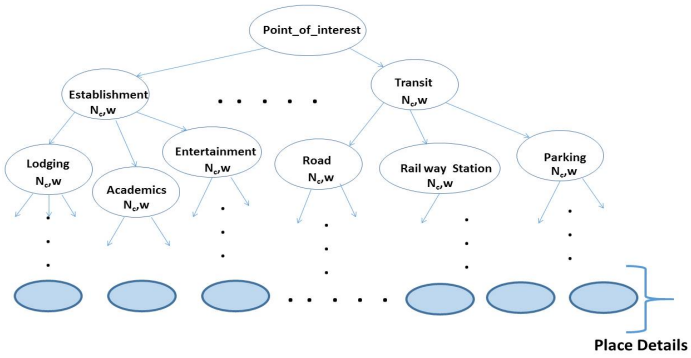


Figure 3: Snapshot of Semantic Stay Point Taxonomy

For geo-tagging of stay-points, we extracted geo-tagged location information from Google-Map using iterative reverse geocoding and refinement of the geo-tagged data. To use this geo-tagged data in trajectory analysis, a systematic model to capture the semantics of the data is needed. We converted the geo-tagged data to a $SSP_{Taxonomoy}$ and weighted each node of the taxonomy with the visit-frequency all of the users. Figure 1 shows the overall architecture of the framework. The 'GPS traces Pre-processing' module generates user-trajectory segments from input raw GPS log, 'Trajectory Summarization' module builds the place taxonomy of the region and creates UTS for each user and 'Trajectory Knowledge Mining' module clusters users based on semantic analysis of movement patterns and categorizes when user-labelled data is available of the region. As mentioned earlier,

the main challenge is to get labelled GPS trace of mobile users for categorization. We present a method to transfer knowledge from source, where labelled data available to target region where labelled user trace is insufficient. By measuring the place taxonomy and analysing user trace feature, we can transfer human movement behavioral knowledge to another spatial region.

4. GENERATING USER-TRACE SUMMARY

Although human movement pattern is unpredictable, there is always some intent behind how user moves. Bayesian network model is deployed to capture the intents of human movement patterns and measure the similarity with other movement patterns.

4.1 Modeling of User-Trace Summary

The behavioral patterns of the human movement traces (collected through GPS footprints) can be effectively modelled using Bayesian network. Given a set of user trajectory segments of individuals or a set of users, how to construct Bayesian network to capture the movement pattern from individual as well as aggregate level? Consider a finite set of random variables, $S = \{S_1, S_2, ..., S_n\}$, where each variable denotes a stay-point in the user trace. Each variable has two values, 'visited' (denoted as 1) and 'not-visited' (denoted as 0). Let P be the joint probability distribution over the variables in S. Here, we aim to build UTS in a form of a Bayesian network, which encodes joint probability distribution over a set of stay-points in the region based on the movements user makes. Given a set of source and destination points, paths followed by an user or a group of users are not pre-defined and it is dependent on the users' personal choices. For example, 80% user-trajectory segments (only stay points are depicted) of user1 is - $[< Hall, 8 >, < Academics, 9 >, < Hall, 12 >]$, and 20% user-trajectory segments of user1 is $[< Hall, 8 >, < cafe, 9 >, < Academics, 10 >, < Hall, 12 >]$. Now, to build the user trace summary, we need to reflect both the trajectory segments along with the ratio (4:1) of following above two segments. Hence, the probability to follow different segments can not be completely captured in graph based approaches [6]. To tackle this challenge, we generate a probabilistic graphical model (UTS) which is capable to represent all movement patterns of an user without any loss of information.

