IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 29, 2018, accepted October 24, 2018, date of publication November 9, 2018,
date of current version December 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2879634

Efficient Similarity Search for Travel Behavior

LEI TANG, (Member, IEEE), YALING ZHAO, ZONGTAO DUAN, AND JUN CHEN
School of Information Engineering, Chang'an University, Xi'an 710064, China

Corresponding author: Lei Tang (tanglei24@chd.edu.cn)

This work was supported in part by the Project of the National Natural Science Fund of China under Grant 61303041, in part by the Key Science and Technology Innovation Team of Shaanxi Province under Grant 2017KCT-29, in part by the Key Research and Development Plan Project of the Shaanxi Province under Grant 2017GY-072, in part by the International Scientific and Technological Cooperation Project of the Shaanxi Province under Grant 2017KW-015.

ABSTRACT To provide travel recommendations and planning in the intelligent transportation system (ITS), we must have the ability to find similar travel patterns among users based on their real mobility traces. To measure the similarity of user's travel behavior, various methods have been proposed, but they usually only rely on a single attributes-related metric. In comparison, studies of the semantic relationships between travel attributes remain scarce, making it difficult to construct a complete mobility pattern that reveals the relevance between users or groups. In this paper, we introduced the heterogeneous information network to build a weighted travel network with spatial–temporal GPS trajectories. The heterogeneous network allows clustering the similar users based on the connections between different attributes instead of attribute values. On this basis, we defined the meta-paths for travel and used each meta-path to formulate a similarity measure over users by improving existing PathSim (Meta-path-based similarity measures) and SimRank. Next, we aggregated different similarities, where each meta-path was automatically weighted by the learning algorithm to make predictions. The experimental results showed that the recall of the similarity measurement algorithm using multiple meta-paths has improved, which yielded better results than the performance of the algorithm using a single meta-path. The performance of the improved PathSim model under different scales of data was 15% higher than the performance of the improved SimRank model in terms of precision and 21% higher in terms of recall. Due to the area under curve values, our experiments also show that a meta-path combination is more effective than the state-of-the-art approaches and can be efficiently computed.

INDEX TERMS Travel, similarity, heterogeneous information network, meta-path.

I. INTRODUCTION

To improve both traffic and travel, it is necessary for ITS to address the challenges presented by the increasing number of motor vehicles, while at the same time satisfying increasing demands for more and better traveling [1]. ITS has proven to be valuable tools for sorting through a large number of available transportation services while traveling. To promote the use of ITS for making individual trips easier and more enjoyable, much attention has been paid to understanding and analyzing similarities in travel patterns among users [2]. In the case of two individuals with similar travel patterns, we might conclude quickly that recommending nearby services along similar routes would be straight-forward as well as highly desirable [3]. For example, city planners can closely monitor such patterns and compare mobile usage, identify regularities across regions and user groups [4]. A local, daily ridesharing within the city can also be promoted by recognizing matching rides along similar patterns [5].

Generally, individuals on a trip will generate a considerable quantity of spatial data with time markers that can describe his or her travel behavior, such as taxi data [6], mobile phone data [7], and social media data [8]. To exploit travel attributes, e.g., to improve travel time, pick-up location, and transport mode, travel services generally use GPS trace data and check-in data from social media. By analyzing the check-in records for a specified location, the hot spots of an area and their popularity can be determined. In addition, users can give consent for services to obtain their personal preferences and frequent access patterns by mining their check-in histories.

However, existing methods focus on mining and fusing only typical attributes from GPS or social media data [9], [10]. For this reason, it is difficult to characterize complex travel attributes and the associations between them, such as when a user buys fuel, uses a car maintenance service, or arrives at a local university by car or public transportation. Also, to measure the relevance of two

users, neighborhood-based measures such as common neighbors and Jaccard's coefficient were proposed. Other graph-theoretic measures that are based on random walks between objects include Personalized PageRank and SimRank. These measures do not consider distinct relationships connecting users with their attributes, that form a heterogeneous information network(HIN) [11]. The rich semantics encoded in different types of relations can then bring more information. We have illustrated an example of similarity measurement in carpooling. To optimize the use of private cars, each car must carry more people. Usually, existing approaches can match drivers and riders by assessing the similarity of their origin-destination (O-D) [12], or by determining demographic information (e.g., gender and estimated income) and individual interests [13]. These approaches have not yet taken into account the relationships between different objects in travel, such as a user and his travel time, that would enable extraction of significant semantic knowledge for ride-matching. In contrast, by constructing a HIN, users are connected with each other through social links, and they are also connected with a set of locations, timestamps, and text contents through online activities. We can thus mine various relationships, and extract the user's attributes for measuring the similarity between drivers and riders. For example, by understanding the relationship between users in terms of location (e.g., staying at the same places), an application could find user preferences for various spots, and determine the departure time and pick-up locations. Also, the ability to identify a passenger's friends, or friends of friends, by mining a relation between users via activities (e.g., shopping after work) would be helpful when grouping similar passengers for sharing a trip.

In this study, we presented a heterogeneous travel network based on a Geolife dataset that contained 182 travers and covered a period of five years. Then we defined three types of travel meta-paths to illustrate the special relations between users. Using this approach, the similarity measurement became a matter of classifying users with the different meta-paths quantified. Our contributions are as follows:

To the best of our knowledge, this is the first work to use structured HIN for measuring travel behavior. We described the semantic relationships between users by constructing a heterogeneous travel network and mining multiple meta-paths for travel. This approach is different from prior work that searched for similar users by grouping users according to attributes recognized, instead of the relations among users.

Considering that different top-k measures would be derived from different relations, we are able to combine such measures automatically. To estimate how a single meta-path influences the effectiveness of the combination, we employed a machine learning technique that allowed for qualifying the weight of each meta-path.

For labeling the similarity among users in the dataset, we trained a model to estimate the variance of occurrences between positive and negative samples, instead of manually labeling a small subset of data. We fitted a polynomial of degree 3 by 10-fold cross validation using all data and determined a relevance score at two levels: 0-non relevant and 1-relevant.

The remaining sections of this paper are presented as follows. In Section II, we present an analysis of existing research related to the similarity of travel behavior. In Section III, we propose a heterogeneous travel network and describe how the meta-path is constructed. Next, Section IV puts forward our similarity calculation method based on the meta-path. Section V describes the evaluation of the experiment and its results, and Section VI concludes the paper.

II. RELATED WORK

A. TRAVEL BEHAVIOR ANALYSIS

Similarity measurement for travel behavior focuses mainly on travel behavior analysis, mobility pattern recognition, and pattern-based similarity calculations. Travel behavior characterizes user's attributes in different scenarios (e.g., commuting, family travel). Kang *et al.* [14] explored the purposes of household travels. By recognizing travel patterns, including travel modes and times, they constructed a model for selecting a destination based on routes and schedules. They could then apply this model to determine potential travel locations and forecast daily household travel plans in spatiotemporal terms. Pan *et al.* [15] introduced the bounded-rationality individual decision-making model in the selection of travels. They also compared and improved the relative utility model and random regret model. Based on these models, with contexts taken into account, the bounded rationality in determining travel attributes was revealed, and the model's superior performance in obtaining the bounded rationality in individuals' travel selection behavior was verified.

1) TRAJECTORY PATTERN MINING

Other researchers considered the travel attributes of individuals or groups, which was helpful for constructing a trajectory model that could describe the pattern of past travel [16], [17] and plan future travel [18], [19]. When users' daily trajectory data are collected by a GPS receiving device or social media, the amount of spatial-temporal data is massive, and activities, such as staying at a special location, can be captured. In a study by Lian *et al.*, [20] a point of interest (POI)-based matrix was used to make travel recommendations through the decomposition of the weighted matrix. In the decomposition model, the vectors of the activity area and POI-influenced area were used to improve the connections between users and their points of interest. Moreover, spatial clustering was accomplished by using the two-dimensional kernel, which solved the sparsity problem of the POI matrix and improved recommendation performance. However, a classification was rough in the decomposition model, which affected the accuracy of the inferred information categories. He *et al.* [21] inferred frequent routes by mining GPS trajectories and using a frequency-related quality of service (QoS)-constrained method for mining the frequent paths. Next, they

employed a route mining strategy to optimize the candidate similar paths. Based on this work, users were provided with carpooling recommendations. Furthermore, spatial-temporal pattern mining was proposed to incorporate both regions of interest and travel time between movements [22].

