IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 28, 2018, accepted January 13, 2019, date of publication January 31, 2019, date of current version March 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2896577

A Semantic User Distance Metric Using GPS Trajectory Data

ZEDONG LIN[1], QINGTIAN ZENG[1,2], HUA DUAN[3], CONG LIU[1], (Student Member, IEEE), AND FAMING LU[1]

[1]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[2]College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[3]College of Mathematics and System Science, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding authors: Qingtian Zeng (qtzeng@163.com) and Hua Duan (huaduan59@163.com)

This work was supported in part by the NSFC under Grant 61472229, Grant 61602278, Grant 61602279, Grant 71704096, Grant 61702306, and Grant 31671588, in part by the Sci. and Tech. Development Fund of Shandong Province of China under Grant 2016ZDJS02A11, Grant 2014GGX101035, and Grant ZR2017MF027, in part by the Humanities and Social Science Research Project, Ministry of Education, under Grant 16YJCZH154, Grant 16YJCZH041, Grant 18YJAZH017, and Grant 16YJCZH012, in part by the Taishan Scholar Climbing Program of Shandong Province, in part by the Shandong Province Postdoctoral Innovation Project under Grant 201603056, in part by the Fund of Oceanic Telemetry Engineering and Technology Research Center, State Oceanic Administration, under Grant 2018002, and in part by the Shandong University of Science and Technology (SDUST) Research Fund under Grant 2015TDJH102.

ABSTRACT User similarity measure plays an important role in various location-based services including location prediction and recommendation. However, existing similarity computation methods fail to meet the distance metric axioms. In addition, existing works also suffer from some deficiency when identifying indoor stay regions and representing semantic information. To address these issues, this paper proposes a new method to evaluate user similarity by analyzing the global positioning system (GPS) trajectory data. Specifically, a more accurate algorithm for indoor stay region identification is proposed by taking velocity into account. Word embedding technique is used to compute the semantic distance between two stay regions. After that, stay regions are clustered and the user's GPS trajectories are represented as a multiset of semantic label sequences. Then a distance metric of these multisets satisfying the distance metric axioms is proposed on the basis of the balanced transportation problem. Finally, the effectiveness of the proposed method is evaluated by experiments using both synthetic and real-life datasets.

INDEX TERMS GPS trajectory data, location-based service, user similarity, mobility profile, distance metric.

I. INTRODUCTION

With the popularity of GPS-enabled devices, large amount of users' spatiotemporal data are recorded. These data contain valuable information for user behavior analysis. For instance, many efforts have been devoted to analyze users' spatiotemporal data to provide location-based services, such as location prediction [1], [2], location recommendation [3], [4], friend recommendation [5], [6], community discovery [7], [8] and link prediction [9]. For the above applications, the user similarity measure is an extremely important step. For example, similarity results serve as a basis for user clustering or classification algorithms. However, existing user similarity computation methods do not meet distance metric requirements properly, i.e. they are not symmetric, positive, or not holding the triangular inequality. Metric violations prohibit the normal use of some machine learning algorithms, which typically have been formulated for metric data only [10], for example, the nearest neighbor search [11], fast clustering [12] and large-interval classifier [13] accelerate the performance by using the properties of distance metric of input data.

In addition, the performance of many learning and data mining algorithms heavily relies on the representation of the input space. For example, principal component analysis [14] and support vector machine [15] are based on a vector representation of the input space. Existing methods measure similarity between two users in a pairwise manner. The non-metric pairwise comparison as the output of these algorithms cannot be embedded in a vector (Euclidean) space without distortion. Therefore, forcing non-metric pairwise data to be embedded in vector space is typically equivalent to twisting data to measure, which destroys the distance structure of the input space.

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

To address the above problems, this paper presents a new user distance measure that satisfies the distance metric axioms to facilitate the GPS trajectories-based user behavior similarity computation. The main contributions are summarized as follows:

- Firstly, proposed a new method to identify stay regions by considering not only the duration of stay and the size of regions, but the moving velocity between locations, based on which the accuracy of indoor stay region identifications is improved greatly.
- Secondly, proposed a new method to semantically represent a stay region as a multiset of words of POI (Point Of Interest). In this way, the semantic distance between two stay regions is defined by word embedding technique by representing semantic of words as vectors. Based on the proposed method, semantic representation of stay regions is much easier and more reasonable compared to existing POI-based technique [16] and Term Frequency Inverse Document Frequency (TF-IDF) vector-based technique [17].
- Next, introduced the concept of semantic trajectory to model users' GPS trajectories in a semantic space. User mobility profile is represented as a multiset of semantic trajectories. The distance between user mobility profiles is computed based on the optimal solution of a balanced transportation problem. The proposed distance measure is proved to meet all the distance metric axioms.
- Lastly, performed comprehensive experiments on both synthetic and real-life datasets to compare the proposed method to existing benchmark methods. The experimental results show that our method achieves the best performance and accuracy.

The rest of this paper is organized as follows: Section II surveys some related works and highlights their limitations. The basic idea and some common methods are described in Section III. Then, the proposed framework of computing distance of user mobility profiles, together with the detailed algorithms, is described in Section IV. Section V evaluates the proposed approach with two experiments. Section VI summarize the paper.

II. RELATED WORK

This paper identifies similar users by comparing their mobility profiles discovered from GPS-based trajectories. We mainly introduce related work on users' mobility profiles construction and users' mobility profiles comparison.

A. USER'S MOBILITY PROFILE CONSTRUCTION

Yoshida *et al.* [18] introduce the concept of delta patterns that are represented as ordered lists of items with time intervals and present a heuristic algorithm to find frequent delta patterns. Giannotti *et al.* [19] propose the concept of temporally annotated sequences (TASs). Essentially, TASs are an extension of frequent sequential patterns (FSPs) by adding transition time information between their elements to sequences. Furthermore, Giannotti *et al.* [20] define the concept of trajectory patterns to represent the behavior patterns of moving objects. The trajectory pattern contains the same sequence of POIs with similar transition time. Lv *et al.* [21] first use place preference vectors to represent users' activities in one day, and then users' long-term activity regularities are discovered by a hierarchical clustering algorithm. Li *et al.* [22] introduce the stay point to represent a geographic area where a user stayed for a period of time. A density-based clustering algorithm is used to cluster these stay points into geospatial regions in a divisive manner such that each user refers to a hierarchical graph. Each level in this hierarchical graph is composed of a sequence of geospatial areas with corresponding granularity. The common geospatial areas are identified to analyze users' mobility. However, the above-mentioned works model the users' mobility without considering the semantics of locations. Alvares *et al.* [23] integrate geographic information into trajectories to extract more meaningful moving patterns of objects. Xiao *et al.* [24] expand the work of [22] where a stay region is represented as a rectangular geographic region centered on a stay point. The semantics of a stay region is represented by one TF-IDF vector that is constructed by categories of all POIs within the same geographic region. Then, a hierarchically clustering algorithm is used to group these feature vectors into semantic locations and a hierarchical graph that is similar to [22] is used to model user's mobility profile. Mazumdar *et al.* [16] introduce the concept of significance score to facilitate the identification of stay points. The POI category of one single stay point within a stay region is used to represent the semantics of the stay region. The frequent sequential patterns of POI categories are obtained by a sequential pattern mining method and are used to analyze moving patterns for each user. However, these methods are not accurate enough for identifying indoor stay regions.