We define UTS, $N_B =< G, \theta >$, where $G =< V, E >$, $v_i \in V$ consists of (N_c, t_i); where N_c is the code-place of the $SSP_{Taxonomy}$ and t_i is the temporal value at the node. $e_i \in E$ represents dependencies between the vertices, i.e stay-points of the trace. We assume that each variable in S is independent of its non-descendants in G given its parents. This independence assumption holds because two stay-points are independent when there is no direct movement from one to another in the given user trajectory segment. θ contains the set of parameters

$$\theta_{s_i|Pa_{s_i}} = P(s_i|Pa_{s_i}) \quad (1)$$

, which quantifies the network, for each possible values of s_i (in our case '0' or '1') in S_i and Pa_{s_i} of Pa_{S_i} (the parent set of S_i in G). The joint probability distribution is given by

$$P_N(S_1, S_2, ..., S_n) = \prod_{i=1}^{n} P_N(S_i|Pa_{S_i}) = \prod_{i=1}^{n} \theta_{S_i}|Pa_{S_i} \quad (2)$$

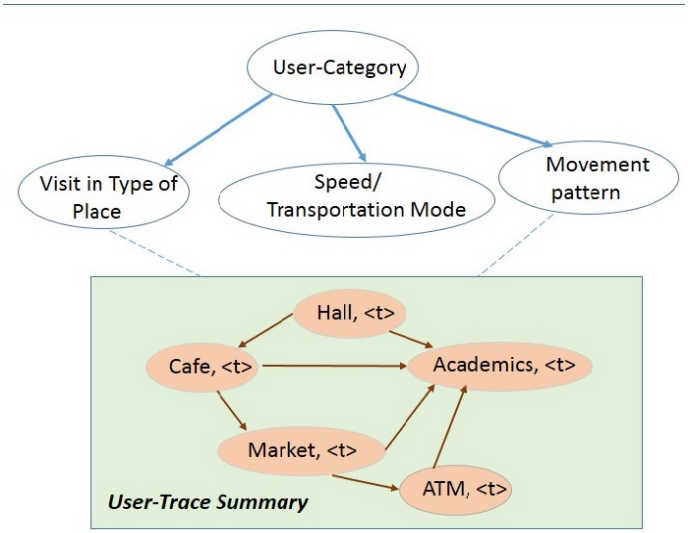


Figure 4: Bayesian Network for User-Classification

Each variable (S_i, stay-points in our problem) in the network has a associated conditional probability distribution. The conditional probability distribution of S_i, given its parent Pa_{S_i} is denoted by $P(S_i|Pa_{S_i})$. It is calculated from the GPS-log of the users by analysing the frequency of visiting S_i after visiting its parent Pa_{S_i} and visiting S_i from other stay-points which are not included its parent set. To cluster or grouping trajectory segments, location information, time duration in a stay-point, speed of movement, movement patterns of the trajectories need to be considered. We have used *Temporal Common Sub-sequence, (TempCS)* clustering algorithm [6], which captures common subsequences among the trajectories. In our problem, given a User-Trace Summary $N_{B_1} =< G_1, \theta_1 >$; and a set of N_B, similarity matrix among the networks is required. To measure the similarity between two probability distribution, we have used the Bhattacharyya distance [1] given as:

$$D_B(S_i, S_j) = -ln \sum_{x \in [0,1]} \sqrt{S_i(x) S_j(x)} \quad (3)$$

Using $TempCS$ on complete UTSs (say, N_{B_1} and N_{B_2}), we generate a list (L_c) of all common staypoints and the common subsequences (say L_s). Using D_B in equation 3, it may be observed the similarity between two probability distribution of each two consequent staypoints in L_s. Hence, we come up with the similarity measures between N_{B_1} and N_{B_2} as:

$$Sim_{Sequence}(N_{B_1}, N_{B_2}) = \frac{|L_s|}{|L_c|} \sum_{S_i \in L_s} D_B(S_i, S_{i+1}) \quad (4)$$

and

$$Sim_{Temporal}(N_{B_1}, N_{B_2}) = \sum_{S_i \in L_c} Min(T^1_{S_i}, T^2_{S_j}) \quad (5)$$

where $T^j_{S_i}$ denotes time-duration spent in stay-point S_i of N_{B_j}.