2) TRAVEL TIME ESTIMATION

Travel time is an important factor that influences travel behavior. Some methods first estimate the travel time of individual road segments and then sum up the travel times of the road segments belonging to a path [23]. However, it is difficult to explicitly model the complex factors for crossing two road segments, e.g., intersections, and traffic lights [24]. Other researches try to find more optimal concatenations of road segments to estimate the travel time of a path, for example using a joint probability model [25], or a dynamic Bayesian network [26]. However, these methods do not consider the interactions between the length of sub-path and the number of trajectories passing in travel time estimation.

3) USER PREFERENCES MODELING

Shang *et al*. [27] observed that the spatial similarity on trajectories itself is not sufficient to capture the relationship between users due to the more specific preferences of users. Most existing approaches directly predict a user's preference on a location [28]. Personalized Trajectory Matching (PTM) took into account the significance of each sample point in a trajectory. Different weights have been assigned to sample points based on the user's preferences [29]. Zhou *et al.* [30] split the check-in records to construct a spatial-temporal trajectory dataset. Then they formally incorporated multiple context information of trajectory data into the proposed model, including user-level, trajectory-level, location-level, and temporal contexts. User-level contexts characterize user preference such as shopaholic or music fan in an embedding vector. By learning the representation of user preference, Zhou *et al.* implemented the location recommendation and social link prediction. Liu *et al.* [31] provided insights into users' preference transitions over location categories. By splitting a user's check-in history into several non-overlapping sequences, the probability denoting a user would follow a given preference transition can be predicted to recommend a set of POIs.

B. SIMILARITY MEASUREMENT

The calculation of similarity depends on the travel attributes selected when seeking patterns, and the results determine the performance of the patterns. Trasarti *et al*. [32] mined users' public mobility profiles and adopted a mobility modeling algorithm to measure the similarity of user travel behaviors, using a path similarity function. This method reduced the spatial-temporal complexity of the data, and it was robust. However, only a small number of data points were used and validated, so non-deterministic situations in the real world were not well considered and simulated. Elbery *et al*. [17] proposed a recommendation system that used the individual's check-in history and homepage locations to act as a user. In addition, a fast Fourier transform (FFT) was employed to represent the user's check-in data, and to measure the similarity between users. In this system, the weighted hierarchical clustering method was adopted to estimate user locations, and to further recommend carpool services. Although the preprocessing of data was not required when using FFT, only ten coefficients of FFT were determined by the system, so the similarity needed to be improved.

In existing models, usually, symbol-based schemes have been used for estimating similarity using GPS data or social media data. It is difficult to characterize the associations between users by treating travel attributes as the symbols, also we cannot mine the similarity among users according to the potential relationships connecting users with their attributes. The study of HINs started in 2010 for analyzing the implied connections between different objects. HINs have been used widely in data mining for similarity measurement, clustering, classification, link prediction, and recommendation. Sun *et al*. [33] defined the symmetric meta-path to find the different linkage paths among the same type of objects in a network. PathSim was then proposed to measure the meta-path-based similarity. PathSim was useful for both homogeneous and heterogeneous networks, and it also described the semantic meanings behind paths. However, the performance of different meta-paths should be tested to provide accurate similarity measures in real systems. Heterogeneous collective link prediction (HCLP) [34] allows for predicting the relationships among multiple types of links using the meta-path in Bioinformatics. Yet, choosing appropriate meta-paths automatically should be improved and validated in different application scenarios. SemRec [35] improves the meta-paths for describing the uncertain weight of each link in the recommender system. It helps depict the path semantics to predict the rating scores of users on items.

Because of the comprehensive information integration and rich semantic information provided by a HIN, it promises to generate better similarity measurement [35]. However, the attribute values on links, and the wide use of only one meta-path in a HIN may fail to capture accurately the semantic relations among objects [36]. For example, a ''user-user'' meta-path denotes the links between users, and a ''user-location-user'' meta-path implies that the same spot was visited by two users. Each meta-path may suggest similar users in terms of different travel attributes that lead to a difference of similarity measurement. Thus, in practice, it is necessary to make a meta-path combination and define the effect of each path to optimize the measurement.

In our study, as distinct from the work by Sun and Han [37], we focused on the relationships among travel attributes. We adopted the HIN theory and defined three types of semantic travel relationships based on the HIN nodes. Also, we carried out a supervised learning process to constitute a multiple meta-path, instead of using the linear combination of several meta-paths with the predefined parameters, such as

Path Constrained Random Walk (PCRW), for analyzing the similar travel behaviors.

III. MODELING TRAVEL BEHAVIOR BASED ON MULTIPLE ATTRIBUTES

A. ANALYSIS OF TRAVEL BEHAVIOR ON TRAJECTORY DATA

Users' GPS trace data usually contain both geographical location and temporal information. Consequently, temporal attributes, POI (points of interest), and other types of information such as transport mode are required for the description of travel behavior. Through the flexible use of these detailed travel attributes, an individual's travel can be described accurately, and various travel mobility patterns can be recognized and classified.

For this study, first, we preprocessed the GPS points, which included recognizing the stay points using k-means clustering, and modeling via space-time series. To determine users' preferences exactly, we needed to know their service-visiting behaviors. Therefore, each point was marked to indicate which service was accessed. This procedure allowed us to mine the positive correlation between service-visiting behaviors and GPS trajectories. According to our previous work [38], the users' generated travel trajectories (travel points) were labeled with the user's identification (ID), the locations visited during a certain time of day, and the available services corresponding to the locations.

The dataset for the quantitative analyses was provided from Geolife and comprised the POI data covering 60% of the area in Beijing, including 385,734 data records. We selected only the 21 most representative types of services, including educational training, shopping, and cultural media. Stay points included more than 4000 locations in Beijing, such as the Old Summer Palace, Tsinghua University, National Stadium, and the "Branch bank" building, as shown in Figure 1. In terms of travel time, a day (24 hours) was divided into four periods based on consistent intervals: 00:00:00–07:00:00, 07:00:00–12:00:00, 12:00:00–19:00:00, and 19:00:00–24:00:00.

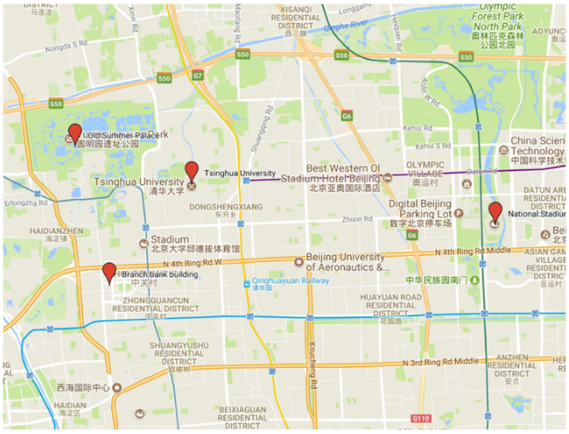


FIGURE 1. Illustration of some of the stay points in Beijing, where there are four stay points marked in red.

B. PROBLEM DEFINITION

Considering the diversity and complexity of travel attributes, we adopted the theory of HINs to construct a travel network based on the trajectories by serializing travel points. In such a network, nodes and relations are of different types. This work utilizes HIN as defined in Definition 1, which is similar to the definition by Sun *et al.* [33]

Definition 1: HIN A HIN is defined as a triad $G =< V, E, A >$, with an object type mapping function $\varnothing : V \rightarrow A$ and a link type mapping function $\psi : E \rightarrow R$, where each $v \in V$ denotes a set of nodes with different properties, including different objects. Each link $e \in E$ describes the multiple semantic relationships between different objects. A denotes the set of node types, representing the types to which the objects belong.

When the types of objects $|A| > 1$ or the types of relations $|R| > 1$, the network is a heterogeneous information network; otherwise, it is a homogeneous information network.

As we know, conventional HIN do not consider the attribute values on links. However, this travel network can contain attribute values on links. Concretely, the users can visit a place several times during a day. The times that a user appeared at a certain place can be used as an indicator on the link between user and location which user has visited in the past. For this study, we allowed for the existence of complex relationships between multiple types of nodes, and the HIN was expanded. Users' stay points, the services users accessed at stay points, and travel times were chosen as the node types of the model.