B. USER'S MOBILITY PROFILE COMPARISON

Horozov *et al.* [25] use users' votes on the POIs to construct a vector, and then compute the user similarity through the Pearson coefficient. Li *et al.* [22] propose a framework to estimate the similarity between users, referred to as hierarchical-graph-based similarity measurement. The framework determines the degree of similarity using the sequence property of users' movements and the hierarchy property of geographic spaces. Mazumdar *et al.* [16] propose a Check-in Distribution based Similarity measure (CDS), which compares the similarity between users by considering not only the length and support of common location sequences, but also the distribution of users' check-ins performed on each day in a week. However, all these methods depend heavily on geographical overlaps, and therefore it is extremely challenging to evaluate the similarity between two users who have similar interests but live far away. For this reason, Zheng *et al.* [26] extend the work in [22] to compute the similarity of users by integrating the content to a geospatial region by exploring the categories of POIs within the region. Ying *et al.* [27] propose a method to compare user similarity semantically

based on maximal semantic trajectory pattern (MSTP). This method transforms a sequence of stay regions into a semantic trajectory by using landmark categories in geographic information databases. The similarity between users is defined as the weighted average of maximum semantic trajectories. Chen *et al.* [28] find that in some cases the MSTP measure fails to produce maximum similarity between two identical users. In addition, a similarity measure, named as Maximal Trajectory Pattern (MTP), is proposed to improve the MSTP where the similarity between two users is computed using the length of the longest common sequences between the maximal trajectory patterns and their supports. However, two users can never be considered to be identical even if they have similar patterns with distinguished frequencies. To address this limitation, Chen *et al.* [29] propose a method to estimate the similarity between users based on Common Pattern Set (CPS), based on which the common pattern between two trajectories is identified. The length and support of the common patterns are used to compute the relative importance. Both the relative importance and the distribution of support values of the common patterns are used to compare the similarity between two users. More recently, Mazumdar *et al.* [16] propose a common Patterns Distribution-based Similarity measure (PDS) to compute the similarity between pairwise users. The relative importance to a user is computed using the length and support of common patterns. The similarity between two users is the relative importance weighted by the ratio of the difference between the number of checked-in days by the users and the maximum number of checked-in days. However, these methods cannot satisfy the distance metric axioms.

III. METHOD OVERVIEW

This section explains the research methodology and techniques used in this paper. In addition, the rationality behind is also explained.

A. PRINCIPLES OF THE PROPOSED METHOD

Firstly, we present the limitations of existing works in stay region identification, semantic representation of stay regions, and mobility profile distance computation. Then the basic principles to address these issues are given. Before that, the basic concepts are defined as follows.

Definition 1 (GPS Point): A GPS point is a 2-tuple $l = (lng, lat)$*, where lng represents longitude and lat represents latitude.*

Definition 2 (GPS Trajectory and Spatiotemporal Point): A GPS trajectory $tr = < p_0, \ldots, p_n >$ *is a sequence of spatiotemporal points that are fully ordered by timestamps, where* $p_k = (l_k, t_k)(0 \leq k \leq n)$ *is a spatiotemporal point,* l_k *is a GPS point, and* t_k *is a timestamp* $(\forall 0 \leq k < n, t_k < t_{(k+1)})$.

Definition 3 (Stay Region): A Stay region sr is a geographic area where a user stayed over a time threshold δT *within a distance threshold* δD.

Existing works cannot identify indoor stay regions accurately. The basic idea of these works is to extract a consecutive

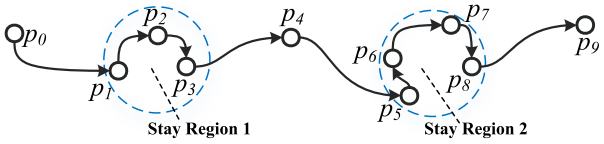


FIGURE 1. The stay regions.

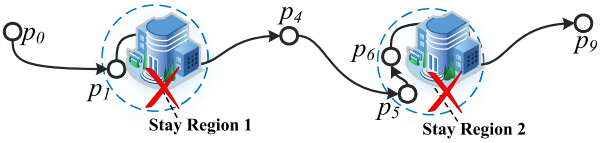


FIGURE 2. The indoor stay regions.

spatiotemporal points $sr = \langle p_i, \ldots, p_j \rangle$ from a GPS trajectory such that $\forall i \leq z \leq j$, $dist(p_i, p_z) \leq \delta D$, $dist(p_i, p_{j+1}) > \delta D$, $time(p_i, p_j) \geq \delta T$, where $dist(p_i, p_z)$ denotes geographical distance between p_i and p_j, $time(p_i, p_j)$ denotes a time interval between p_i and p_j. Figure 1 shows the identification results for a GPS trajectory $<p_0, p_1, \ldots, p_9>$ by existing stay region identification methods. Consider the example of the GPS trajectory shown in Figure 1. Based on the results in Figure 1, we have $dist(p_1, p_4) > \delta D$, $dist(p_5, p_9) > \delta D$ and $time(p_5, p_6) < \delta T$. Suppose that, as shown in Figure 2, indoor spatiotemporal points p_2, p_3, p_7 and p_8 cannot be collected. For a single p_1, that $dist(p_1, p_4) > \delta D$ causes the stay region 1 cannot be identified. For $<p_5, p_6>$, although $dist(p_5, p_6) \leq \delta D$, $time(p_5, p_6) < \delta T$ still prevents the formation of the stay region 2 due to the absence of p_7, p_8. In Section IV-A.1, we present a more accurate identification algorithm of indoor stay regions by taking the moving velocity between locations into account.

In addition, the semantics of users' stay regions is important for revealing user's interest. Therefore, reference [16] uses a single pre-defined category of a location to represent the semantics of a stay region. However, it is almost impossible to accurately determine a single POI category to represent the semantics of users' activities in a stay region. In fact, GPS contains random10-meter or more errors to the real position. Sometimes, there could be multiple POIs that correspond to the same location in high-rise buildings. Reference [17] employs TF-IDF technology to construct a feature vector to represent the semantics of a stay region according to categories of all POIs in the stay region. Usually there are a few POIs in a stay region, so the set of categories of all POIs in the stay region can be a short text. Existing work have shown that TF-IDF technology is not suitable for semantic representation of short texts [30]. In this work, each POI category label is treated as a word set. Then we use a multiset of those words to represent the semantics of user's stay regions. Based on this semantic representation, the distance between two stay regions is defined by applying word embedding technique. Details can be found in Section IV-A.2

Finally, existing similarity computation methods of users' GPS trajectories fail to meet the distance metric axioms. To address the problem, we convert users' distance to an

optimal solution of a balanced transportation problem. Based on the conversion, a user distance metric is able to satisfy all distance metric axioms in IV-B.

B. COMMON TECHNIQUES

In this paper, the semantics of user's activities in a stay region and the user mobility profile are expressed as a multiset. One of the key tasks is to propose a novel metric that satisfies all distance axioms to measure distance between two multisets.

Definition 4 (Feature Multiset): A feature multiset is a set of tuples (p_i, w_i), where p_i is a feature, and w_i is the multiplicity of feature p_i.

Definition 5 (Feature Multiset Distance): Given two feature multisets $F = \{(p_1, w_1), \ldots, (p_m, w_m)\}$ and $F' = \{(p'_1, w'_1), \ldots, (p'_n, w'_n)\}$, where (1) c_{ij} represents the mutual transportation amount between p_i and p'_j; and (2) $d(p_i, p'_j)$ is a pre-defined cost function per unit of mutual transportation between p_i and p'_j, called ground distance. The feature multiset distance between F and F' is defined as the minimum cost of mutual transportation between them:

$$d_F(F, F') = \arg\min_{c_{ij}} \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} d(p_i, p'_j)$$

$$s.t. \begin{cases} \sum_{j=1}^{n} c_{ij} = \frac{w_i}{w} & 1 \le i \le m \\ \sum_{i=1}^{m} c_{ij} = \frac{w'_j}{w'} & 1 \le j \le n \\ \sum_{i=1}^{m} w_i = w \\ \sum_{j=1}^{n} w'_j = w' \\ c_{ij} \ge 0 \end{cases}$$

Given two feature multisets $F = \{(A, 1), (B, 1)\}$ and $F' = \{(A, 1), (B, 3)\}$, a pre-defined symmetric cost per unit function $d(\cdot, \cdot)$ with $d(A, A) = 0$ and $d(A, B) = 1$. Since $d(\cdot, \cdot)$ is symmetric, the feature multiset distance is also symmetric. Therefore, the features in F as sources or sinks will not influence the results of feature multiset distance between F and F'. Without loss of generality, we assume that the features in F are the sources and the features in F' are the sinks. The multiplicities of the features in F and F' are normalized. Each feature p_i in F can be transported to any feature in F' in part or in whole with the normalized multiplicities. $d(p_i, p'_j)$ is a cost function per unit of transportation. The goal of the whole transportation process is to find an optimal plan c_{ij} to minimize the total cost of transportation from F to F'. An optimal plan is shown in Figure 3 where we have $d_F(F, F') = 0.25 \times 0 + 0.5 \times 0 + 0.25 \times 1 = 0.25$.