4.2 User Categorization

We aim to classify users into pre-defined categories based on the features of the GPS traces. Bayes classifier is used as our

goal is to get the probability of multiple user-categories of the target user. The classification task for a user u provides a output probability vector, $PV_u = \{p_1, p_2, ..., p_i\}$ where i is the user-category and p_i is the probability of the user u to be categorized in i^{th} class. The user-classification is done based on three observable features, i) *visit in types of places* (f_1), ii) *Speed of movement or transportation mode* (f_2), iii) *User Movement patterns* (f_3). f_2 can take values 'Bi-Cycle' or 'Four-wheeler' or any other transportation mode (when transportation mode is available) or speed can be computed and value can be discretized into 'high', 'medium' and 'low' value. f_1 and f_3 can be deduced from User-Trace Summary generated from the GPS log and the corresponding similarity measures given in equation 4 and 5. For each user-trace, we find out the weighted feature vector and create a Bayesian network [as depicted in Figure 4] to classify the user to a pre-defined user-category.

Considering our problem, say a new instance (GPS traces of user u) u having three feature values f_1, f_2, f_3; C be the target feature which takes value c. u is classified to the class with maximum posterior probability as defined:

$$U_{category} = argmax_c P(c)P(f_1, f_2, f_3|c) \tag{6}$$

As all features are independent of each other,

$$P(f_1, f_2, f_3|c) = \prod_{i=1}^{3}(f_i|c) \tag{7}$$

In our problem-scenario, it is obvious that user classification is more dependent on user movement patterns or stay-points rather than transportation mode, i.e. the assumption that the features having equal importance in classification does not hold. The Bayesian classification with feature weighting [7] is defined:

$$U_{w-category}(u, w(i)) = argmax_c P(c) \prod_{i=1}^{3}(f_i|c)^{w(i)} \tag{8}$$

where $w(i) \in \mathbb{R}^+$, i.e each feature i has its own weight $w(i)$. We have used *Kullback-Leibler* measure [7] for weighting each of the features in our Bayesian learning. In this section, we describe how user categorization can be done from labelled GPS traces using the GPS trace features.

5. MAPPING KNOWLEDGE FROM ONE REGION TO ANOTHER OF SIMILAR TYPE USING TRANSDUCTIVE TRANSFER LEARNING

Our objective is to extract knowledge from the GPS traces of source region (in our experiment IIT Kharagpur, India) and transfer the knowledge, more specifically labelled data of Kharagpur region to target region (Dartmouth College, Hanover city, USA) for user-categorization at the target region. Both the regions (source and target) are academic institutes/universities.

According to [14], *Transductive Transfer Learning* is defined on a source domain D_S and a corresponding learning task T_S, a target domain D_T and a corresponding learning task T_T, where it aims to improve the learning of the target predictive function $f_T(.)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$ and $T_S = T_T$ It is mapped in our problem as analysing and learning categorical movement patterns from labelled GPS traces or ground truth data of source region, and carry out the user categorization task in the target region, where we have only unlabelled GPS traces of users. The formulation of our problem is stated as Given a semantic source region of interest S_{ROI} and user-classification task T_S; different but related semantic target region of interest T_{ROI}, how to learn the classifier from S_{ROI} and T_S to perform the classification or learning task in T_T. We define semantic source region of interest as $S_{ROI} = (SSP^S_{Taxnomy}, P(w_S))$ and semantic target region of interest as $T_{ROI} = (SSP^T_{Taxnomy}, P(w_T))$. The former term in the tuple represents the taxonomy of stay-points of the particular region and the later term presents the probability distribution of visiting frequencies in the nodes of the taxonomy. Given two different region of interests, (here, Kharagpur, India and Dartmouth, Hanover, USA), the problem of transferring the knowledge of movement patterns becomes more challenging due to different life-styles of people. For example, we have observed, wide variation in probability distribution of footprints in the staypoints namely, Gym, Stadium, Restaurants. Further, there are differences in temporal pattern of visiting in the two regions of our experiment. Basically, we aim to find proper feature-representation that minimizes the divergence of two regions and the classification error. We aim to reduce the classification error in the target region of interest:

$$error(h) = \frac{1}{|U_T|} \times \sum_{<u,c> \in |U_T|} Pr(h(u) \neq c) \tag{9}$$

