DOI：10. 13471/j. cnki. acta. snus. ZR20250097

耦合 Word2Vec 和动态语义地图的车辆轨迹相似性度量*

黄敏[1]，梁宁晨[1]，王晓聪[1]，陈淋[1]，陈开颖[2]

1. 中山大学智能工程学院 / 广东省智能交通系统重点实验室，广东 深圳 518107
2. 佳都科技集团股份有限公司，广东 广州 510665

摘　要：提出一种耦合改进 Word2Vec 方法和动态语义地图的轨迹相似性度量方法。借助增加目的地约束的 Word2Vec 模型学习卡口序列关联关系，并与目的地建立显式联系；同时，动态语义地图可以作为时间和出行行为维度的相似性度量方法构建基础。实验结果表明，城市功能区在一天之中呈现出显著的动态变化特征。并且，在轨迹层次聚类任务中，本文方法的平均 AC 值较对比方法降低了 0.36，体现出其更强的相似性度量能力与稳健性。

关键词：轨迹相似性；目的地预测；Word2Vec；动态语义地图

中图分类号：P208　　**文献标志码：**A　　**文章编号：**2097 - 0137(2025)06 - 0076 - 10

Vehicle trajectory similarity measures for coupling Word2Vec and dynamic semantic map

HUANG Min[1], *LIANG Ningchen*[1], *WANG Xiaocong*[1], *CHEN Lin*[1], *CHEN Kaiying*[2]

1. *School of Intelligent Systems Engineering, Sun Yat-sen University / Key Laboratory of Intelligent Transportation System in Guangdong Province, Shenzhen 518107, China*
2. *Jiadu Technology Group Company Limited, Guangzhou 510665, China*

Abstract: A trajectory similarity measure combining improved Word2Vec method and dynamic semantic map is proposed. The method learns the bayonet sequence correlation by the Word2Vec model with added destination constraint and establishes explicit connection with the destination; on the other hand, the dynamic semantic map can build the basis for the similarity measure of time and travel behavior dimension. The experimental results indicate that urban functional zones exhibit significant dynamic changes throughout the day. In the trajectory hierarchical clustering task, the average AC value of the method in this paper is 0.36 lower than that of the method chosen for comparison, further demonstrating its superior similarity measurement capability and robustness.

Key words: trajectory similarity; destination prediction; Word2Vec; dynamic semantic map

随着通信和定位技术的发展，可以便捷地获取和存储海量的移动物体轨迹数据，作为疫情防控（Schlosser et al.，2021）、交通规划（Cao et al.，2022；Du et al.，2019）等方面的研究基础。但目的地预测

* **收稿日期：**2025-06-05　　**录用日期：**2025-07-05　　**网络首发日期：**2025-09-23
基金项目：广东省自然科学基金（2023A1515240046）
作者简介：黄敏（1975年生），女；**研究方向：**可计算路网；E-mail：Huangm7@mail.sysu.edu.cn

增强出版　ZR20250097　　全文阅读　ZR20250097

任务中面临着一个亟待解决的问题，即轨迹样本稀疏（晋广印等，2024；江婧等，2019）。为了应对这一挑战，研究者通过子轨迹合成（Xue et al.，2013）、前后缀轨迹分段（余丹青等，2023）等方法扩充轨迹样本数量。而轨迹相似性度量能够正确判断轨迹的相似性，在完备轨迹的基础上挖掘轨迹的潜在移动模式（Xia et al.，2011；吴晨昊等，2023），扩充样本并应用于交通预测的研究中。

近年来，自然语言处理模型在相似性任务中的成功应用，为轨迹相似性度量提供了新的思路。李威（2018）提出将车辆轨迹看作“文本”，将轨迹位置看作组成轨迹的“词”，利用 Word2Vec 方法进行建模。Kang et al.（2021）利用 Doc2Vec 模型将路段序列转化为固定维数的轨迹段特征向量，提高了不同长度轨迹段之间相似度的计算效率。罗月童等（2021）利用 Word2Vec 方法将道路卡口监控过车数据进行轨迹卡口映射。这些方法基于车辆出行模式，识别不同路线轨迹点之间的关联性，并将它们映射到相似的向量空间位置，从而更适用于大规模数据挖掘应用，但它们大多着眼于空间维度（Li et al.，2018）和时空维度（Shang et al.，2018）。

地理位置语义（Furtado et al.，2016）推动了车辆出行行为在轨迹相似性度量中的应用。在实践中，从 POI 数据提取的功能区用地性质（郑国强等，2023；Gao et al.，2017）可以作为轨迹点的地理位置语义。然而，地理位置语义受人类活动影响所体现的功能性是动态变化的，需要考虑合理的时间窗划分方式。因此，本文提出一种耦合改进 Word2Vec 和动态语义地图的车辆轨迹相似性度量方法，从空间、时间和车辆出行行为三个维度进行轨迹相似性度量。并基于宣城市的车辆出行轨迹数据，验证方法面向目的地预测任务中稀疏数据场景的可行性。