The model's edge $E = \{E_{ul} \cup E_{ls} \cup E_{lt}\}$ was constructed between node types. E_{ul} connects users and the stay points, indicating that users stayed in the identified place; E_{ls} connects stay points and services, indicating the categories of services utilized by users at stay points; and E_{lt} denotes the time periods during which users remained at stay points. In addition, we allowed weighting in the heterogeneous travel network to quantify the degree of association between different node types. The conceptual diagram of the proposed model is shown in Figure 2.

Definition 2 Heterogeneous travel network. A heterogeneous travel network is defined by a six-tuple $TN =< U, L, S, T, E, W >$.

(1) $U = \{u_1, u_2, \ldots, u_n\}$ denotes the set of types of travel users, and u_i denotes user nodes;

(2) $L = \{l_1, l_2, \ldots, l_m\}$ denotes the set of types of stay points;

(3) $S = \{s_1, s_2, \ldots, s_{21}\}$ denotes the set of types of services accessed at stay points;

(4) $T = \{t_1, t_2, t_3, t_4\}$ denotes the set of travel time series;

(5) $E = \{E_{ul} \cup E_{ls} \cup E_{lt}\}$ denotes the set of all edges in the network. $E_{ul} = \{e(u, l) | u \in U, l \in L\}$ describes the semantic relationship between users and locations, i.e., users stay at places. $E_{ls} = \{e(l, s) | l \in L, s \in S\}$ contains the semantic relationship between locations and service types, i.e., users

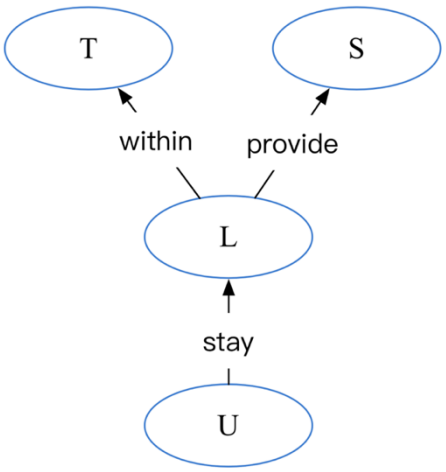


FIGURE 2. Example of HIN Schema. U: users; L: locations; S: category of service; T: time period. "Stay" means that a user stays a location, "provide" means that a user chooses the service categories at a certain location and "within" represents that a user travels within a time period.

stay at places and use provided services. $E_{lt} = \{e(l,t)|l \in L, t \in T\}$ describes the semantic relationship between time periods and locations, i.e., users stay within time periods.

(6) $W = \{W_{ul} \cup W_{ls} \cup W_{lt}\}$ denotes the weighted set of edges in the model. $W_{ul} = \omega(e(u,l))$ denotes the times that users appeared at a certain stay point; $W_{ls} = \omega(e(l,s))$ denotes the number of categories of service provided at stay points; $W_{lt} = \omega(e(l,t))$ denotes the number of stay points that were visited within a time period.

The applicability of a heterogeneous travel network can be described by the following example. Mike lives in Beijing, and his travels in a day are as follows. He eats breakfast in SAGA Mall at 07:40, arrives at work at Tsinghua University at 09:00, and watches a movie in SAGA Mall at 19:30. By using a heterogeneous travel network, the following nodes and sets of edges can be constructed:

$$U = \{Mike\},\ L = \{'\text{SAGA Mall}','\text{ Tsinghua University}'\},$$
$$S = \{'\text{meal}','\text{entertainment}','\text{ work}'\},$$
$$T = \left\{ \begin{array}{l} '(00:00:00, 07:00:00)', \\ '(07:00:00, 12:00:00)', \\ '(12:00:00, 19:00:00)', \\ '(19:00:00, 24:00:00)' \end{array} \right\}$$
$$E_{ul} = \left\{ \begin{array}{l} e(Mike, '\text{SAGA Mall}'), \\ e(Mike, '\text{Tsinghua University}') \end{array} \right\}.$$

$W_{ul} = \{2, 1\}$ indicates that Mike stayed twice at SAGA Mall and once at Tsinghua University, respectively.

$$E_{ls} = \left\{ \begin{array}{l} e('SAGA\ Mall', 'meal'), \\ e('Tsinghua\ University', 'work'), \\ e('\text{SAGA Mall}', 'shopping') \end{array} \right\}.$$

$W_{ls} = \{1, 1, 1\}$ indicates that there is one type of services Mike performed each time when he stayed at a certain place.

$$E_{lt} = \left\{ \begin{array}{l} e('\text{SAGA Mall}', '(07:00:00, 12:00:00)'), \\ e('\text{Tsinghua University}', '(07:00:00, 12:00:00)'), \\ e('\text{SAGA Mall}', '(19:00:00, 24:00:00)') \end{array} \right\}$$

$W_{lt} = \{1, 1, 1\}$ indicates that Mike stayed twice at t_2, and once at t_4. Furthermore, by connecting users with different attributes, the network was extended to mine the symmetric travel behavior and enable users to be fitted into the same class as those having similar travel attributes. Figure 3 provides an illustrated example. In the network, users are connected with a set of locations, timestamps, and text contents through online service-visiting activities. For instance, TOM and William stayed at the same place (i.e., 'hostel') and used different provided services (i.e., 'accommodation', 'food service') within the time period from 08:00 hrs. (8 a.m.) to 19:00 hrs. (7 p.m.).

FIGURE 3. User connection graphs under different travel attributes.

After computing the relatedness between users on which similarities are to be performed based on the heterogeneous travel network, we perform a weighted combination of those similarities. Then similarity search is performed to label the unlabeled user-pair. The problem of similarity search for travel behavior is formally defined in Definition 3.

Definition 3 Similarity search. Given a heterogeneous travel network $TN =< U, L, S, T, E, W >$ and a subset of user-pairs $H = (u_i, u_j) \subset U \times U$, predict the class (i.e., similar or dissimilar) labels for the unlabeled pairs $(U \times U - H)$. H is labeled with values $C = \{C_{ij}|0 \le C_{ij} \le 1\}$ denoting each pair is similar or not.

IV. SIMILARITY MEASUREMENT BASED ON A META-PATH

A. META-PATH FOR TRAVEL

A meta-path [11] was proposed to describe the paths that exist between the different nodes in the HINs. In the context of

traffic and travel, the travel attributes of different users had specific semantic connections. For example, user A arrived at an identified location and used a certain type of service, and user B arrived at the same location as well. The former's semantics could be described by a triad <travel user, stay point, accessed service>. Similarly, a triad < travel user, stay point, travel user> could be used to indicate that user A and user B arrived at the same location. Therefore, by searching the node sets in the heterogeneous travel network, the semantic paths of different lengths and different types between nodes could be constructed to obtain a meta-path. Firstly, we formally define the meta-graph in HIN for travel.

Definition 4 Travel relationships. According to the types of nodes in the heterogeneous travel network, three types of travel relationships R are defined: (1) users stay at a certain place (a location-based travel relationship) represented by R_1; (2) users access a certain type of service during their stay (a service-based travel relationship) represented by R_2; (3) users stay at a certain place within a certain time period (a time-based travel relationship) represented by R_3.

Definition 5: Travel meta-path. In the heterogeneous travel network TN, a travel meta-path is described as $meta-Path = \{U \xrightarrow{R_1} L,\ U \xrightarrow{R_1} L \xrightarrow{R_2} S, U \xrightarrow{R_1} L \xrightarrow{R_3} T\}$. R denotes the travel relationship between nodes, wherein $R = \{R_1^{\circ}R_2,\ R_1^{\circ}R_3\}$ defines the composite relations between node types.

For example, the location-based travel relationship can be described using the travel meta-path $U \xrightarrow{stay} L$, or short as UL(U is the start type and L is the end type). Hence the similarity search based on UL will give the user-pairs with the same stay points.

Definition 6: Instance of travel meta-path. For meta-path $meta - Path(mP)$, if real path $p = \{v_i, v_{i+1} \in U \cup L \cup S \cup T | v_i \xrightarrow{R_j} v_{i+1}\}$ exists and the relationship between node v_i and v_{i+1} is R_j for any i, the path p is called an instance of a travel meta-path. The set of all p that satisfies the condition is called the set of instances of the meta-path.