Given the training data set (or labelled data) of source region and data (unlabelled) distribution of target region, we aim to estimate the correct label or category of users in the target region. To tackle the above issue, as depicted in Figure 1, we carry out the trajectory summarization module on the target region data and generate the $SSP^T_{Taxnomy}$ of the region and UTS for each user. We extract geo-tagged information of the GPS points (latitude, longitude) from Google Map by reverse geocoding technique. For example, for a particular POI, say Department of Mathematics, geo-tagged information is in the form: [point of interest, establishment, academic, university, department] or another POI, say Gile Hall has geo-tagged information array: [point of interest, establishment, lodging, Student Housing Complex, Hall of Residence, Undergraduate Students' Hall]. Clearly, it stores the places in a hierarchical form, we convert the array in a taxonomy where each of the nodes is given an unique code to store the semantic meaning of the places. In our proposed coding scheme, we extract the parent's code (say, c_p) of a node (say c), and check whether the particular node has siblings. Let a node (c) has n siblings then we append $n+1$ along with the parent's code. Hence, the node gets a code $c_p n + 1$. Before assigning new code, we check whether the same place-type appears in the $SSP^S_{Taxnomy}$ and assign the same code if it is present. From the coding scheme, it is obvious that all the nodes in level l of the taxonomy have unique codes of length l. Also, the common sequence of two such codes represents common hierarchy of two nodes, i.e, similar semantic types of places. While inserting each node (n), we recursively add the visit frequency of each node from n to the root following the code of the node.

After the generation of $SSP^T_{Taxnomy}$ (destination taxonomy),

we compare it with the $SSP^S_{Taxnomy}$ (source taxonomy). Because, transductive transfer learning method does not work when the target and source region of interests and distribution of GPS traces are completely different. For example, we can't simply compare an academic ROI and a commercial ROI as clearly the movement patterns are completely different which will lead to poor classification result. Hence, we need to analysis where it is feasible to transfer knowledge between two domains. Otherwise it may lead to 'negative transfer' [14]. Following the coding scheme of the taxonomy, we find out the maximum common sequence and compare the GPS footprint distribution at each of the level of the taxonomy.

We aim to find a good feature representation that will minimize the differences between two regions. By comparing two such structures, we get the common taxonomy (say $SSP^C_{Taxonomy}$) features of source and target region of interest. We generate the UTS of users in target region and training data of source region using the $SSP^C_{Taxonomy}$. The Bayes classifier, as explained in section 4, computes the probability vector of a user u being categorized in all of the pre-defined classes. It depends on the distribution of the training data. Here, along with the training data of source region, we use test data distribution of target region for user-classification. We have implemented Transductive Bayes classifier [2] algorithm, where at each iteration weight of the classifier based on the labelled test data is increased.

In this section, we propose how spatial trajectory knowledge can be refined and transferred by capturing the common place taxonomy of two regions and probability distribution of GPS trace data.

5.1 Motivating Example Scenario: Extending to other Regions of Interest

One of the promising application of the problem is to build up various $SSP_{Taxonomy}$ from GPS traces collected from different types of users and from varied locations. If we can build several such stay-point taxonomy (University Campus/Commercial places/ Residential areas) and construct a knowledge base of visiting patterns, we can train Bayes classifier using more categorical movement patterns. Further, we can extend our framework for better urban planning.

6. EXPERIMENTAL EVALUATIONS

In this section, we present brief description of the dataset used and some results representing the performance of the framework.

6.1 Dataset

The GPS dataset is collected voluntarily from 56 university students/professors from their GPS-enabled mobile devices and Google Map's timeline. The dataset is mainly generated in Kharagpur, India region and captures daily movement patterns of the subjects from October 2015 to November 2016. Also, the dataset is manually labelled (ex. category of users - student/professors, daily movement patterns described by them etc.) by the subjects. We also carry out our experiment with Dartmouth data [11],[18],[12] consisting GPS traces of students and professors, employees of the university. Table 1 shows the final user-distribution of the datasets.