Based on Definition 5, we have $\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} = \sum_{i=1}^{m} \frac{w_i}{w} = \frac{\sum_{i=1}^{m} w_i}{w} = \frac{w}{w} = 1$. Similarly, we have $\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} = 1$. Therefore, the feature multiset distance can be regarded as a balanced transportation problem.

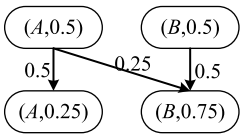


FIGURE 3. The distance between two feature multisets.

Algorithm 1 Feature Multiset Distance Computation

Input: Two feature multisets, F and F' respectively
Output: $d_F(F, F')$, the distance between F and F'
Data: G is a residual graph; $b(v)$ is the supply ($b(v)>0$) or the demand ($b(v)<0$) of a vertice v; $f(e)$ is the transportation amount of an edge e; $c(e)$ is the cost of an edge e

1 $w := \sum_{(f_j, w_j) \in F} w_j$; $w_i := w_i / w$ for all $(p_i, w_i) \in F$;
2 $w := \sum_{(f_j, w_j) \in F'} w_j$; $w_i := w_i / w$ for all $(p_i, w_i) \in F'$;
3 Multiply by α to make weights in F and F' integers;
4 **for** $(p_i, w_i) \in F$) **do**
5 Create a vertice s for G; $b(s) := w_i$;
6 **for** $(p_j, w_j) \in F'$) **do**
7 Create a vertice t for G; $b(t) := -w_j$;
8 Create a directed edge $e(s, t)$ for G;
9 cost $c(e) := d(p_i, p_j)$;
10 $f(e) := 0$ for all $e \in E(G)$;
11 $\gamma := 2^{\lfloor \log \beta \rfloor}$, where $\beta = max\{b(v): v \in V(G)\}$;
12 **if** $b = 0$ **then**
13 **go to** 22;
14 **else**
15 Choose vertices s and t with $b'(s) \ge \gamma$ and $b'(t) \le -\gamma$;
16 **if** *there is no such pair* (s, t) **then go to** 21;
17 Find an edge $e = (s, t)$ in G of minimum weight;
18 $b(s) := b(s) - \gamma$ and $b(t) := b(t) + \gamma$;
19 $f(e) := f(e) + \gamma$;
20 **go to** 12;
21 $\gamma := \gamma / 2$; **go to** 12;
22 **return** $\sum_{e \in E(G)} \frac{c(e) f(e)}{\alpha}$;

For this problem, there always exists an optimal solution. The process to solve the feature multiset distance is shown in Algorithm 1. Specifically, the multiplicities of features in F and F' are first normalized, and the multiplicities are made integral using a common coefficient (lines 1-3). Then, a residual graph G (lines 4-9) is constructed based on the two feature multisets. The sources and the sinks in the residual graph G represent the features in F and F' respectively. A directed edge is added between any pair of source p_i and sink p'_j. The cost of the edge is the feature distance $d(p_i, p'_j)$. Next, the minimum cost flow f (lines 10-21) is solved using the Capacity Scaling Algorithm [31]. Finally, the feature multiset distance between F and F' is computed according to the minimum cost flow f (line 22).

The time complexity of the Algorithm 1 is mainly determined by the capacity scaling algorithm. According to [31], the time complexity of the capacity scaling algorithm is $O(N^2(N + 2\log 2N)log(2 + \beta))$, where N represents the maximum number of features in two feature multisets.

To investigate the properties of the feature multiset distance, we need to first introduce the concept of distance metric.

Definition 6 (Distance Metric): Given a set S, a function $d : S \times S \rightarrow R^+ \cup \{0\}$ is a distance metric or distance function if it satisfies the following four axioms. For all $x, y, z \in S$, (1) Reflexivity, i.e., $d(x, y)=0 \Leftrightarrow x = y$; (2) Non-negativity, i.e., $d(x, y) \geq 0$; (3) Symmetry, i.e., $d(x, y) = d(y, x)$; (4) Triangle inequality, i.e., $d(x, z) \leq d(x, y) + d(y, z)$.

Theorem 1: The feature multiset distance is a distance metric if the feature distance is a distance metric.

Proof: According to [32], the balanced transportation problem is a distance metric if the ground distance is a distance metric. Definition 5 shows that the feature multiset distance can be regarded as a balanced transportation problem. The feature distance as ground distance is a distance metric, therefore, the feature multiset distance is a distance metric.

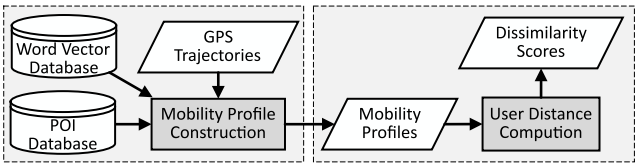


FIGURE 4. The proposed framework.

IV. PROPOSED FRAMEWORK

As depicted in Figure 4, the proposed framework mainly includes two phases: mobility profile construction and user distance computation. More specifically, as users' activities in different or non-overlapping geographic spaces may show the same interests and hobbies, a POI database and a word embedding database are employed to construct users' mobility profiles in a semantic space instead of a geographic space. Then, the distance of mobility profiles is introduced to measure the distance between users.

A. MOBILITY PROFILE CONSTRUCTION

As shown in Figure 5, the process of constructing mobility profile includes three steps: (1) stay region identification; (2) semantic representation of stay region; and (3) modeling individual mobility profile. Firstly, the geographical areas that are stayed for a period time are extracted from each user's spatiotemporal data. Then, the semantics of a user's activities in a geographical area are expressed through the categories of POIs within the area. Finally, each user's mobility profile is constructed in semantic spaces.

1) STAY REGION IDENTIFICATION

This paper introduces a velocity threshold to facilitate the identification of indoor stay regions. Details of stay point identification are depicted in Algorithm 2.

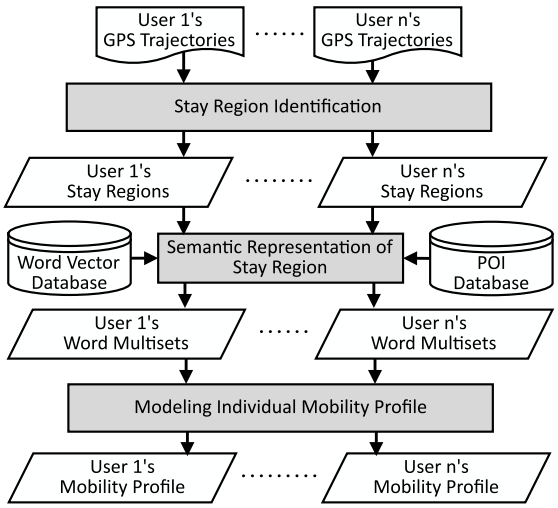


FIGURE 5. The construction of mobility profile.