1　方　法

耦合改进 Word2Vec 和动态语义地图的车辆轨迹相似性度量方法，如图 1 所示。其基本过程是：首先，基于人类出行的交通流量变化趋势，划分自适应时间窗；然后，根据 POI 属性识别不同时段的功能区性质，构建动态语义地图用于识别卡口节点的地理位置动态语义；接着，导入轨迹数据的空间序列、时间戳和地理位置动态语义属性，构建基于空间、时间和车辆出行行为三个维度的轨迹相似性度量方法；最后，依据轨迹数据集进行轨迹层次聚类，验证方法的有效性和不同降采样率下的稳健性。

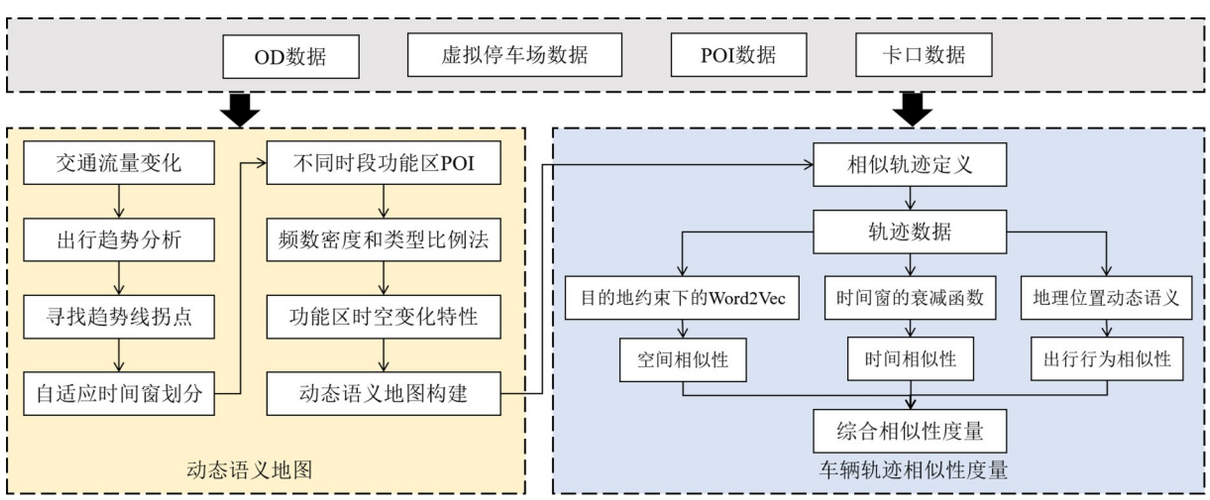

图 1　本文方法流程

Fig. 1　Flowchart of the method proposed in this paper

1.1　相关定义

相关定义如下：（1）卡口节点是指在实际道路中对车辆个体进行识别的监控设备位置点。（2）轨迹 OD 点指轨迹的起始卡口节点和到达卡口节点。（3）卡口序列轨迹。车辆在行驶过程中经过一系列卡口节点，按时间顺序对卡口进行排序，得到卡口序列轨迹。因此，轨迹序列中每个轨迹点都是卡口节点。（4）相似卡口序列轨迹。对于两条不同的卡口序列轨迹，若被判定为相似轨迹，应该具有以下 3 个特征：其一，同一方向的两条轨迹通过了不同路径，但两者的卡口序列有较强关联关系，并且到达地点一致或接近；其二，两条轨迹在同一活动模式的时间范围内，即起始时间、到达时间及行程时间相近；其三，两条轨迹出发点和到达点的地理位置

语义应当相似。

1.2 动态语义地图

1.2.1 自适应时间窗划分 地理位置动态语义的变化与人类活动息息相关(Gui et al.,2021),因此根据交通流变化时间划分动态语义的变化点,得到不同时间窗 Windows_k(k表示时间窗的编号)的动态语义地图。

对交通流随时间的变化图进行LOESS平滑拟合,可以构建表示居民活动的趋势线。LOESS是一种基于局部回归分析的非参数方法,通过对复杂的波动数据设定带宽因子f,根据总样本数计算得到局部拟合的窗口长度,对每个窗口内的样本x_h分配权重。本文使用高斯权重函数

$$\omega(x_h)=\exp\left[-\frac{(x_h-x'_h)^2}{2f^2}\right],$$

其中$\omega(x_h)$表示当前目标点x_h的权重,x'_h表示训练数据中的其他点。利用加权最小二乘法为每一点x_h局部拟合一个多项式函数,最后将拟合点连接在一起,即为LOESS回归曲线,其具有连续性。对于一个连续函数$f(x)$,若其坐标点前后满足二阶导数正负变化,则该坐标点称为函数的拐点。根据此性质判断连续函数LOESS回归曲线的拐点。之后,根据拐点之间的时间间隔自适应地划分时间窗,将时间窗划分为以拐点为分割点的窗口。