In Figure 4, we identify the heterogeneous relationships between each user-pair. Specifically, two users are similar in travel if they have visited the same stay points, or used the same services, or traveled at the same time. These relations can be characterized using three types of meta-paths.

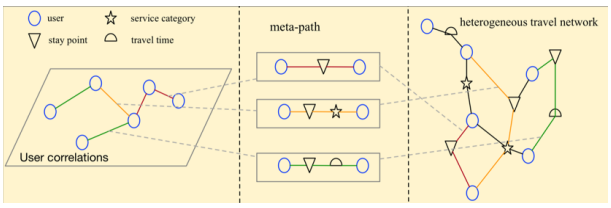


FIGURE 4. A heterogeneous travel network schema with meta-path. The edges in red denote two users have stayed at the same place, ones in yellow imply two users have chosen the same service categories at the same location, ones in green indicate two users have visited the same location within the same time period..

By associating different users via meta-paths, we can build a heterogeneous travel network that may capture the essential semantics of the real world.

B. SIMILARITY MEASURES BASED ON TRAVEL META-PATH

The similarity of travel behaviors can be determined by characterizing the travel meta-paths. In existing methods, the similarity is captured based mainly on each of a few characteristics and fused into a single representation. In this study, the stay points, service categories, and travel times were chosen as travel attributes, and a meta-path-based method was applied to the similarity measures by satisfying the attribute value constraint

1) SELECTION OF A META-PATH

In our work, we searched for similar users based on their travel relationships (e.g., spatial-temporal trajectories and service-visiting behaviors). According to the results of the study by Sun *et al.* [11], the correlation between users will be significantly reduced whenever the number of nodes contained in a meta-path is greater than 4. Moreover, we assumed that if two users have similar travel relationships, there must be a symmetric meta-path(e.g., '$U \xrightarrow{stay} L \xrightarrow{stay^{-1}} U$', '$U \xrightarrow{stay} L \xrightarrow{provide} S \xrightarrow{provide^{-1}} L \xrightarrow{stay^{-1}} U$', '$U \xrightarrow{stay} L \xrightarrow{within} T \xrightarrow{within^{-1}} L \xrightarrow{stay^{-1}} U$', as shown in Table 1). Therefore, only meta-paths with fewer than 4 nodes were constructed in this study.

TABLE 1. Definition of meta-path based on travel relationship.

Travel meta-path	Semantics of meta-path (travel relationship)	Similarity travel relationship based on meta-path
U→L	Users stay at a certain location (travel location)	If there is U→L→U, the two users have visited the same stay points, and the travel relationship between them is similar.
U→L→S	Users stay at a certain location and access a service type (use travel service)	If there is U→L→S→L→U, the two users have used the same services, and the travel relationship between them is similar.
U→L→T	Users stay at a certain location within a certain time period (travel time)	If there is U→L→T→L→U, the two users have traveled at the same time, and the travel relationship between them is similar.

2) QUANTITATIVE ANALYSIS FOR A META-PATH

Based on the selected meta-paths, we used the PathSim [33] method for measuring similarity by calculating the eigenvalue of the meta-path between users, as shown in (1). PathSim enables finding similar objects in the network when they are connected and share the same field.

$$\begin{aligned} PW_{x,y}(mp) &= \frac{2 \times \left|\left\{P_{x-y} : P_{x-y} \in mP\right\}\right|}{\left|\left\{P_{x-x} : P_{x-x} \in mP\right\}\right| + \left|\left\{P_{y-y} : P_{y-y} \in mP\right\}\right|} \\ &= \frac{2 \times (i \times j : i, j \in W_{ul}(E_{ul}(x, L) \cap E_{ul}(y, L)))}{\sum^{i^2} : i \in W_{ul}(x, L) + \sum^{j^2} : j \in W_{ul}(y, L)} \end{aligned} \quad (1)$$

where P_{x-y} denotes the number of meta-paths between user x and user y; P_{x-x} and P_{y-y} denote the number of meta-paths that connect user x and user y with themselves, separately; $E_{ul}(x, L)$ and $E_{ul}(y, L)(L \in \{\text{'}University\text{'}, \text{'}Park\text{'}, \text{'}Gym\text{'}\})$ denote the set of meta-paths for user x and user y to reach a particular location L; and $W_{ul}(x, L)$ *and* $W_{ul}(y, L)$ denote their corresponding weights. $E_{ul}(x, L) \cap E_{ul}(y, L)$ denotes the set of meta-paths for user x and user y to reach the same location. i and j represent the times of the visits of user x and user y to the particular locations, respectively.

We need to measure the relatedness between same-typed objects i.e., source and target object type of a meta-path would be the same. In this case, we measure the relatedness between users. An example is given in Figure 5. Following the meta-path $U \rightarrow L \rightarrow U$ (abbreviated as ULU), we compute the relatedness between users Tom and William. The weight on each path represents the number of times that a user arrives at a known location in a day. According to (1), $PW_{TOM,\ William}(ULU) = \frac{2\times(2\times8)}{(5\times5+2\times2)+(2\times2+8\times8)} = 0.330$.

The SimRank [39] method was also applied to estimate similarity by iteratively propagating similarity to neighbors until convergence (no similarity changes), as shown in (2). SimRank is improved over PageRank, and provides the similarity measurement based on network structure, with the intuition that two nodes are similar if nodes' in-neighbors are similar.

$$PW_{x,y}(mp)^{k+1} = \begin{cases} 1 & x = y \\ 0.8 \times \dfrac{\sum_{h\in I(x), l\in I(y)} PW_{h,l}(mp)^k}{|I(x)|\,|I(y)|} & x \neq y \end{cases} \quad (2)$$

where $I(x)$ *and* $I(y)$ denote the set of users x and y reaching particular locations. k denotes the number of iterations. According to Figure 5, $I(Tom) = \{University, Gym\}$ and $I(William) = \{Park, Gym\}$. Therefore,

$$\begin{aligned} PW_{TOM,William}(ULU) = 0.2 \times (&PW_{University,Park}(LUL) \\ &+ PW_{University,Gym}(LUL) \\ &+ PW_{Park,Gym}(LUL) + 0.2 \end{aligned}$$

Furthermore,

$$\begin{aligned} PW_{University,Gym}(LUL) &= 0.4 \times PW_{TOM,William}(ULU) \\ &\quad + 0.4 \\ PW_{University,Park}(LUL) &= 0.8 \times PW_{TOM,William}(ULU), \end{aligned}$$

and

$$PW_{Park,Gym}(LUL) = 0.4 \times PW_{TOM,William}(ULU) + 0.4$$

The iteration of $PW_{TOM,William}(ULU)$ continues until convergence.

3) MUTI-PATH SIMTRAVEL (MPST) ALGORITHM

The similarity between users in travel relationships can be determined by quantifying the instances of travel

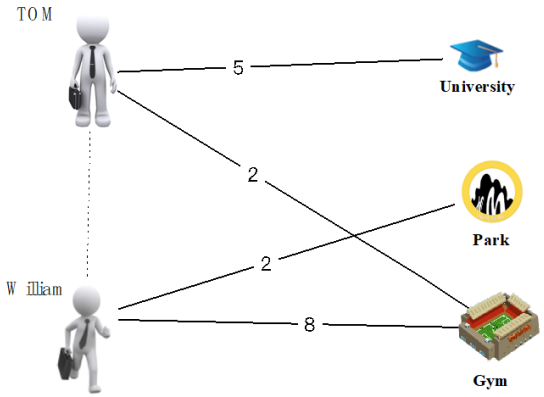


FIGURE 5. An example of the heterogeneous travel network model.

meta-paths between any two users. First, the similarity between user x and user y in mobility was determined by sample training. Then, according to the constructed symmetrical meta-path set $mP_{sym} = \{ULU, ULSLU, ULTLU\}$, we applied the quantitative analysis proposed in Section Section IV.B to determine an eigenvector on each metal-path in a weighted combination, where $\vec{\alpha} = \left(PW_{x,y}(ULU), PW_{x,y}(ULSLU), PW_{x,y}(ULTLU)\right)$.