User Category	Kharagpur Region	Dartmouth Region
Undergraduate Student	8	2
Graduate Student	11	15
Graduate Student (Research Student)	22	NA
Employees/Professors	8	3
Non-Residential Students	7	NA

Table 1: GPS Data-set Description of Kharagpur and Dartmouth region

6.2 Accuracy of User-Classification

We generate a confusion matrix and find out the recall and precision for each of the class. Precision is the measure how accurate the classifier is and recall is how well it can classify a positive class label; $Precision = \frac{TruePositive}{TruePositive+FalsePositive}$ and $Recall = \frac{TruePositive}{TruePositive+FalseNegative}$ Table 2, represents the evaluation measure of the classifier. Here, in Kharagpur region, we divide the dataset into training (70%) and test dataset and learn the Bayesian classifier for user categorization. In Dartmouth region, we did not use any labelled data of the region to train the classifier, rather we used the knowledge gained from Kharagpur region to classify users in Dartmouth region.

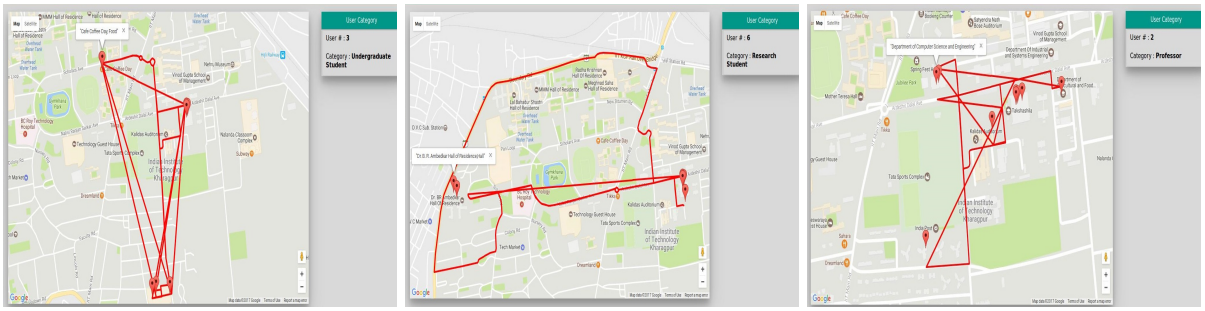
	Kharagpur		Dartmouth	
User Category	Precision	Recall	Precision	Recall
Undergraduate Student	0.75	0.75	0.78	0.82
Graduate Student	0.846	0.95	0.9	0.82
Graduate Student (Research Student)	0.6315	0.545	NA	NA
Employees/Professors	0.555	0.625	0.62	0.56
Non-Residential Students	0.714	0.714	NA	NA

Table 2: Precision/Recall values of experiment on Kharagpur and Dartmouth region

We have developed a demonstration system to visualize the results. Figure 5 depicts summarized movement patterns of student and professor category of users in Kharagpur region.

7. CONCLUSION AND FUTURE WORK

The paper addresses the user categorization problem from the GPS traces of the users. User categorization or similarity measurement of users based on movement patterns provides insights of common needs or interests of people and may help in various application like car-pooling, business settlements, city planning etc. We propose a framework to model individuals' movement patterns, analyzing human movement patterns both from semantic and spatio-temporal context and extracting implicit information. Finally, we attempt to present an insight on the transferring knowledge base from one city domain to another unknown city where data is insufficient.