Algorithm 2 Stay Region Identification

Input: tr, a trajectory; δD, the distance threshold; δT, the time threshold; δV, the velocity threshold
Output: ssr, the set of identified stay regions
Data: sr is a candidate stay region; p_s, p_e are the first and the last spatiotemporal points of the candidate stay region

```
sr := {p_0}; p_s := p_0; p_e := p_0;
for p ∈ tr/{p_0} do
    if dist(p_s, p) ≤ δD then
        sr := sr ∪ {p};
        p_e := p; continue;
    if time(p_s, p_e) ≥ δT then
        ssr := ssr ∪ {sr}; sr := {p};
        p_s := p; p_e := p; continue;
    if time(p_s, p) < δT then
        sr := sr ∪ {p}/{p_s};
        p_s := p_{s+1}; p_e := p; continue;
    if v(p_e, p) := dist(p_e, p)/time(p_e, p) < δV then
        ssr := ssr ∪ {sr}; sr := {p};
        p_s := p; p_e := p; continue;
    sr := sr ∪ {p}/{p_s};
    p_s := p_{s+1}; p_e := p;
return ssr;
```

Figure 6 demonstrates the identification process of stay regions from the trajectory in Figure 2 by Algorithm 2. At the beginning, the candidate stay region is $sr = \{p_0\}$. Then, we check whether sr is a stay region or not by using the three thresholds δD, δT and δV. sr is not a stay region as we have $dist(p_0, p_1) > \delta D$, $time(p_0, p_0) < \delta T$ and $v(p_0, p_1) > \delta V$. By removing the first element p_0 from sr and adding p_1 to sr, we obtain $sr = \{p_1\}$. Suppose $v(p_1, p_4) < \delta V$, then $dist(p_1, p_4) > \delta D$, $time(p_1, p_1) < \delta T$ and $time(p_1, p_4) > \delta T$,

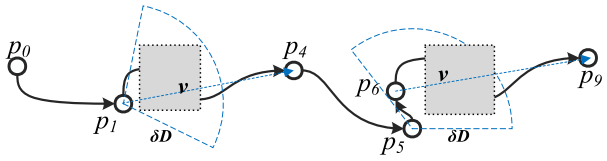


FIGURE 6. The identification of indoor stay regions.

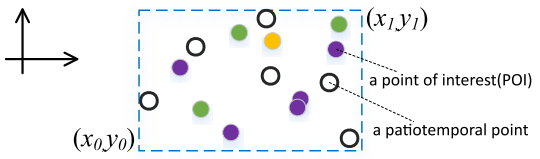


FIGURE 7. The semantic representation of stay regions.

which ensure $sr = \{p_1\}$ to be a stay region. Next, sr is assigned to $\{p_4\}$. A stay region $\{p_5, p_6\}$ can be made in the same way worked on other spatiotemporal points. Finally, two stay regions $\{p_1\}$ and $\{p_5, p_6\}$ are obtained.

2) SEMANTIC REPRESENTATION OF STAY REGION

To handle this problem, we propose a new semantic representation for stay regions. After the identification process of stay regions, a stay region is represented as a set of spatiotemporal points. Based on the spatiotemporal points in the stay region, a minimum rectangular geographic region is determined which contains all these spatiotemporal points. As shown in Figure 7, there are two diagonal coordinates (x_0, y_0), (x_1, y_1), where x_0 and y_0 are the minimum values of longitude and latitude of all spatiotemporal points in a stay region; x_1 and y_1 are the maximum values of longitude and latitude of them. The semantics of a stay region is defined by a set of tuples (W_i, w_i), where W_i is a word which forms the categories of POIs in the stay region, w_i denotes the number of occurrences of W_i in the stay region.

The categories of POIs are obtained by querying Google Places API. A POI database is composed of a set of POI instances that each contains information such as POI category, latitude and longitude. A POI category label can be regarded as a word set. For example, a POI category ''shopping mall'' can be seen as a word set $\{shopping, mall\}$ with its constituent words ''shopping'' and ''mall''.

An example is given to illustrate the concept of semantics of stay regions. Suppose there is a stay region shown in Figure 7 which includes one shopping mall, three banks and five hotels, the semantics of this stay region can be expressed as $\{(bank, 5), (hotel, 3), (shopping, 1), (center, 1)\}$.

The semantics of stay regions can be regarded as a feature multiset, so the semantic distance between stay regions can be computed by using the feature multiset distance. We employ the distance between *word embeddings* to be the ground distance to measure the semantic distance between two words. We use *Euclidean distance* to be the distance between *word embeddings*. Since *Euclidean distance* is also a distance metric [33], it is easy to prove that the semantic distance between stay regions is a distance metric according to Theorem 1.

3) MODELING INDIVIDUAL MOBILITY PROFILE

Based on the semantic distance between stay regions, stay regions can be clustered and each cluster is called semantic location. By clustering stay regions, each GPS trajectory is regarded as a sequence of semantic locations $a_0 \xrightarrow{\triangle t_1} a_1 \ldots \xrightarrow{\triangle t_m} a_m$, where $\forall 1 \leq i \leq m$, $\triangle t_i = a_i.t_x - a_{i-1}.t_n$, a_i is a semantic location, t_x, t_n are the maximum and minimum timestamps of the spatiotemporal points in the stay region corresponded by a_{i-1} and a_i, $\triangle t_i$ represents the interval time from a_{i-1} to a_i. If $\triangle t_i$ is extremely small, it means that a_{i-1} and a_i should belong to a bigger stay region, but they are identified as two different stay regions. Therefore, if $\triangle t_i$ is less than a certain threshold, a_{i-1} and a_i are merged into a new group. On the contrary, if $\triangle t_i$ is extremely big with respect to a threshold, the semantic trajectory should be split into two subsequences of semantic location. The 3σ-rule is a simple and widely used method for outlier detection. In this paper, we use the 3σ-rule to determine the abnormal threshold of $\triangle t_i$: the abnormal minimum threshold of $\triangle t_i$ is set to μ-3σ, the abnormal maximum threshold of $\triangle t_i$ is set to $\mu + 3\sigma$, where μ is the average value of $\triangle t_i$ and σ is the standard deviation of $\triangle t_i$.

After de-noising by the 3σ-rule, a semantic location sequence can be represented as a symbol sequence. In this way, each user is represented as a multiset of symbol sequences. Then, a sequential pattern mining algorithm is applied to mine frequent patterns from the multiset. These frequent patterns are used to model each user's mobility profile. Each frequent pattern is called one semantic trajectory.

In this paper, mobility profile of a user is modelled by a set of tuples (T_i, w_i), where T_i is a semantic trajectory and w_i is the number of occurrences of T_i.

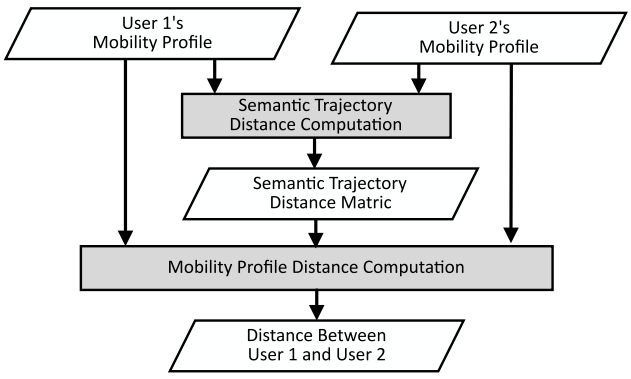


FIGURE 8. The computation procedure of user distance.

B. USER DISTANCE COMPUTATION

As shown in Figure 8, the process of user distance computation includes two steps: semantic trajectory distance computation and mobility profile distance computation. First, the semantic trajectory distance is defined based on longest

common subsequence. A distance matrix of semantic trajectories is generated based on two users' mobility profiles. Then, the distance between mobility profiles is computed by combining the distance matrix of semantic trajectories with the frequency of semantic trajectories, i.e., user distance.

1) SEMANTIC TRAJECTORY DISTANCE COMPUTATION

It is observed that the longer common subsequence of two semantic trajectories, the smaller the distance between them. Therefore, we define the distance between semantic trajectories as follows:

Definition 7 (Semantic Trajectory Distance): The distance between two semantic trajectories T_1 and T_2 is defined as follows:

$$d_t(T_1, T_2) = 1 - \frac{|lcs(T_1, T_2)|}{|T_1| + |T_2| - |lcs(S_1, S_2)|}$$

where $|T|$ represents the length of the semantic trajectory T, $lcs(T_1, T_2)$ represents the longest common subsequence between two semantic trajectories T_1 and T_2.