Based on the logistic regression, the prediction results of similarity between two users are shown in Equation. (3):

$$\mathbf{Y} = \frac{1}{(1 + e^{-\vec{\alpha}\ \Theta})} \quad (3)$$

where vector $\mathbf{Y}$ denotes the similarity between user x and user y in the training set. We marked its value as 0 to represent the dissimilarity of users x and y, otherwise $\mathbf{Y} = 1$. Meanwhile, based on the training set, a supervised learning method was adopted to generate the weight vector Θ of the meta-path eigenvector. On this basis, different users in the testing set were selected for the prediction of similarity, namely vector $\mathbf{Y}$.

Algorithm 1 below describes the MPST algorithm. The meta-paths were selected by rows 1–9 of the algorithm from the set of users with labeled similarity, and either (1) or (2) was used to calculate their eigenvector in a meta-path combination. To facilitate the comparison, the algorithm that used (1) for calculating the characteristics was named MPS, and the algorithm that used (2) was named MSR. The generated eigenvector was used as input for the machine learning process by row 10 of the algorithm to obtain the weight Θ through data training. The set of users whose similarity was to be measured was selected from the testing set by rows 12–23. The eigenvector was calculated following the above method, and then, in the case of weight Θ, the similarity vector was obtained, which can determine the similar users.

V. PERFORMANCE ANALYSIS

A. EXPERIMENTAL SETUP

We used the Geolife trace dataset provided by Microsoft Research Asia, which contains 182 users and covers a period

Algorithm 1 MPST Algorithm

Input : Heterogeneous travel network TN; symmetrical meta-path set mP_{sym}; users' training set $U_{train} = \{< u_i, u_j >, (i, j = 1, 2, \hat{a}\check{A}\c{e}n, m)\}$; element number K; the similarity vector $\boldsymbol{S_{train}}$ of U_{train}; users' testing set $U_{test} = \{< u_h, u_l >, (h, l = 1, 2, \hat{a}\check{A}\c{e}n, m)\}$, weight Θ.

Output: The similarity vector $\boldsymbol{S_{test}}$ of U_{test}.

foreach $< u_i, u_j > \in (U_{train})$ **do**
  $k = 0$;
  foreach $P_{u_i - u_j} \in mP_{sym}$ **do**
    $PW_{u_i,u_j}\left(P_{u_i-u_j}\right) := 0$;
    foreach $p \in getInstandPatSet\left(P_{u_i-u_j}\right)$ **do**
      $PW_{u_i,u_j}(p)$;
    end
    $\alpha[k++] := PW_{u_i,u_j}(p)$;
  end
  $\Theta^* := argmax \sum\{S_{train} \cdot \Theta\alpha - \ln(1 + e^{\alpha\Theta})\}$;
end
foreach $< u_h, u_l > \in (U_{test})$ **do**
  $k = 0$;
  foreach $P_{u_h - u_l} \in mP_{sym}$ **do**
    $PW_{u_h,u_l}\left(P_{u_h-u_l}\right) := 0$;
    foreach $p \in getInstandPatSet\left(P_{u_h-u_l}\right)$ **do**
      $PW_{u_h,u_l}(p)$;
    end
    $\alpha[k++] := PW_{u_h,u_l}(p)$;
  end
  $S_{test} := e^{\alpha\Theta^*}/(e^{\alpha\Theta^*} + 1)$;
end
Return $\boldsymbol{S_{test}}$.

of five years. We extracted 3,891 stay points and then performed operations such as data cleansing and normalization to obtain the required spatial-temporal trajectories.

The cosine similarity was used for labeling the relevance of all user-pairs in the dataset. If the result was greater than, or equal to a threshold, it was treated as to be positive samples (i.e., each user-pair can be clustering), while the remaining ones were negative. For optimizing the thresholds for ground truth labels, we performed 10-fold cross validation. After we obtain groups of positive and negative samples, we defined a vector $x^n = x^{(1)}, \ldots, x^{(l)}, \ldots, x^{(L)}$ from all of the L candidate thresholds. The similarity rating for x^n based on a polynomial filter was computed as follows:

$$\hat{y}^n\left(x^n, w\right) = w_0 + \sum_{j=1}^{M} w_j (x^n)^j, \tag{4}$$

where M is the degree of the polynomial, w_0 is the global bias, $w \in \mathbb{R}^M$, representing the weights for the thresholds. The parameters can be learned by minimizing the mean square loss for each degree:

$$\min_{x^n,w,M} \sum_{n=1}^{L} (\hat{y}^n\left(x^n, w\right) - y^n)^2, \tag{5}$$

where y^n is an observed similarity rating, estimated by the variance of occurrences between positive and negative samples. Figure 6(a) shown that error loss of polynomial under different degrees. We fitted a polynomial of degree 3 using all data, and minimized the testing error and identifying the optimal threshold as 0.75, as shown in Figure 6(b).

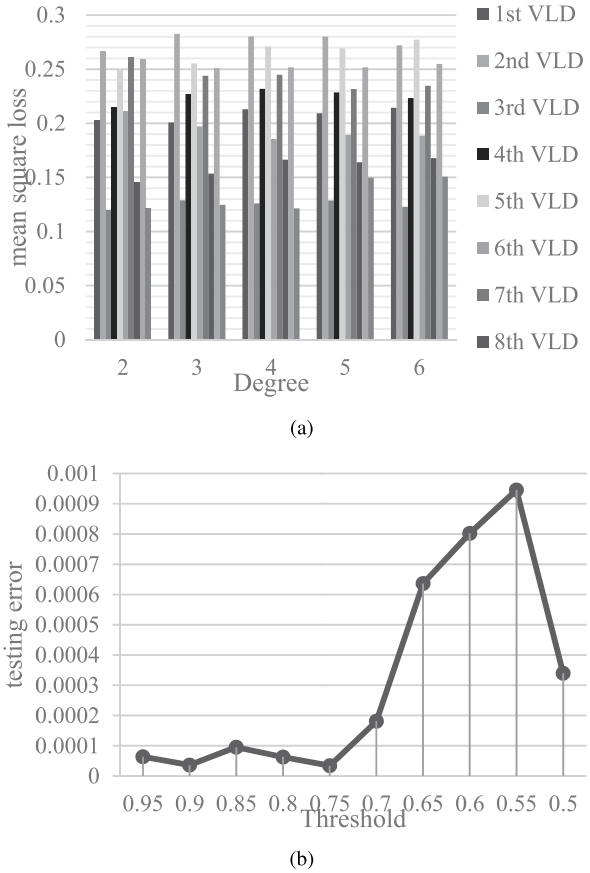


FIGURE 6. Illustration of 10-fold cross validation for minimizing the testing error and identifying the threshold for labelling as 0.75. (a) Error of polynomial of different degrees in K^{th} validation. Testing error of a polynomial of degree 3 on all dataset.

We could label the similarity of any two users, and get 1000 such pairs, containing 182 different persons. We consider these as our dataset, and label their ground truth, i.e., $\{< u_i, u_j >_h, Sim_h\}_{h=1}^{1000}$, where $< \cdot, \cdot >$ is the dot product of two users in the samples, and $Sim_h \in \{1, 0\}$ is the label.

We randomly split the above dataset into training and test ones by the ratio 8:2, i.e., 80% of the whole data are used for training and the remaining 20% are for testing. The labeled pairs are then used as the input to a logistic regression classifier for determining the corresponding weight of each meta-path. Finally, we selected the users in the testing set to

evaluate the performance of (1) and (2) with a single travel meta-path and MPST, separately.

B. PERFORMANCE EVALUATION OF THE PROPOSED METHOD

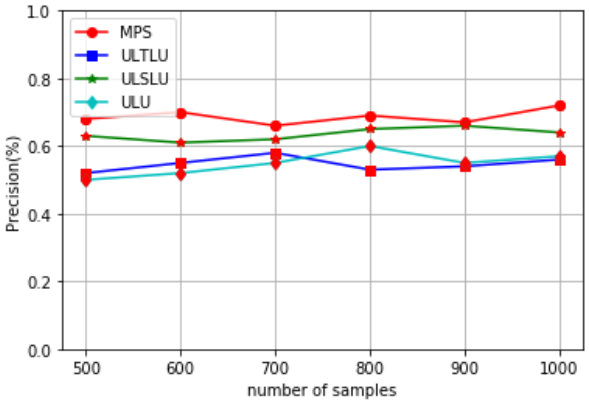
We evaluated the similarity measures performance of different methods using the indicators shown in Table 2.