1.2.2 城市区域单元 轨迹点的地理位置语义依赖于所属功能区用地性质,因此需要对城市区域单元进行划分。以城市路网为基础单元是常见的划分方式(周杭等,2022),且本文所涉及的轨迹点都为卡口位置的道路节点,因此引入虚拟停车场(余志等,2022)概念。如图2所示,虚拟停车场是一个由不重复的节点g_i和有向边$e_i^{i'}$所围成的区域。其中,g_i表示第i个卡口节点,$e_i^{i'}$由起始节点g_i和到达节点$g_{i'}$连成。

1.2.3 POI用地类型划分 基于POI数据可以对城市功能区进行定量识别(池娇等,2016)。结合宣城市的实际情况,POI数据的用地类型可以分为居住用地、公共用地、商业用地、绿地广场用地、工业用地和交通设施用地共6大类,对其进行编号为后续识别功能区的用地性质提供必要信息。功能区用地性质的变化取决于POI属性的开放与否,研究区域POI开放时间点如增强出版:附表1所示。

1.2.4 动态语义地图构建 构建动态语义地图的本质是对地图中各个功能区的构建。如图2所示,根据划分好的自适应时间窗,得到对应时间窗 Windows_k 开放的POI。利用POI数据的经纬度,在空间上匹配虚拟停车场p,并结合时空的POI数据得到功能区G_p及其用地性质,最后得到功能区的属性。在空间上,功能区G_p是一个由虚拟停车场p表示的区域,其中包含道路卡口节点g_i。在动态语义上,功能区G_p中不同时间窗 Windows_k 对应不同的POI,映射得到功能区用地性质,用于表示轨迹点在该时间窗下的地理位置语义。

根据功能区G_p中的POI,通过频数密度和类型比例得到G_p用地性质。频数密度

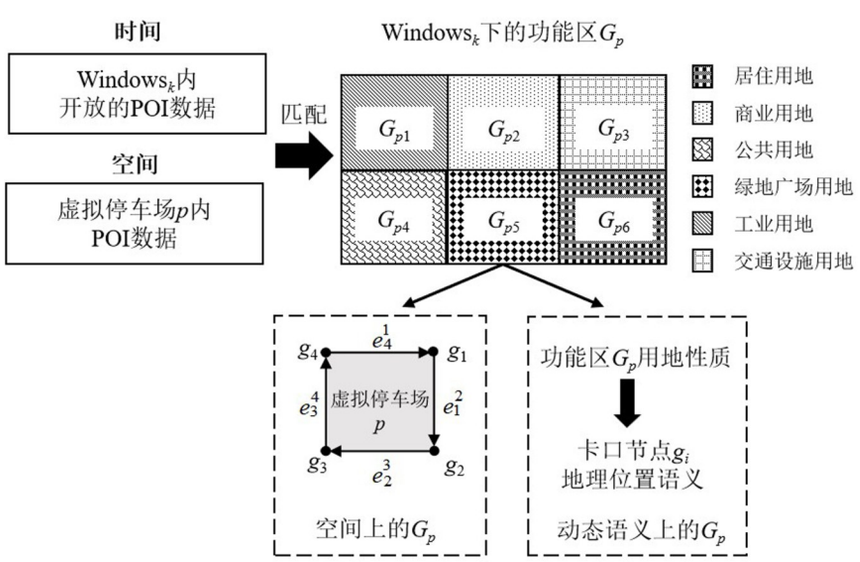

图2 功能区构建及属性

Fig. 2 Functional area construction and attributes

$$F_p^v = \frac{n_p^v}{\sum_p n_p^v},$$

其中 v 表示用地类型，$v = 1, 2, \cdots, 6$；n_p^v 表示功能区 G_p 内第 v 种用地类型数量；F_p^v 表示第 v 种用地类型与该用地类型的频数总数之比。

第 v 种用地类型的频数密度占功能区 G_p 内所有用地类型频数密度的比例，即类型比例

$$C_p^v = \frac{F_p^v}{\sum_{v=1}^{6} F_p^v} \times 100\%.$$

Li et al.（2022）确定类型比例值为 50% 是判断功能区 G_p 用地性质的标准。当 G_p 内某一种用地类型比例 C_p^v 占到 50% 及以上时，即确定该 G_p 为单一功能区。当 G_p 内所有用地类型比例 C_p^v 均没有达到 50% 时，则该功能区 G_p 为混合功能区。

受 POI 变化的影响，不同时间窗中功能区 G_p 的用地性质有所不同。因此，在第 k 个时间窗 Windows_k，功能区总和 $\bigcup_p G_p$ 结合各个功能区 G_p 的用地性质，组成第 k 个语义地图 $(\bigcup_p G_p)_k$。汇总 k 个语义地图，得到随时间窗变化的动态语义地图

$$\text{MAP}_{\text{dynamic}} = \left\{(\bigcup_p G_p)_1, (\bigcup_p G_p)_2, \cdots, (\bigcup_p G_p)_k, \cdots\right\}.$$