When T_1 equals to T_2, we have $lcs(T_1, T_2) = 1$, $d_t(T_1, T_2) = 0$. On the contrary, there is no common subsequence between T_1 and T_2, and therefore, we have $lcs(T_1, T_2) = 0, d_t(T_1, T_2) = 1$.

Theorem 2: Given two semantic trajectories T_1 and T_2, their semantic trajectory distance, i.e. $d_t(T_1, T_2) = 1 - \frac{|lcs(T_1,T_2)|}{|T_1|+|T_2|-|lcs(S_1,S_2)|}$, is a distance metric.

Proof: According to Definition 7, we have:

$$\begin{aligned} d_t(T_1, T_2) &= 1 - \frac{|lcs(T_1, T_2)|}{|T_1| + |T_2| - |lcs(T_1, T_2)|} \\ &= \frac{|T_1| + |T_2| - 2|lcs(T_1, T_2)|}{|T_1| + |T_2| - |lcs(T_1, T_2)|} \\ &= \frac{2|T_1| + 2|T_2| - 4|lcs(T_1, T_2)|}{2|T_1| + 2|T_2| - 2|lcs(T_1, T_2)|} \end{aligned} \quad (1)$$

According to [34], the following equation holds:

$$edit(T_1, T_2) = |T_1| + |T_2| - 2|lcs(T_1, T_2)| \quad (2)$$

where $edit(x, y)$ is the string edit distance between x and y. The following is obtained when Eq. (2) is substituted in Eq. (1)

$$d_t(T_1, T_2) = \frac{2edit(T_1, T_2)}{|T_1| + |T_2| + edit(T_1, T_2)} \quad (3)$$

Eq. (3) has been proved to be a normalized distance metric by [35]. Therefore, the semantic trajectory distance is a distance metric.

2) MOBILITY PROFILE DISTANCE COMPUTATION

User's mobility profile can be regarded as a feature multiset, and therefore, the distance between mobility profiles can also be computed by the feature multiset distance. When computing the distance between two mobility profiles, the ground distance is the semantic trajectory distance. Theorem 2 shows that the semantic trajectory distance is a distance metric, and it is easy to prove that the distance between two mobility profiles is also a distance metric according to Theorem 1.

V. EXPERIMENTAL EVALUATION

The proposed method is called as Distance Metric (DIM), and we evaluated this method based on both synthetic and real-life datasets. In the following, we compared our method with MSTP [27], MTP [28], CPS [29] and PDS [16]. In order to evaluate these methods on distance metric axioms, we compute the dissimilarity values by the sum of values of similarity and dissimilarity equaling to 1, because such computation will not change the properties of distance metric [36]. The experiment enviorenment is as follows:

- Programming Language: Python 2.7
- CPU: four Intel Xeon Processor E7-4830 v1 with 8 cores 2.13GHz
- Memory: 64GB
- Operation System: Ubuntu 16.04 LTS

A. EVALUATION ON SYNTHETIC DATASET

A synthetic dataset shown below is constructed, which contains seven users' mobility profiles, with the first four being from literature [29]. Based on the dataset, the four distance metric axioms in Definition 6 are used to evaluate existing methods and demonstrate their insufficiencies.

$M_1 = \{(a, 2), (b, 2), (c, 4), (ab, 2)\}$;
$M_2 = \{(a, 1), (b, 1), (c, 4), (ab, 3)\}$;
$M_3 = \{(a, 1), (b, 1), (c, 4), (ba, 3)\}$;
$M_4 = \{(a, 1), (d, 1), (c, 4), (ad, 3)\}$;
$M_5 = \{(a, 1)\}$; $M_6 = \{(a, 1), (b, 3)\}$;
$M_7 = \{(a, 10)\}$

A distance matrix between mobility profiles produced by different methods is shown in Table 1. Let d_{ij} be the matrix element with row index i and column index j. Based on Table 1, we have the following observations: (1) for MSTP, $d_{11} \neq 0$ and $d_{76} + d_{75} < d_{65}$ indicate that MSTP cannot guarantee the reflexivity and the triangle inequality; (2) for MTP, $d_{12} + d_{13} < d_{23}$ indicates that MTP cannot guarantee the triangle inequality; (3) for CPS, $d_{12} + d_{13} < d_{23}$ indicates that CPS cannot guarantee the triangle inequality; (4) for PDS, $d_{13} \neq d_{31}$ and $d_{12} + d_{13} < d_{23}$ indicate that PDS cannot guarantee the symmetry and the triangle inequality.

The performance of these methods on the distance metric axioms is shown in Table 2. It can be found that DIM hold all distance metric axioms while other methods only hold some of these axioms.

B. EVALUATIONS ON REAL-LIFE DATASET

We conducted a series of experiments over a real-life dataset. For each user, the dissimilarity scores with other users are computed based on different methods. The top K nearest neighbors (KNNs) are predicted for each user and ranked according to dissimilarity in an increasing order, based on which we observed the rank of the nearest neighbor in the prediction results. In addition, the performance of different methods is evaluated using the two well-known criteria, i.e., Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG), borrowed from information retrieval.

TABLE 1. Comparison on the distance metric axioms.

	MSTP						
	M_1	M_2	M_3	M_4	M_5	M_6	M_7
M_1	0.5	0.5	0.6	0.6	0.75	0.73	0.69
M_2	0.5	0.5	0.61	0.61	0.7	0.7	0.68
M_3	0.6	0.61	0.5	0.61	0.7	0.7	0.68
M_4	0.6	0.61	0.61	0.5	0.7	0.88	0.68
M_5	0.75	0.7	0.7	0.7	0	0.67	0
M_6	0.73	0.7	0.7	0.88	0.67	0.5	0.54
M_7	0.69	0.68	0.68	0.68	0	0.54	0
	MTP						
	M_1	M_2	M_3	M_4	M_5	M_6	M_7
M_1	0	0	0.07	0.24	0.32	0.19	0.29
M_2	0	0	0.08	0.21	0.32	0.21	0.29
M_3	0.07	0.08	0	0.21	0.32	0.21	0.29
M_4	0.24	0.21	0.21	0	0.32	0.65	0.29
M_5	0.32	0.32	0.32	0.32	0	0.33	0
M_6	0.19	0.21	0.21	0.65	0.33	0	0.25
M_7	0.29	0.29	0.29	0.29	0	0.25	0
	CPS						
	M_1	M_2	M_3	M_4	M_5	M_6	M_7
M_1	0	0.03	0.29	0.53	0.8	0.53	0.71
M_2	0.03	0	0.33	0.56	0.81	0.56	0.73
M_3	0.29	0.33	0	0.56	0.81	0.56	0.73
M_4	0.53	0.56	0.56	0	0.81	0.91	0.73
M_5	0.8	0.81	0.81	0.81	0	0.5	0.82
M_6	0.53	0.56	0.56	0.91	0.5	0	0.91
M_7	0.71	0.73	0.73	0.73	0.82	0.91	0
	PDS						
	M_1	M_2	M_3	M_4	M_5	M_6	M_7
M_1	0	0	0.25	0.5	0.75	0.5	0.75
M_2	0	0	0.33	0.56	0.78	0.56	0.78
M_3	0.33	0.33	0	0.56	0.78	0.56	0.78
M_4	0.56	0.56	0.56	0	0.78	0.78	0.78
M_5	0	0	0	0	0	0	0
M_6	0	0	0	0.75	0.75	0	0.75
M_7	0	0	0	0	0	0	0
	DIM						
	M_1	M_2	M_3	M_4	M_5	M_6	M_7
M_1	0	0.11	0.24	0.38	0.7	0.5	0.7
M_2	0.11	0	0.22	0.33	0.72	0.61	0.72
M_3	0.24	0.22	0	0.33	0.72	0.61	0.72
M_4	0.38	0.33	0.33	0	0.72	0.82	0.72
M_5	0.7	0.72	0.72	0.72	0	0.75	0
M_6	0.5	0.61	0.61	0.82	0.75	0	0.75
M_7	0.7	0.72	0.72	0.72	0	0.75	0

TABLE 2. Comparison on the distance metric axioms.