Training data under different scales were used for the training and testing of the heterogeneous travel network. Precision and recall over all classes were computed by first summing TP, FP and FN over all partitions in testing data for each class separately, then applying the formulas shown in Table 2 for precision, recall and AUC on the sums, as argued in. Equation (1) from the PathSim method was used to analyze the precision and recall of an individual travel meta-path (i.e., 'ULU', 'ULSLU', and 'ULTLU') and multiple meta-path combinations under different data scales (Figure 7). Because of the large number of positions (i.e., 3,891 stay points) used in the experiment, for the 'ULU' meta-path, users could visit only some of the locations in one day. Consequently, the data describing users' location-based travel relationships showed sparsity and uneven distribution. For this reason, it was difficult for the travel network based on the 'ULU' meta-path to provide accurate similarity measurements, and the performance was the worst among all the meta-path-based models. In contrast, because there were 21 sample of service types and 4 sample of travel times in the dataset, the sparsity of data for the 'ULSLU' and 'ULTLU' meta-path-based models was improved, and their performance were improved as well.

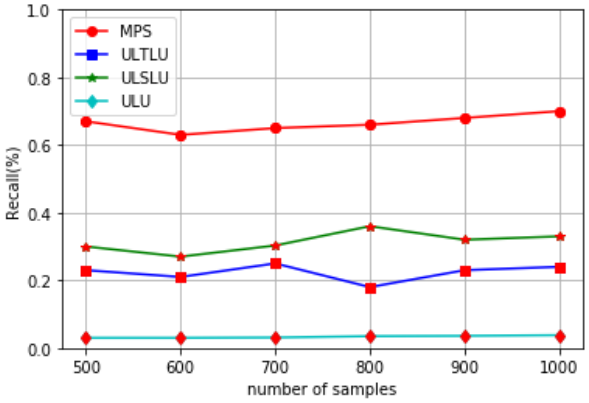
TABLE 2. Performance indicators of similarity measures.

Indicators	Description
TP	# of user-pairs correctly classified as being similar
FP	# of user-pairs mistakenly classified as being similar
FN	# of user-pairs mistakenly classified as being dissimilar
TN	# of user-pairs correctly classified as being dissimilar
Precision	TP/(TP +FP)
Recall	TP/(TP+FN)
AUC	probability that TPR>FPR, TPR=TP/(TP +FN), FPR=FP/(FP +TN)

However, in the travel network constructed based on the 'ULTLU' meta-path, a large number of paths were produced. The effect was that the proportion of meta-paths with similar time-based travel relationships was lowered, enabling this meta-path-based measures to perform between the other two meta-path-based measures. As shown in Figure 7, the MPST-based measures performed better than other methods in terms of precision and recall. This is consistent with our expectation because the fusion of multiple single paths and the use of machine learning allow to train and adaptively adjust the influence of meta-paths on the determination of travel behavior. Here, MPST limits the results to those persons who utilized the same services at the same period of time, which cannot be represented by 'ULTLU' or 'ULTLU' alone.



(a)

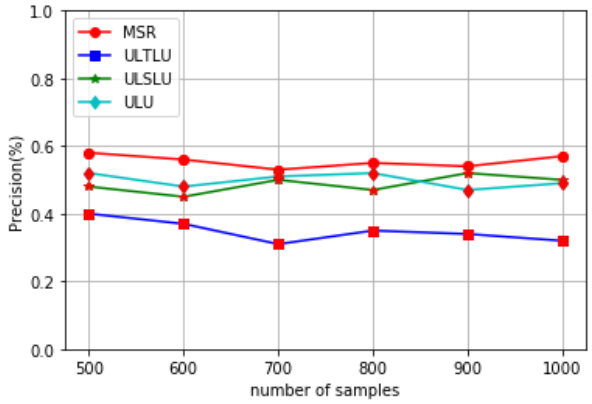


(b)

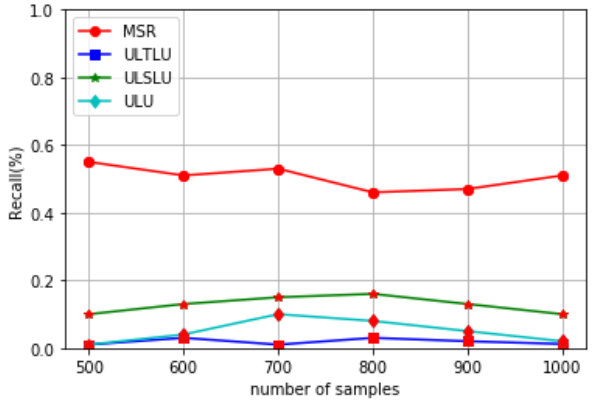
FIGURE 7. Comparison of a single travel meta-path and a meta-path combination in measuring similarity using PathSim. (a) Precision comparison. (b) Recall comparison.

Equation (2) from the SimRank method was also used to analyze the precision and recall under different scales of data, as shown in Figure 8. The MPST-based measures performed better than single meta-path-based measures on the whole. When the '*ULTLU*' meta-path was used for similarity measurement, for the SimRank method the similarity needed to be calculated iteratively based on the number of paths between the users and travel time periods. A large number of paths allowed for improving the user similarity. However, when calculating the similarity between different time periods, the in-degree node data set for different time periods increased drastically, resulting in a decrease in similarity and a poorer performance compared to other meta-path-based measures. Thus, the meta-path '*ULSLU*' and '*ULU*' yield better results than the meta-path '*ULTLU*' on all the SimRank-based measures. Also, the scales of in-degree node data generated using meta-paths '*ULSLU*' and '*ULU*' remained essentially the same when iterated at different times. Thus, the relative performance was basically similar.

To analyze the performance of the PathSim and SimRank method in similarity analysis, multiple travel meta-paths were chosen to build the heterogeneous travel network. On this

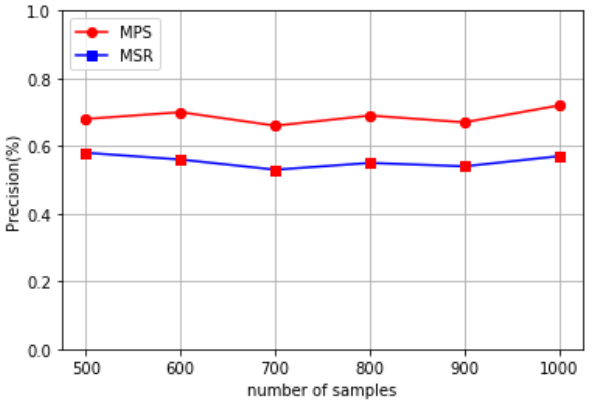


(a)

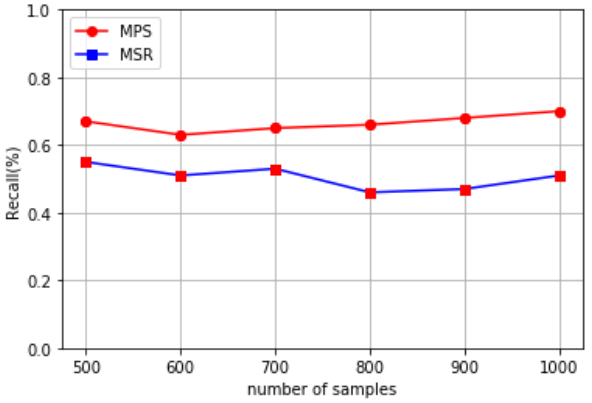


(b)

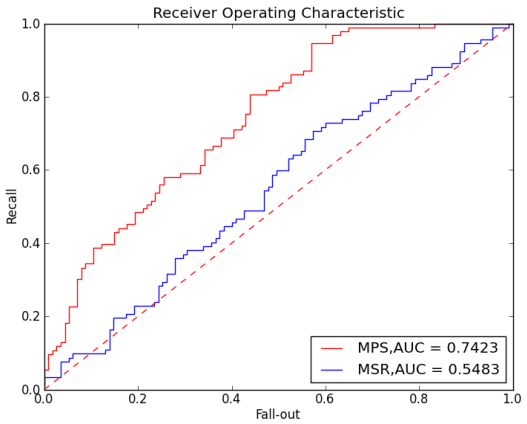
FIGURE 8. Comparison of a single travel meta-path and a meta-path combination in measuring similarity using SimRank. (a) Precision comparison. (b) Recall comparison.