1.3　轨迹相似性度量

轨迹序列中，每个轨迹的卡口节点 g_i 都有对应的虚拟停车场 p。轨迹序列 T 以按时间顺序经过的 g_i 表示，即 $T = \{g_1, g_2, \cdots, g_i, \cdots\}$。$g_i$ 包括了对应卡口地理位置 (x_i, y_i)（即经度和纬度）、经过该卡口的时间戳 t_i。

对于任意 t_i 时刻的卡口节点 g_i，可以在动态语义地图中查询到对应 Windows_k 的功能区 G_p，然后得到 G_p 用地性质，用于表示 g_i 的地理位置语义。于是，g_i 拥有了第三个属性，即地理位置动态语义属性。本文以向量形式描述 g_i 的地理位置动态语义属性，即

$$\overrightarrow{\text{BSim}_p^{(t_i)}} = \left\{C_p^1, C_p^2, \cdots, C_p^6\right\}, \quad C_p^v \in [0, 1],$$

其中 p 表示 g_i 所属的虚拟停车场编号，t_i 表示卡口节点 g_i 对应的时间戳。

对单一和混合功能区情形下的地理位置动态语义属性 $\overrightarrow{\text{BSim}_p^{(t_i)}}$ 进行区分。若为单一功能区，则 g_i 的地理位置动态语义属性 $\overrightarrow{\text{BSim}_p^{(t_i)}}$ 由独热向量表达，即占比最大的用地类型比例 C_p^v 索引值为 1，其余位置的值都为 0。

例如，当功能区 G_p 居住用地占比最大且超过 50% 时，则 $\overrightarrow{\text{BSim}_p^{(t_i)}} = \{1, 0, 0, 0, 0, 0\}$。若功能区 G_p 为混合功能区，则 $\overrightarrow{\text{BSim}_p^{(t_i)}}$ 由各类型比例 C_p^v 实际值的向量形式表达。例如，功能区 G_p 中居住用地、公共用地、商业用地、绿地广场用地、工业用地和交通设施用地分别占比 30%、20%、10%、20%、10% 和 10%，则地理位置动态语义属性

$$\overrightarrow{\text{BSim}_p^{(t_i)}} = \{0.3, 0.2, 0.1, 0.2, 0.1, 0.1\}.$$

根据相似卡口序列轨迹对的特征，对轨迹序列第 m 条轨迹 $T_m = \{g_1^{(m)}, g_2^{(m)}, \cdots, g_i^{(m)}, \cdots\}$ 和第 n 条轨迹 $T_n = \{g_1^{(n)}, g_2^{(n)}, \cdots, g_i^{(n)}, \cdots\}$ 从空间、时间和出行行为维度进行相似性度量。

1.3.1　空间相似性度量

Word2Vec（Mikolov et al.，2013）是一种用于从原始文本中训练单词向量的数据驱动方法，其中的 Skip-Gram 模型认为文本中的每个词都与上下文有着紧密关联，基于此设计了一个自监督学习任务，因此通过 Skip-Gram 模型能够学习到稀疏的卡口数据关联关系。但由于其只考虑上下文约束，没有直接将目的地与相似轨迹建立显式联系，可能不足以让目的地一致或接近的轨迹向量更相似。

因此，引入目的地约束参数 σ 以提升 Word2Vec 模型对轨迹卡口与目的地的关联约束。若轨迹 T_m 和 T_n 的目的地卡口 $g_{\text{D}}^{(m)}$ 与 $g_{\text{D}}^{(n)}$ 一致，则参数 $\sigma = 1$；若轨迹 T_m 和 T_n 的目的地卡口 $g_{\text{D}}^{(m)}$ 与 $g_{\text{D}}^{(n)}$ 不一致，即存在地理空间距离，则以基于球面距离 d 的指数衰减函数作为目的地约束参数 σ 的表达。

$$d = 2R\arcsin\left[\sin^2\left(\frac{y_{\text{D}}^{(m)} - y_{\text{D}}^{(n)}}{2}\right) + \cos y_{\text{D}}^{(m)} \cos y_{\text{D}}^{(n)} \sin^2\left(\frac{x_{\text{D}}^{(m)} - x_{\text{D}}^{(n)}}{2}\right)\right],$$

式中地理位置 $(x_{\text{D}}^{(m)}, y_{\text{D}}^{(m)})$ 和 $(x_{\text{D}}^{(n)}, y_{\text{D}}^{(n)})$ 与轨迹 T_m 和 T_n 的目的地卡口 $g_{\text{D}}^{(m)}$ 与 $g_{\text{D}}^{(n)}$ 分别对应。两种目的地约束参数

$$\sigma = \begin{cases} 1, & g_{\text{D}}^{(m)} = g_{\text{D}}^{(n)}, \\ \mathrm{e}^{-\rho d}, & g_{\text{D}}^{(m)} \neq g_{\text{D}}^{(n)}, \end{cases}$$