	Reflexivity	Non-negativity	Symmetry	Triangle inequality
MSTP	×	√	√	×
MTP	√	√	√	×
CPS	√	√	√	×
PDS	√	√	×	×
DIM	√	√	√	√

Legend: the axiom holds (√); the axiom does not hold (×).

1) DATASET DESCRIPTION

The dataset is collected from the Geolife [37] project of Microsoft Research Asia, which includes 182 users' 18,670 GPS trajectories over five years (2007-2012). Each user has some trajectory files that each file represents a GPS trajectory. We selected the top 100 users with large number of trajectories as some users in the dataset have fewer check-ins. To evaluate the performance of each method in top K nearest neighbor prediction, the ground truth is required. For this purpose, we divide the trajectories of each user into two parts evenly and regard them as trajectories of two different users. After de-noising by the 3σ-rule in Section IV-A.3, the trajectories of a user u can be represented as a single sequence $s_0 \rightarrow s_1 \ldots \rightarrow s_m$ of semantic locations that are fully ordered. A timestamp t can be determined to divide the sequence of semantic locations into two even parts. All the trajectories with check-in time less than t are assigned to $u^{\#}$. Similarly, all the trajectories with check-in time greater than t are assigned to u^*. Because $u^{\#}$ and u^* are generated from the trajectories of u, they are defined to be the nearest neighbors to each other. Finally, 200 users are obtained. The detailed description of the generated working dataset is shown in Table 3.

TABLE 3. The detailed description of the working dataset.

Users	Trajectories	Check-ins	Distance(km)	Duration(h)
200	17,942	23,711,587	1,246,820	47,362

TABLE 4. Parameter settings.

Parameter	δD	δT	δV	sup_{min}	d_c	k
Value	200m	30min	0.5m/s	0.3	0.5%	35

2) EVALUATION CRITERIA

MAP and NDCG are applied to evaluate the performance of our method. MAP is a commonly used evaluation criterion in the field of information retrieval. The definition of MAP is given as follows:

$$MAP = \sum_{i=1}^{N} (s_i)/N \tag{4}$$

where $N = 200$ is the number of all users in the working dataset, and s_i is the score of predicting K nearest neighbors of user i. The definition of s_i is given as follows:

$$s_i = \begin{cases} \frac{1}{r_i} & r_i \leq K \\ 0 & r_i > K \end{cases}$$

where r_i is the rank of the nearest neighbor of user i in the predicted results.

The definition of nDCG@K is the same as Eq. (4), and the main difference is in the form of s_i. For this experiment, the definition of s_i is as follows:

$$s_i = \begin{cases} \frac{1}{log_2(r_i+1)} & r_i \leq K \\ 0 & r_i > K \end{cases}$$

3) PARAMETER SETTINGS AND ANALYSIS

Some parameters used in our method are set as shown in Table 4.

δD, δT and δV are the distance, time and velocity thresholds in the stay region identification algorithm. sup_{min} is

the minimum support threshold for PrefixSpan algorithm. With loss of generality, we assign δD, δT and sup_{min} the same values as those in [16]. The Density Peaks Clustering (DPC) [38] is applied to cluster stay regions based on the semantic distance between stay regions. Note that d_c is the percentage of average neighbors of each data point in DPC algorithm, and k is the final clustering number of DPC algorithm.

For each data point i, the DPC computes two quantities: its local density ρ_i and its distance δ_i from points of higher density. All data points are drawn in a decision graph as shown in Figure 9. The cluster centers are recognized as points for which the two values are relatively large.

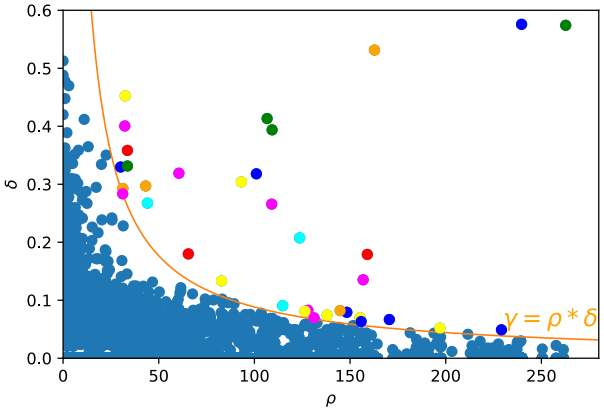


FIGURE 9. Decision graph.

The DPC clusters data points by truncating a cutoff distance d_c. When the value of d_c becomes small or large, the cluster centers are concentrated on the upper left or lower right angles of the decision graph respectively. According to the reference range of d_c in DPC and the distribution of data points in the decision graph, d_c is set to 0.5%. In addition, the cluster centers are decided by the product of two quantities $\gamma = \rho * \delta$ and sort data points in a decreasing order according to the value of γ. The first k data points with larger γ are taken as cluster centers. If the value of γ of the kth data point is γ_k, all data points above a curve $\gamma_k = \rho * \delta$ in the decision graph are cluster centers. By varying the value of k, we observe the evaluation results of DIM algorithm at nDCG@K (the results at MAP are similar). Figure 10 shows that, with the increase of the number k of cluster centers, the result at nDCG@K becomes higher. When the value of k is bigger than 35, the result at nDCG@K remains basically unchanged. Therefore, the number of cluster centers k is finally set to 35.

4) EXPERIMENT RESULTS

Figures 11 (a) and (b) show the obtained MAP and nDCG@K scores using different methods which predict the top K nearest neighbors (KNNs). A higher score normally indicates an accurate predicted result. The horizontal axis is the number K of predicted nearest neighbors. It can be observed that DIM

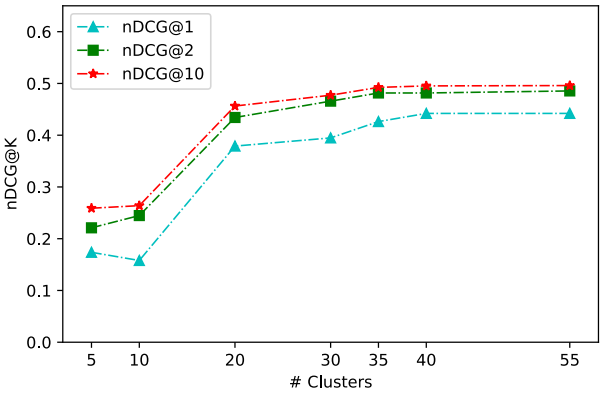


FIGURE 10. nDCG@K vs the number of clusters.

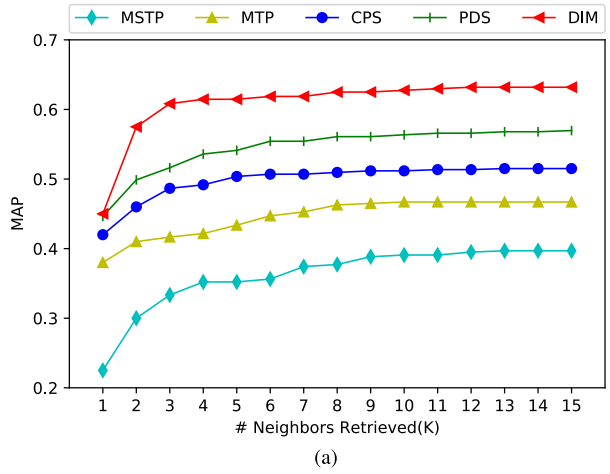


FIGURE 11. Comparison of KNNs prediction. (a) MAP. (b) nDCG@K.

achieves significantly higher scores than existing methods with respect to our evaluation criteria.