(a)



(b)



(c)

FIGURE 9. Influence of different similarity measures. (a) Precision comparison. (b) Recall comparison. (c) Comparisons of ROC and AUC.

basis, the performances of MPS and MSR in precision and recall were compared. As shown in Figure 9, the method of using PathSim to quantify meta-path characteristics and assess similarity had a better performance than the method using SimRank. The performance of the improved PathSim was 15% higher than the performance of the improved SimRank in terms of precision and 21% higher in terms of recall. Specifically, due to the sparsity of data, the eigenvector on each metal-path with MSR was iteratively calculated to be rather small, where some of those were very close to 0. Therefore, the recall value fluctuated widely as the number of samples increased. For example, the value when the number of sample was equal 1000 was higher than that was equal 500. But, the eigenvector with MPS was stable for each metal-path, which were affected little by the sparsity. It is well able for MPS to learn the semantic relations between users well, the recall value thus improved as the number of samples increased. This outcome was the result of SimRank's focus on analyzing models based on a bipartite graph structure, and determining the similarity between objects having strong connections in different sets by multiple iterations. However, SimRank cannot describe the semantic relationships between objects, so it provided lower accuracy in calculating similarity than PathSim.

C. COMPARISONS OF MPST AND OTHER SIMILARITY MEASURES

In this set of experiments, based on the sample set described in Section V.A, we compare MPST with two other typical

TABLE 3. My caption.

Metric	MPST		BSCSE	PCRW
	PathSim	SimRank		
AUC	0.7423	0.5483	0.5640	0.2606

measurement methods, i.e., the BSCSE-based [40] and the PCRW-based measures. For each similarity measures, we plotted the ROC (receiver operating characteristic) Curve. We then compute the area under the curve, i.e., AUC. The experiment results are shown in Table 3. Observe that MPST-based measures using PathSim (i.e., MPS) further outperform these alternative methods. Although the PCRW-based measures is also a linear combination of single meta-paths, its AUC value, is just 0.2606, and is worse than MPST(i.e, 0.7423). This is because MPST combines the meta-paths by learning the influences of a single meta-path (i.e., 'ULSLU', 'ULU' or 'ULTLU') on the effectiveness, instead of varying the parameter values indicating each path' influence. The BSCSE-based measures is better than using PCRW as it can express the common nodes in the meta-paths. However, it needs to produce a new meta-structure each time along comes to another object (e.g., travel mode) taken into consideration. The PCRW and BSCSE-based measures also cannot be directly applied to weighted meta-paths proposed in MPST, because they do not consider the attribute value constraint on relations.

D. DISCUSSION

The information gathered from GPS trajectories, in general, is used to analyze the travel attributes [41]. Choosing the fine-grained information markers (e.g., semantics in the spatial layout) are especially useful in capturing latent relationships among users. We model the behavior similarities along two distinct makers: semantic properties of the locations and temporal duration of the trajectory. The similarities are computed by applying appropriate meta-paths to extract the key relevant attributes.

From the results on the datasets, we can understand that meta-path based similarity measures on travel attributes and classification on target users is effective. However, viewing the semantic contained by a meta-path as the sole determinant of similarity [42] may miss the real relationships or falsely identify the non-existent groups. Therefore, we should leverage the semantics of various meta-paths simultaneously for effective results.

The proposed method, MPST-based measures demonstrates that leveraging a weighted combination of various semantics in the travel network can improve the accuracy on similarity search. How to tune this weight is challenging because it would lead to the limited flexibility in processing real data. Some researchers conducted the online survey [43] on potential target users to assign the weights to different meta-paths. Such a survey needs incorporate a larger number of features into their user interfaces, at greater cost with respect to implementation time and code. Due to the problem of similarity search with multiple meta-paths, we plan a diffusion by learning the weight to guarantee the maximal consistency of different semantic and effective information fusion.

VI. CONCLUSIONS

For measuring the similarity of travel behavior, we described the semantic relationships between users by constructing a heterogeneous travel network and introducing multiple meta-paths for travel. We validated the similarity measures based on PathSim and SimRank. With a meta-path combination, we employed the logistic regression to learn the weight of each meta-path automatically, and further predicted the similarity between different users.

In this experimental study, we first performed the necessary operations such as data cleansing and normalization. Then we evaluated three types of travel meta-paths and MPST algorithms separately under different scales of data. Comprehensive experiments on real sample collections from Microsoft Research Asia were conducted to compare various similarity measurement approaches. Promising experimental results demonstrated that our proposed method outperforms other alternative measurement techniques. The reason behind this is that in MPST we use a more expressive representation for the data, and build the connection between the higher-level semantics of the data and the final results.

In this study, only the stay points, travel times, and categories of chosen services were considered in the proposed heterogeneous travel network. However, more attributes can be introduced into future studies to construct a travel network that describes more complicated relationships. When more objects and their relationships with each other are taken into consideration, such as the series relationship between locations, the accuracy of the similarity calculation will be improved. However, a challenge of the measures is its high computational cost because of the relevance computation for many object pairs over a HIN. It is important to ensure that these similarity measures can be efficiently evaluated in future work. Currently, only two similarity measurements for link-based structures were selected for performance analysis: PathSim and SimRank. In future studies, the heterogeneous travel network will be validated, and the performances of similarity measurements without the meta-paths, e.g., personalized PageRank will be evaluated to illustrate the universality of the proposed travel network.

ACKNOWLEDGEMENT

L. Tang thanks LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

REFERENCES

[1] K. Uchimura, H. Takahashi, and T. Saitoh, "Demand responsive services in hierarchical public transportation system," *IEEE Trans. Veh. Technol.*, vol. 51, no. 4, pp. 760–766, Jul. 2002, doi: 10.1109/tvt.2002.1015354.

[2] M. J. Roorda and T. Ruiz, "Long- and short-term dynamics in activity scheduling: A structural equations approach," *Transp. Res. A, Policy Pract.*, vol. 42, no. 3, pp. 545–562, 2008, doi: 10.1016/j.tra.2008.01.002.

[3] Z. Chen, J. C. Xia, B. Irawan, and C. Caulfied, "Development of location-based services for recommending departure stations to park and ride users," *Transp. Res. C, Emerg. Technol.*, vol. 48, no. 48, pp. 256–268, 2014, doi: 10.1016/j.trc.2014.08.019.

[4] H. Senaratne *et al.*, "Urban mobility analysis with mobile network data: A visual analytics approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1537–1546, May 2018, doi: 10.1109/tits.2017.2727281.

[5] N. Bicocchi, M. Mamei, A. Sassi, and F. Zambonelli, "On recommending opportunistic rides," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 12, pp. 3328–3338, Dec. 2017, doi: 10.1109/TITS.2017.2684625.

[6] C. Chen, D. Zhang, N. Li, and Z.-H. Zhou, "B-Planner: Planning bidirectional night bus routes using large-scale taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1451–1465, Aug. 2014, doi: 10.1109/tits.2014.2298892.

[7] G. Zhong, X. Wan, J. Zhang, T. Yin, and B. Ran, "Characterizing passenger flow for a transportation hub based on mobile phone data," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1507–1518, Jun. 2017, doi: 10.1109/tits.2016.2607760.

[8] S. M. Grant-Müller, A. Gal-Tzur, E. Minkov, S. Nocera, T. Kuflik, and I. Shoor, "Enhancing transport data collection through social media sources: Methods, challenges and opportunities for textual data," *IET Intell. Transp. Syst.*, vol. 9, no. 4, pp. 407–417, 2014, doi: 10.1049/.2013.0214.

[9] J. Lin, W. Yu, X. Yang, Q. Yang, X. Fu, and W. Zhao, "A real-time en-route route guidance decision scheme for transportation-based cyberphysical systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2551–2566, Mar. 2017.

[10] A. Schulz, P. Ristoski, and H. Paulheim, "I see a car crash: Real-time detection of small scale incidents in microblogs," presented at the 10th Extended Semantic Web Conf., Montpellier, France, 2013.