σ 的取值范围为 $(0, 1]$。ρ 为衰减系数，决定了函数的衰减速度。$\rho > 1$ 时，指数函数衰减过快，将会大大衰减卡口关联关系对空间相似性的贡献。为了避免远距离目的地对模型的干扰，以及过于迅速使

轨迹对之间的关联关系衰减，本文ρ取0.05。将改进Word2Vec中的Skip-Gram模型作为空间相似性度量的评估模型，其工作原理如图3所示。

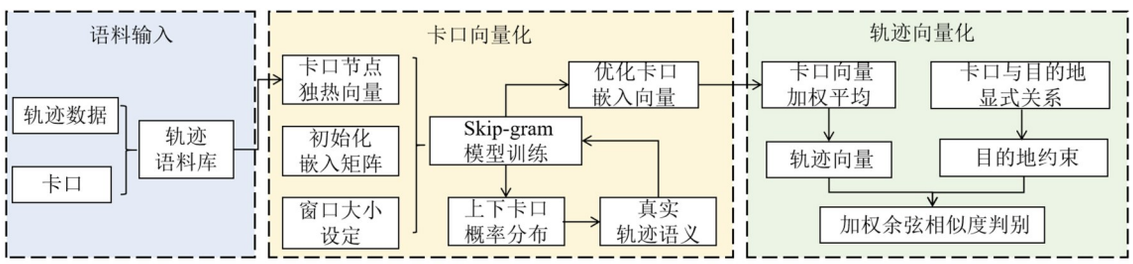

图3 改进Word2Vec方法的工作原理

Fig. 3 The working principle of the improved Word2Vec method

语料库以卡口为单词，轨迹序列为句子输入，以给定卡口点$g_i^{(m)}$为中心，设定窗口大小为u，初始化$g_i^{(m)}$的嵌入矩阵，得到初始的卡口嵌入向量$\overrightarrow{g_i^{(m)}}$，再放入Skip-Gram模型结合真实的轨迹语义训练，得到上下卡口序列最可能出现的$2u$个卡口概率分布，优化后的卡口嵌入向量$\overrightarrow{g_i^{(m)}}$的均值

$$\overrightarrow{T_m}=\frac{\overrightarrow{g_1^{(m)}}+\overrightarrow{g_2^{(m)}}+\cdots+\overrightarrow{g_i^{(m)}}+\cdots}{\left|\overrightarrow{g_1^{(m)}}\right|+\left|\overrightarrow{g_2^{(m)}}\right|+\cdots+\left|\overrightarrow{g_i^{(m)}}\right|+\cdots}.$$

引入σ的Word2Vec方法，对轨迹T_m和T_n进行了余弦相似性度量，可以定义空间相似性

$$\mathrm{SSim}_{m,n}=\sigma\times\frac{\overrightarrow{T_m}\cdot\overrightarrow{T_n}}{\left\|\overrightarrow{T_m}\right\|\cdot\left\|\overrightarrow{T_n}\right\|},$$

$\mathrm{SSim}_{m,n}\in[-1,0)\cup(0,1]$。$\mathrm{SSim}_{m,n}\in[-1,0)$时，轨迹$T_m$和$T_n$卡口序列存在一定相反性，不符合本文对轨迹相似性进行度量的目的，因此不考虑该类轨迹，只选取$\mathrm{SSim}_{m,n}\in(0,1]$的轨迹对。$\mathrm{SSim}_{m,n}$越接近0表示相似性越低，$\mathrm{SSim}_{m,n}=1$表示两条轨迹的卡口序列表达完全一致。

1.3.2 时间相似性度量 常见的时间相似性度量主要基于两条轨迹之间的共同轨迹段、开始时间和轨迹持续时间等特征（Xia et al.，2011），关注轨迹时间点本身，忽略了时间带来的交互语义信息。而依据交通流随时间的趋势变化得到的时间窗，强化了时间与居民活动模式之间的联系。因此，定义基于时间窗$\mathrm{Windows}_k$的指数衰减函数来量化两条轨迹之间的时间相似性。$f(\Delta t)=\mathrm{e}^{-\lambda|\Delta t|}$表示以$\Delta t$为自变量的时间相似性。其中，$\Delta t$分别由轨迹$T_m$和$T_n$中对应O点时间窗的时间差$\Delta t_{m,n}^{\mathrm{O}}=t_{\mathrm{O}}^{(m)}-t_{\mathrm{O}}^{(n)}$、对应D点时间窗的时间差$\Delta t_{m,n}^{\mathrm{D}}=t_{\mathrm{D}}^{(m)}-t_{\mathrm{D}}^{(n)}$和轨迹持续时间差值$\Delta t_{m,n}^{\mathrm{O}\to\mathrm{D}}=\left|t_{\mathrm{D}}^{(m)}-t_{\mathrm{O}}^{(m)}\right|-\left|t_{\mathrm{D}}^{(n)}-t_{\mathrm{O}}^{(n)}\right|$表示。对于O点时间窗的时间差$\Delta t_{m,n}^{\mathrm{O}}$，若在同一时间窗中，则$\Delta t_{m,n}^{\mathrm{O}}$为0；若在不同时间窗中，$\Delta t_{m,n}^{\mathrm{O}}$为时间窗的中心点之间的差值。D点时间窗之间的时间差$\Delta t_{m,n}^{\mathrm{D}}$同理。