To evaluate our proposed stay region identification method, existing methods are improved by using our stay region identification method. Similarly, in this experiment the K nearest neighbors are predicted for each user by different methods. The performance of methods is evaluated by MAP and nDCG@K respectively. The experimental results are shown in Figure 12 where methods named with ".V"

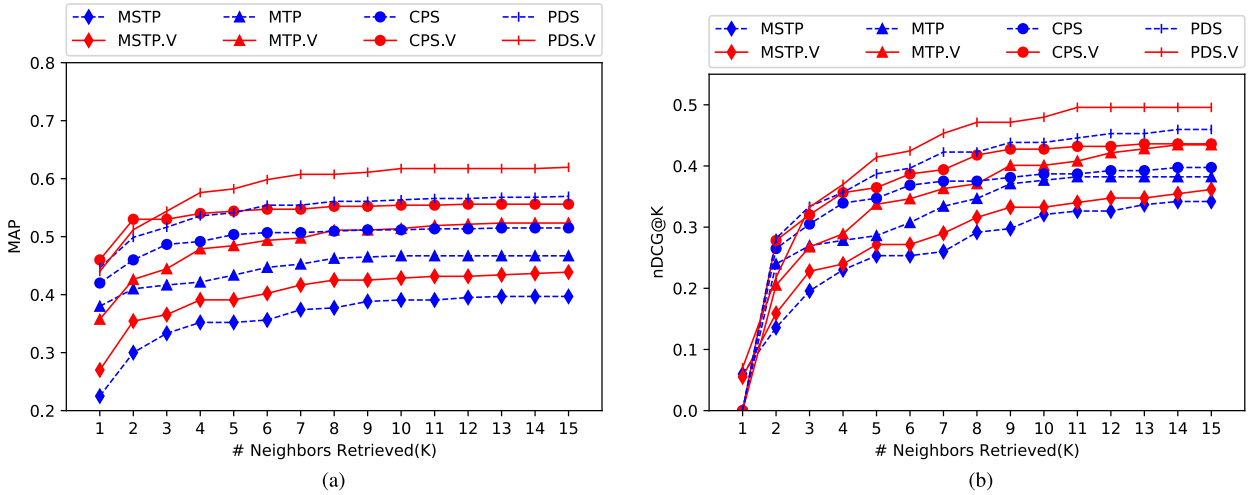


FIGURE 12. The performance demonstration of the proposed identification method of stay regions. (a) MAP. (b) nDCG@K.

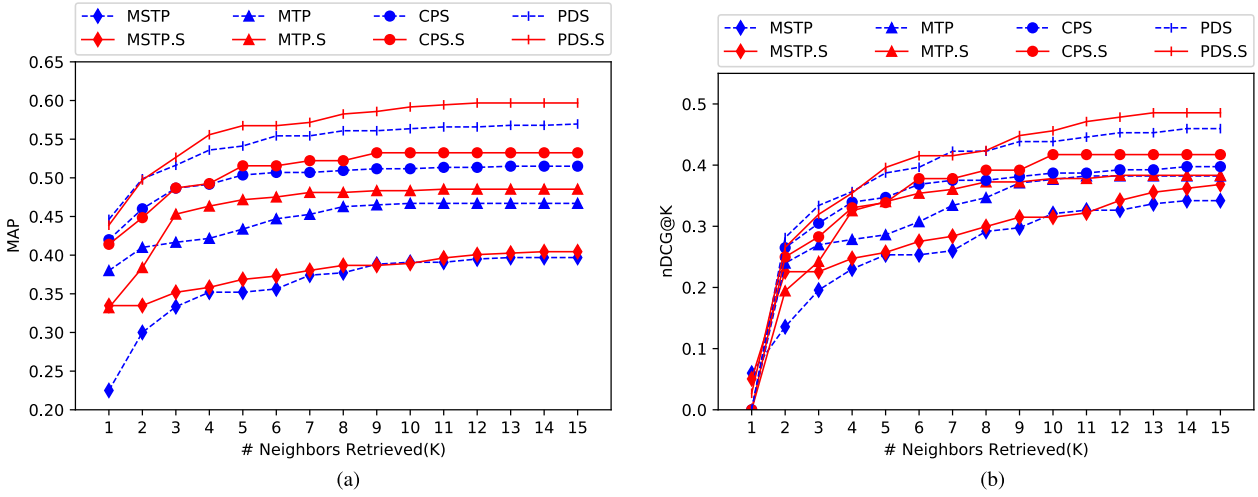


FIGURE 13. The performance demonstration of the proposed semantic representation method of stay regions. (a) MAP. (b) nDCG@K.

suffix are improved methods. It is observed that, all improved methods achieve performance improvement, which indicates the effectiveness of the proposed stay region identification method.

Similarly, existing methods are improved by our proposed stay region semantic representation method. The experimental results are shown in Figure 13. Methods named with ".S" suffix are improved. It can be found that the performance of the improved methods is slightly worse than existing methods when the retrieved neighbor number is small. When the retrieved neighbor number increases, the performance of the improved methods becomes better than existing methods.

C. DISCUSSION

This work contains the following three main contributions: (1) a new indoor stay region identification method considering velocity; (2) a new semantic representation of stay regions based on word embedding; and (3) a new user distance metric satisfying metric axioms. Detailed comparisons with existing works are summarized in Table 5.

Specifically, only [16] and this work support the identification of indoor stay regions. The experiment results show that our identification method is more accurate compared to [16]. In addition, most of existing works support the semantic representation of stay regions. Different from them, we introduce word embedding technique into this field for the first time. And it brings better experiment results in most cases as shown in Section V-B.4. Finally, our proposed user distance metric satisfies all the distance metric axioms while existing works can only satisfy some of them. In particular, only our method holds the triangle inequality. As a result, for applications that rely on triangular inequalities, our method will be the only choice.

TABLE 5. The summary of experiment results.

Literature	Mobility profile construction			Mobility profile distance computation	
	Indoor stay region identification supported	Stay regions' semantics supported	Techniques used	Distance metric axioms satisfied	Techniques used
Ying et. al. 2010 [27]	Not involved	Yes	Frequent sequential pattern mining, Maximal semantic trajectory pattern	Partially satisfied (Refer to Table 2 for details)	Longest common subsequence, Consider participation ratio, the support or TF-IDF of maximal semantic trajectory pattern
Chen et. al. 2013 [28]	Not supported	No	Trajectory pattern, Spatiotemporal containment, Hierarchical clustering		Maximal sequence patterns, Consider the length, support of sequence patterns
Chen et. al. 2014 [29]	Not supported	Yes	Probabilistic pattern mining, Common pattern		Consider the length, support of common patterns and Bray-Curtis similarity of support values
Mazumdar et. al. 2016 [16]	Fairly supported	Yes	Frequent sequential pattern mining, Significance score, Common pattern		Consider the length, support, distribution of common patterns and the distribution of check-ins
This work	Better supported	Yes	Frequent sequential pattern mining, Word embedding, Multiset, Density peaks clustering	Satisfied	Longest common subsequence, Balanced transportation problem

VI. CONCLUSIONS

This paper proposes a new user distance computation method for GPS trajectories which hold the distance metric axioms. The proposed method first improves the accuracy of indoor stay region identifications. Subsequently, a new method to semantically represent a stay region is proposed. In this way, the semantic distance between two stay regions is defined by word embedding techniques. Finally, the distance between user mobility profiles is computed based on the solution of a balanced transportation problem. The effectiveness of the proposed method is evaluated based on both synthetic and real-life datasets. In the future, we plan to apply the method to the location-based friend recommendation and other personalized recommendation services.

REFERENCES

[1] R. Wu, G. Luo, Q. Yang, and J. Shao, "Learning individual moving preference and social interaction for location prediction," *IEEE Access*, vol. 6, pp. 10675–10687, 2018.

[2] Q. Peng, M. Zhou, Q. He, Y. Xia, C. Wu, and S. Deng, "Multi-objective optimization for location prediction of mobile devices in sensor-based applications," *IEEE Access*, vol. 6, pp. 77123–77132, 2018.

[3] J.-D. Zhang and C.-Y. Chow, "TICRec: A probabilistic framework to utilize temporal influence correlations for time-aware location recommendations," *IEEE Trans. Serv. Comput.*, vol. 9, no. 4, pp. 633–646, Jul./Aug. 2016.