(a) Summarized Undergraduate Student User-Trajectory Segment (b) Summarized Research Student User-Trajectory Segment (c) Summarized Professor User-Trajectory Segment

Figure 5: Summarized User Trajectory Segments of different categories of users in Kharagpur Region

In the future, we would like to extend features of our system to analyze indoor movements and scale up the system to incorporate different methods for activity learning or behavioral pattern mining. Further, we can extend the learning task for different domains and build up a city hierarchy using the transfer learning methods. We would like to improve experimental section of our work using more number and varied GPS traces. Moreover, we would like to explore the parallel computation and fast data access mechanisms which is highly required for the towering growth rate of trajectory traces.

8. REFERENCES

[1] A. Bhattachayya. On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, 35:99–109, 1943.

[2] K. Branson. A naïve bayes classifier using transductive inference for text classification. Technical report, Technical Report, Dept. of Computer Science and Engineering, UCSD, 2001.

[3] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Transferring naive bayes classifiers for text classification. In *Proceedings of the national conference on artificial intelligence*, volume 22, page 540. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[4] B. Du, C. Liu, W. Zhou, Z. Hou, and H. Xiong. Catch me if you can: Detecting pickpocket suspects from large-scale transit records. 2016.

[5] Z. Fang and Z. M. Zhang. Discriminative feature selection for multi-view cross-domain learning. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1321–1330. ACM, 2013.

[6] S. Ghosh and S. K. Ghosh. Thump: Semantic analysis on trajectory traces to explore human movement pattern. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 35–36. International World Wide Web Conferences Steering Committee, 2016.

[7] C.-H. Lee, F. Gutierrez, and D. Dou. Calculating feature weights in naive bayes with kullback-leibler measure. In *2011 IEEE 11th International Conference on Data Mining*, pages 1146–1151. IEEE, 2011.

[8] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 34. ACM, 2008.

[9] M. Lv, L. Chen, and G. Chen. Discovering personally semantic places from gps trajectories. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1552–1556. ACM, 2012.

[10] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646. ACM, 2009.

[11] M. Musolesi, K. Fodor, M. Piraccini, A. Corradi, and A. Campbell. CRAWDAD dataset dartmouth/cenceme (v. 2008-08-13). Downloaded from `http://crawdad.org/dartmouth/cenceme/20080813`, Aug. 2008.

[12] M. Musolesi, M. Piraccini, K. Fodor, A. Corradi, and A. T. Campbell. Supporting energy-efficient uploading strategies for continuous sensing applications on mobile phones. In *International Conference on Pervasive Computing*, pages 355–372. Springer, 2010.

[13] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua. Beyond doctors: Future health prediction from multimedia and multimodal observations. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 591–600. ACM, 2015.

[14] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[15] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, et al. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4):42, 2013.

[16] C. Renso, M. Baglioni, J. A. F. de Macedo, R. Trasarti, and M. Wachowicz. How you move reveals who you are: understanding human behavior by analyzing trajectory data. *Knowledge and information systems*, 37(2):331–362, 2013.

[17] X. Song, L. Nie, L. Zhang, M. Liu, and T.-S. Chua. Interest inference via structure-constrained

multi-source multi-task learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2371–2377, 2015.

[18] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.

[19] Y. Wei, Y. Zheng, and Q. Yang. Transfer knowledge between cities. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1905–1914. ACM, 2016.

[20] Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yu. Heterogeneous transfer learning for image clustering via the social web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 1–9. Association for Computational Linguistics, 2009.

[21] G. Yavaş, D. Katsaros, Ö. Ulusoy, and Y. Manolopoulos. A data mining approach for location prediction in mobile environments. *Data & Knowledge Engineering*, 54(2):121–146, 2005.

[22] D. Zhang, J. He, Y. Liu, L. Si, and R. Lawrence. Multi-view transfer learning with a large margin approach. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1208–1216. ACM, 2011.

[23] J. Zhang and J. Huan. Inductive multi-task learning with multiple view data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 543–551. ACM, 2012.

[24] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma. Understanding transportation modes based on gps data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1):1, 2010.

[25] Y. Zheng, X. Xie, and W.-Y. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[26] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining correlation between locations using human location history. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 472–475. ACM, 2009.

[27] 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. ACM, 2009.