当$\Delta t=\Delta t_{m,n}^{\mathrm{O}}$时，$f(\Delta t_{m,n}^{\mathrm{O}})$用O点相似性$\mathrm{TSim}_{m,n}^{\mathrm{O}}$表示；当$\Delta t=\Delta t_{m,n}^{\mathrm{D}}$时，$f(\Delta t_{m,n}^{\mathrm{D}})$用D点相似性$\mathrm{TSim}_{m,n}^{\mathrm{D}}$表示；当$\Delta t=\Delta t_{m,n}^{\mathrm{O}\to\mathrm{D}}$时，$f(\Delta t_{m,n}^{\mathrm{O}\to\mathrm{D}})$用轨迹持续时间相似性$\mathrm{TSim}_{m,n}^{\mathrm{O}\to\mathrm{D}}$表示。$\lambda$为衰减系数，为了避免时间关联关系的迅速下降，$\lambda$取0.05。

最后，由于轨迹OD点的时间相似性$\mathrm{TSim}_{m,n}^{\mathrm{O}}$，$\mathrm{TSim}_{m,n}^{\mathrm{D}}$及持续时间差的相似值$\mathrm{TSim}_{m,n}^{\mathrm{O}\to\mathrm{D}}$具有相同的范围$(0,1]$，基于这三个观测值，定义时间相似性$\mathrm{TSim}_{m,n}$。$\phi_1$、$\phi_2$、$\phi_3$为权重系数，

$$\mathrm{TSim}_{m,n}=\phi_1\mathrm{e}^{-\lambda\left|\Delta t_{m,n}^{\mathrm{O}}\right|}+\phi_2\mathrm{e}^{-\lambda\left|\Delta t_{m,n}^{\mathrm{D}}\right|}+\phi_3\mathrm{e}^{-\lambda\left|\Delta t_{m,n}^{\mathrm{O}\to\mathrm{D}}\right|}.$$

由于任一特征都无法单独作为时间相似性的唯一指标，且$\mathrm{TSim}_{m,n}^{\mathrm{O}}$、$\mathrm{TSim}_{m,n}^{\mathrm{D}}$和$\mathrm{TSim}_{m,n}^{\mathrm{O}\to\mathrm{D}}$具有相同影响，因此$\phi_1=\phi_2=\phi_3=1/3$。$\mathrm{TSim}_{m,n}$越接近0表示轨迹之间时间毫不相似，$\mathrm{TSim}_{m,n}$接近1表示两条轨迹持续时间一致且位于同一时间窗。

1.3.3 出行行为相似性度量 目的地预测任务中，轨迹起始和到达的地理位置动态语义能够表征其出行行为意图，因此本文不考虑无活动意义的无数据区。在此基础上，动态语义地图$\mathrm{MAP}_{\mathrm{dynamic}}$将研究区域划分成了不同时间窗$\mathrm{Windows}_k$的语义地图$(\bigcup_p G_p)_k$，涵盖了每个功能区$G_p$的用地性质。结合卡口节点$g_i$所在时间戳$t_i$和虚拟停车场编号$p$，可以查询$g_i$对应的地理位置动态语义属性$\overrightarrow{\mathrm{BSim}_p^{(t_i)}}$。因此，根据轨迹$T_m$和$T_n$的出发点和到达点的地理位置动态语义属性$\overrightarrow{\mathrm{BSim}_p^{(t_{\mathrm{O}}^{(m)})}}$，$\overrightarrow{\mathrm{BSim}_p^{(t_{\mathrm{D}}^{(m)})}}$和$\overrightarrow{\mathrm{BSim}_p^{(t_{\mathrm{O}}^{(n)})}}$，$\overrightarrow{\mathrm{BSim}_p^{(t_{\mathrm{D}}^{(n)})}}$衡量轨迹之间的出行行为相似性。出行行为相似性

$$\mathrm{BSim}_{m,n} = \omega_1\left(\overrightarrow{\mathrm{BSim}_p^{(t_O^{(m)})}} \cdot \overrightarrow{\mathrm{BSim}_p^{(t_O^{(n)})}}\right) + \omega_2\left(\overrightarrow{\mathrm{BSim}_p^{(t_D^{(m)})}} \cdot \overrightarrow{\mathrm{BSim}_p^{(t_D^{(n)})}}\right),$$

式中权重系数 $\omega_1 = \omega_2 = 0.5$。$\mathrm{BSim}_{m,n}$ 趋于 0 表示轨迹 T_m 和 T_n 之间的出行行为十分不一致，$\mathrm{BSim}_{m,n} = 1$ 表示轨迹之间的出行行为完全一致。