[4] H. Gao, J. Tang, and H. Liu, "Addressing the cold-start problem in location recommendation using geo-social correlations," *Data Mining Knowl. Discovery*, vol. 29, no. 2, pp. 299–323, 2015.

[5] A. Bhapkar, K. Fegade, R. Ahire, C. Chaudhary, and A. M. Jagtap, "A better semantic based friend recommendation system for modern social networks," *Int. J. Comput. Appl.*, vol. 156, no. 8, pp. 1–5, 2016.

[6] X. Qiao *et al.*, "Recommending nearby strangers instantly based on similar check-in behaviors," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 3, pp. 1114–1124, Jul. 2015.

[7] C. Xu, L. Zhu, Y. Liu, J. Guan, and S. Yu, "DP-LTOD: Differential privacy latent trajectory community discovering services over location-based social networks," *IEEE Trans. Serv. Comput.*, to be published. doi: 10.1109/TSC.2018.2855740.

[8] W.-Y. Zhu, W.-C. Peng, C.-C. Hung, P.-R. Lei, and L.-J. Chen, "Exploring sequential probability tree for movement-based community discovery," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 11, pp. 2717–2730, Nov. 2014.

[9] M. Fire, L. Tenenboim-Chekina, R. Puzis, O. Lesser, L. Rokach, and Y. Elovici, "Computationally efficient link prediction in a variety of social networks," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, p. 10, 2013.

[10] J. Laub and K.-R. Müller, "Feature discovery in non-metric pairwise data," *J. Mach. Learn. Res.*, vol. 5, pp. 801–818, Dec. 2004.

[11] P. Ciaccia, M. Patella, and P. Zezula, "M-tree: An E cient access method for similarity search in metric spaces," in *Proc. 23rd VLDB Conf.*, Athens, Greece, 1997, pp. 426–435.

[12] L.-A. Gottlieb, A. Kontorovich, and R. Krauthgamer, "Efficient classification for metric data," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5750–5759, Sep. 2014.

[13] M. Hein, O. Bousquet, and B. Schölkopf, "Maximal margin classification for metric spaces," *J. Comput. Syst. Sci.*, vol. 71, no. 3, pp. 333–359, 2005.

[14] I. T. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer, 2011, pp. 1094–1096.

[15] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[16] P. Mazumdar, B. K. Patra, R. Lock, and S. B. Korra, "An approach to compute user similarity for GPS applications," *Knowl.-Based Syst.*, vol. 113, pp. 125–142, Dec. 2016.

[17] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring social ties between users with human location history," *J. Ambient Intell. Humanized Comput.*, vol. 5, no. 1, pp. 3–19, 2014.

[18] M. Yoshida, T. Iizuka, H. Shiohara, and M. Ishiguro, "Mining sequential patterns including time intervals," *Proc. SPIE*, vol. 4057, pp. 213–220, Apr. 2000. doi: 10.1117/12.381735.

[19] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli, "Mining sequences with temporal annotations," in *Proc. ACM Symp. Appl. Comput.*, Dijon, France, 2006, pp. 593–597.

[20] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Jose, CA, USA, 2007, pp. 330–339.

[21] M. Lv, L. Chen, and G. Chen, "Mining user similarity based on routine activities," *Inf. Sci.*, vol. 236, pp. 17–32, Jul. 2013.

[22] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining user similarity based on location history," in *Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Irvine, CA, USA, 2008, pp. 1–10.

[23] L. O. Alvares, V. Bogorny, B. Kuijpers, B. Moelans, J. A. de Macedo, and A. T. Palma, "Towards semantic trajectory knowledge discovery," Hasselt Univ., Limbourg, Belgium, Tech. Rep., Oct. 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.9912&rep=rep1&type=pdf

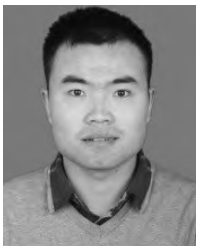
[24] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Finding similar users using category-based location history," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, San Jose, CA, USA, 2010, pp. 442–445.
[25] T. Horozov, N. Narasimhan, and V. Vasudevan, "Using location for personalized POI recommendations in mobile environments," in *Proc. Int. Symp. Appl. Internet*, Phoenix, AZ, USA, Jan. 2006, pp. 124–129.
[26] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history," *ACM Trans. Web*, vol. 5, no. 1, p. 5, 2011.
[27] J. J.-C. Ying, E. H.-C. Lu, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Mining user similarity from semantic trajectories," in *Proc. 2nd ACM SIGSPATIAL Int. Workshop Location Based Social Netw.*, San Jose, CA, USA, 2010, pp. 19–26.
[28] X. Chen, J. Pang, and R. Xue, "Constructing and comparing user mobility profiles for location-based services," in *Proc. 28th Annu. ACM Symp. Appl. Comput.*, Coimbra, Portugal, 2013, pp. 261–266.
[29] X. Chen, R. Lu, X. Ma, and J. Pang, "Measuring user similarity with trajectory patterns: Principles and new metrics," in *Proc. 16th Asia–Pacific Web Conf.*, Changsha, China: Springer, 2014, pp. 437–448.
[30] C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt, "Learning semantic similarity for very short texts," in *Proc. IEEE 15th Int. Conf. Data Mining Workshop (ICDMW)*, Atlantic City, NJ, USA, Nov. 2015, pp. 1229–1234.
[31] K. Bernhard and V. Jens, *Combinatorial Optimization: Theory and Algorithms*. Berlin, Germany: Springer, 2018, pp. 226–227.
[32] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[33] V. Kumar, J. K. Chhabra, and D. Kumar, "Performance evaluation of distance metrics in the clustering algorithms," in *Proc. INFOCOMP*, vol. 13, no. 1, 2014, pp. 38–52.
[34] A. M. Bronstein, M. M. Bronstein, A. M. Bruckstein, and R. Kimmel, "Partial similarity of objects, or how to compare a centaur to a horse," *Int. J. Comput. Vis.*, vol. 84, no. 2, p. 163, 2009.
[35] C. de la Higuera and L. Micó, "A contextual normalised edit distance," in *Proc. IEEE 24th Int. Conf. Data Eng. Workshop (ICDEW)*, Cancún, Mexico, Apr. 2008, pp. 354–361.
[36] S. Chen, B. Ma, and K. Zhang, "On the similarity metric and the distance metric," *Theor. Comput. Sci.*, vol. 410, nos. 24–25, pp. 2365–2376, 2009.
[37] Y. Zheng, X. Xie, and W.-Y. Ma, "GeoLife: A collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, Jun. 2010.
[38] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

ZEDONG LIN received the Ph.D. degree in software engineering from the Shandong University of Science and Technology, Qingdao, China, in 2018.

His research interests include data mining, machine learning, and mobile computing.

QINGTIAN ZENG received the Ph.D. degree in computer software and theory from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He was a Visiting Professor with the City University of Hong Kong, Hong Kong, in 2008. He is currently a Professor with the Shandong University of Science and Technology, Qingdao, China. His current research interests include Petri nets, process mining, and knowledge management.

HUA DUAN received the B.S. and M.S. degrees in applied mathematics from the Shandong University of Science and Technology, Tai'an, China, in 1999 and 2002, and the Ph.D. degree in applied mathematics from Shanghai Jiao Tong University, in 2008. She is currently an Associate Professor with the Shandong University of Science and Technology. Her research interests include process mining and machine learning.



CONG LIU (S'18) received the B.S. and M.S. degrees in computer software and theory from the Shandong University of Science and Technology, Qingdao, China, in 2013 and 2015, respectively. His research interests include business process management, process mining, and Petri nets.

FAMING LU received the Ph.D. degree in computer software and theory from the Shandong University of Science and Technology, Qingdao, China, in 2013, where he is currently an Associate Professor. He has published more than 30 papers in academic journals. His research interests include Petri nets and process mining.