[11] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," *ACM Trans. Knowl. Discovery Data*, vol. 7, no. 3, pp. 1–23, 2013, doi: 10.1145/2513092.2500492.

[12] D. Pelzer, J. Xiao, D. Zehe, M. H. Lees, A. C. Knoll, and H. Aydt, "A partition-based match making algorithm for dynamic ridesharing," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2587–2598, Oct. 2015, doi: 10.1109/2015.2413453.

[13] Y. Li, R. Chen, L. Chen, and J. Xu, "Towards social-aware ridesharing group query services," *IEEE Trans. Services Comput.*, vol. 10, no. 4, pp. 646–659, Jul. 2017, doi: 10.1109/.2015.2508440.

[14] J. E. Kang and W. Recker, "The location selection problem for the household activity pattern problem," *Transp. Res. B, Methodol.*, vol. 55, no. 55, pp. 75–97, 2013, doi: 10.1016/2013.05.003.

[15] X. Pan, S. Rasouli, and H. J. P. Timmermans, "Modeling bounded rationality in choice behavior: Relative utility vs. random regret models," presented at the 13th Int. Conf. Design Decis. Support Syst. Archit. Urban Planning, Eindhoven, The Netherlands, Jun. 2016.

[16] X. Li, M. Li, Y.-J. Gong, X.-L. Zhang, and J. Yin, "T-DesP: Destination prediction based on big trajectory data," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2344–2354, Aug. 2016, doi: 10.1109/tits.2016.2518685.

[17] A. Elbery, M. Elnainay, F. Chen, C.-T. Lu, and J. Kendall, "A carpooling recommendation system based on social vanet and geo-social data," presented at the 21st Int. Conf. Adv. Geograph. Inf. Syst., Orlando, FL, USA, Nov. 2013.

[18] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 308–324, Sep. 2015, doi: 10.1016/j.trc.2015.02.019.

[19] E. Kamar and E. Horvitz, "Collaboration and shared plans in the open world: Studies of ridesharing," presented at the Int. Joint Conf. Artif. Intell., Pasadena, CA, USA, Jul. 2009.

[20] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui, "GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation," presented at the Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, Aug. 2014.

[21] W. He, K. Hwang, and D. Li, "Intelligent carpool routing for urban ridesharing by mining GPS trajectories," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2286–2296, Oct. 2014, doi: 10.1109/2014.2315521.

[22] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proc. ACM SIGKDD*, 2007, pp. 330–339.

[23] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transp. Res. B, Methodol.*, vol. 53, pp. 64–81, Jul. 2013.

[24] Y. Wang, Y. Zheng, and Y. Xue, "Travel time estimation of a path using sparse trajectories," presented at the KDD, New York, NY, USA, Aug. 2014.

[25] A. Hofleitner and A. Bayen, "Optimal decomposition of travel times measured by probe vehicles using a statistical traffic flow model," in *Proc. Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2011, pp. 815–821.

[26] A. Hofleitner, R. Herring, P. Abbeel, and A. Bayen, "Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1679–1693, Dec. 2012.

[27] S. Shang, R. Ding, B. Yuan, K. Xie, K. Zheng, and P. Kalnis, "User oriented trajectory search for trip recommendation," in *Proc. EDBT*, 2012, pp. 156–167.

[28] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie, "We know how you live: Exploring the spectrum of urban lifestyles," in *Proc. ACM Conf. Online Social Netw.*, 2013, pp. 3–14.

[29] S. Shang, R. Ding, K. Zheng, C. S. Jensen, P. Kalnis, and X. Zhou, "Personalized trajectory matching in spatial networks," *VLDB J.-Int. J. Very Large Data Bases*, vol. 23, no. 3, pp. 449–468, 2014.

[30] N. Zhou, W. X. Zhao, X. Zhang, J.-R. Wen, and S. Wang, "A general multi-context embedding model for mining human trajectory data," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 1945–1958, Aug. 2016.

[31] X. Liu, Y. Liu, K. Aberer, and C. Miao, "Personalized point-of-interest recommendation by mining users' preference transition," in *Proc. CIKM*, San Francisco, CA, USA, Oct. 2013, pp. 733–738.

[32] R. Trasarti, F. Pinelli, M. Nanni, F. Giannotti, "Mining mobility user profiles for car pooling," presented at the Int. Conf. Knowl. Discovery Data Mining, San Diego, CA, USA, Aug. 2011.

[33] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," presented at the 37th Int. Conf. Very Large Data Bases, Seattle, WA, USA, Aug. 2011.

[34] B. Cao, X. Kong, and S. Y. Philip, "Collective prediction of multiple types of links in heterogeneous information networks," presented at the IEEE Int. Conf. Data Mining, Shenzheng, China, Dec. 2014.

[35] C. Shi, Z. Zhang, P. Luo, P. S. Yu, Y. Yue, and B. Wu, "Semantic path based personalized recommendation on weighted heterogeneous information networks," presented at the 24th Int. Conf. Inf. Knowl. Manage., Melbourne, VIC, Australia, Oct. 2015.

[36] Y. Ye, S. Hou, Y. Song, and M. Abdulhayoglu, "HinDroid: An intelligent Android malware detection system based on structured heterogeneous information network," presented at the Int. Conf. Knowl. Discovery Data Mining, Halifax, NS, Canada, Aug. 2017.

[37] Y. Sun and J. Han, "Meta-path-based search and mining in heterogeneous information networks," *Tsinghua Sci. Technol.*, vol. 18, no. 4, pp. 329–338, Aug. 2013, doi: 10.1109/2013.6574671.

[38] Z. Duan, L. Tang, X. Gong, and Y. Zhu, "Personalized service recommendations for travel using trajectory pattern discovery," *Int. J. Distrib. Sensor Netw.*, vol. 14, no. 3, pp. 1–12, 2018, doi: 10.1177/1550147718767845.

[39] W. Zheng, L. Zou, L. Chen, and D. Zhao, "Efficient SimRank-based similarity join," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–37, 2017, doi: 10.1145/3083899.

[40] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li, "Meta structure: Computing relevance in large heterogeneous information networks," presented at the 22nd Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, Aug. 2016.

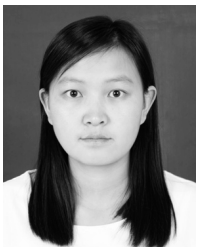
[41] S. Liu and S. Wang, "Trajectory community discovery and recommendation by multi-source diffusion modeling," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 4, pp. 898–911, Apr. 2017.

[42] M. Gupta, P. Kumar, and B. Bhasker, "HeteClass: A meta-path based framework for transductive classification of objects in heterogeneous information networks," *Expert Syst. Appl.*, vol. 68, pp. 106–122, Feb. 2017.

[43] M. Berlingerio, B. Ghaddar, R. Guidotti, A. Pascale, and A. Sassi, "The GRAAL of carpooling: GReen and sociAL optimization from crowdsourced data," *Transp. Res. C, Emerg. Technol.*, vol. 80, pp. 20–36, Jul. 2017.

LEI TANG was born in Jiangyou, Sichuan, China, in 1983. She received the Ph.D. degree in computer science and technology in 2012. She is currently with the School of Information Engineering, Chang'an University, China.

She was a Visiting Researcher at the Chair of Information Systems, Mannheim University, Germany, from 2009 to 2010. Her research interests include the area of intelligent transportation systems and pervasive computing. She is a member of ACM and the China Computer Federation.



YALING ZHAO was born in Tianshui, Gansu, China, in 1995. He is currently pursuing the degree with the School of Computer Science, Chang'an University, China. Her research interests include network embedding. She is a Student Member of CCF.

ZONGTAO DUAN was born in Baoji, Shaanxi, China, in 1977. He received the Ph.D. degree in computer science from Northwestern Polytechnical University, China, in 2006. He is currently a Professor at the School of Information Engineering, Chang'an University, China. He was a Post-Doctoral Research Fellow with the University of North Carolina, USA, from 2009 to 2010. His research interests include context-aware computing in transportation.

Dr. Duan is a member of CCF, CCF High Performance Computing, and Pervasive Computing Technical Committee.

JUN CHEN was born in Xinxiang, Henan, China, in 1994. He received the master's degree in computer science and technology in 2018. His research interests include heterogeneous information networks and machine learning.