1.3.4　综合度量　基于加权系数 α、β、γ，定义综合相似性

$$\mathrm{Sim}_{m,n} = \alpha \times \mathrm{SSim}_{m,n} + \beta \times \mathrm{TSim}_{m,n} + \gamma \times \mathrm{BSim}_{m,n}.$$

Dorosti et al.(2024)认为，在预测任务中空间相似性略大于时间相似性的影响。且时空维度直接反映了轨迹之间的相似性，而出行行为间接反映了车辆之间的驾驶行为相似性，因此认为出行行为相似性的占比最小。

2　试验与讨论

2.1　试验设置

试验数据为 2023-04-17 宣城市主干路网的 58 586 条卡口数据，涉及 7 448 个卡口点、17 404 条路段。根据车辆每次出行的 OD 数据，按照车辆编号以及开始结束时间与卡口数据进行匹配，共生成了 3 082 条轨迹数据，包含经过卡口的时间、经纬度、卡口编号、虚拟停车场等信息。

图 4 为宣城市主城区示意图，粉色区域为主城区虚拟停车场的覆盖图，圆圈点为主干路口卡口点，连线成为主干路网。根据所有轨迹数据中 OD 点对应的虚拟停车场，筛选出 254 个主干路级的虚拟停车场。

同时，从百度地图中爬取 POI 数据（包括 POI 名称、经纬度、所在地址）。查询遥感影像，确定各类 POI 的平均建筑面积或占地面积，如增强出版：附表 2 所示。对于面积越大的 POI，划定权重更大。根据权重分值对原 POI 数据的数量进行重新定义。例如：某停车场原 POI 有 2 个点，而其权重值为 50 分，因此现在该停车场的 POI 点数为 100 个。

试验环境为搭载 16 核 24 线程 i7-13700 CPU 与 intel（R）UHD Graphics 770 显卡的台式机，内存 32GB，软件环境为 Python 3.8.3。

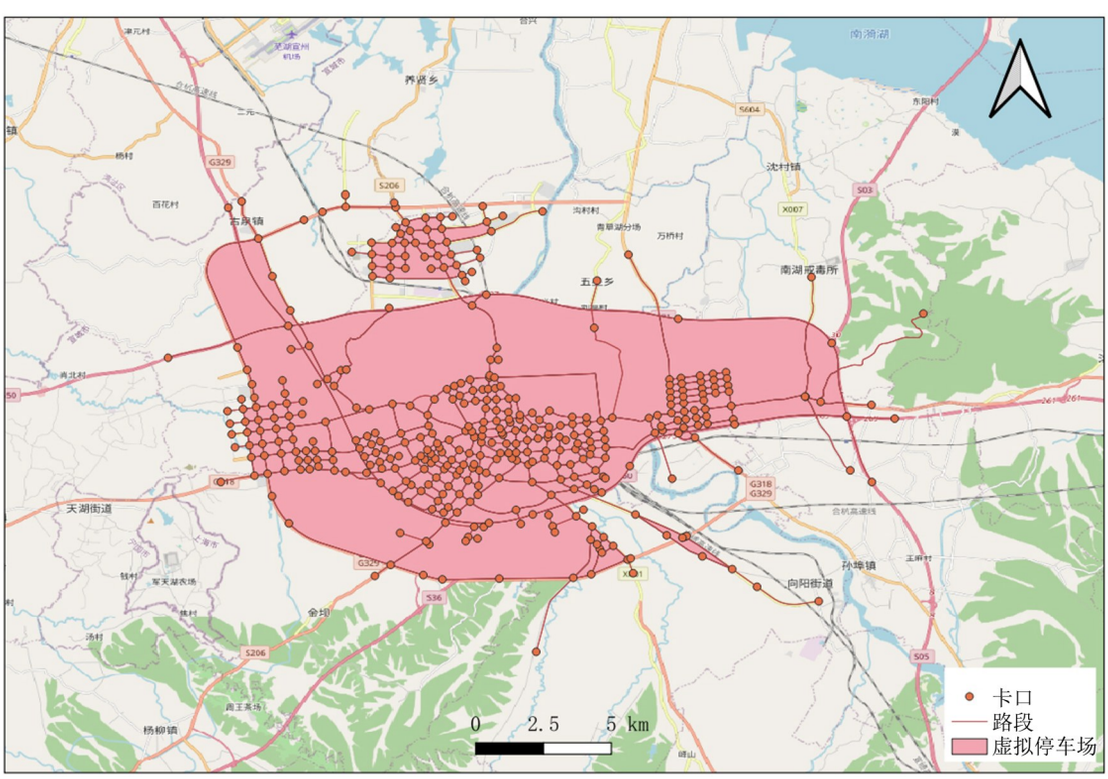

图 4　宣城市主城区示意图

Fig. 4　Schematic diagram of Xuancheng main city area

2.2　自适应时间窗及功能区分析

通过动态语义地图构建试验，得到自适应时间窗的划分结果。并根据构建好的动态语义地图，分析功能区随时间窗的变化情况。

将车辆每次出行的出发点及对应时间作为一次出行，每 5 min 向下取整统计一次交通流量，并进行异常值修正。在此基础上，对折线变化趋势进行 LOESS 平滑拟合。试验表明，拟合效果最好的平滑系数为 0.1。之后，引入最小间隔约束（≥10 个数据点）合并相邻过近的拐点，确保各时段具备统计显著性，得到的结果如图 5 所示。

自适应时间窗的划分结果，如表 1 所示。为了与动态语义地图的时间窗划分结果保持一致，本文根据 $Windows_1$ 至 $Windows_{10}$ 中 00:05:00—23:50:00 的时间段内选择出行且到达的车辆轨迹。

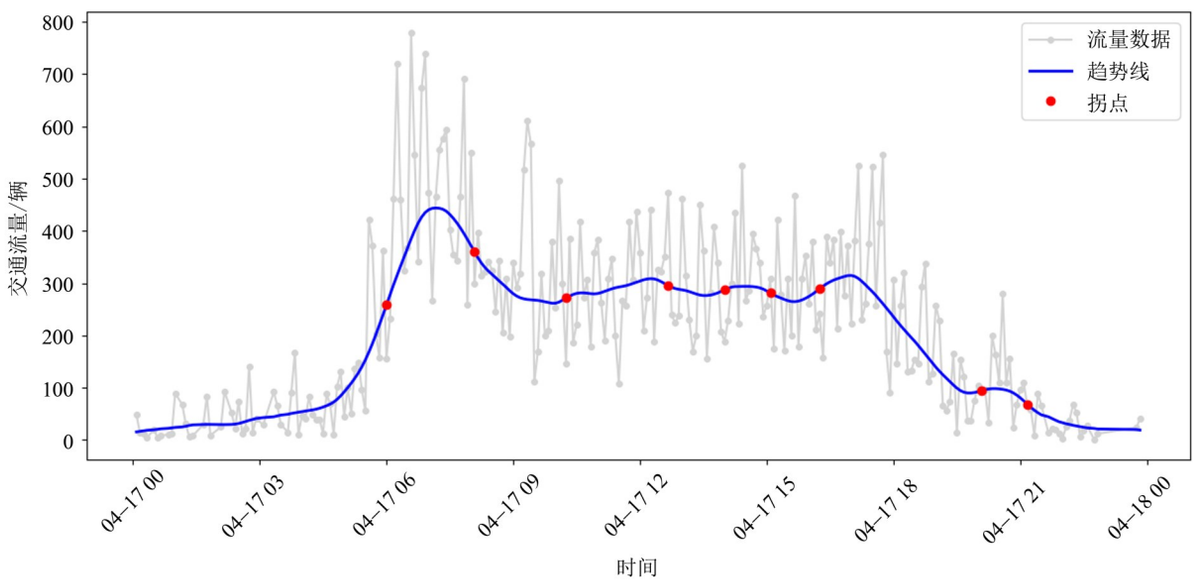

图 5 2023 年 4 月 17 日的拐点-趋势图

Fig. 5 Turning point-trend chart on 2023-04-17

表 1 自适应时间窗划分结果

Table 1 Time window division results

时间	1	2	3	4	5	6	7	8	9	10
起始时间	00:05:00	06:00:00	08:05:00	10:15:00	12:40:00	14:00:00	15:05:00	16:15:00	20:05:00	21:10:00
结束时间	06:00:00	08:05:00	10:15:00	12:40:00	14:00:00	15:05:00	16:15:00	20:05:00	21:10:00	23:50:00

在筛选出的虚拟停车场中除去无数据的功能区后，按时间窗分配，得到 1 087 个功能区，如表 2 所示。按照用地性质划分，可以得到 949 个单一功能区，138 个混合功能区。可以看出，单一功能区比重较大，混合功能区占比较少，反映了宣城市区域功能的集中性。由于不同时段功能区用地性质会发生变化，按时间变化对功能区进行划分，得到一天中随时间窗变化的功能区有 46 个，未发生变化的有 208 个。

表 2 功能区统计

Table 2 Functional area statistics

划分	功能区	个数	占比
按用地性质划分	单一功能区	949	87.3%
	混合功能区	138	12.7%
按时间变化划分	动态功能区	46	18.1%
	固定功能区	208	81.9%

2.3 权重分配

比较空间相似性 $SSim_{m,n}$、时间相似性 $TSim_{m,n}$ 和出行行为相似性 $BSim_{m,n}$ 的权重系数对相似性度量结果的影响。4 种权重组合下的相似性度量值如表 3 所示。可以发现，过度强调空间系数会导致其他特征的失衡，空间权重系数 α 不宜超过 0.5。而较高的时间权重或许会对整体相似性造成负面的影响。与此同时，出行行为权重系数 γ 不宜低于 0.2。$\alpha = 0.5, \beta = 0.3, \gamma = 0.2$ 的组合维持了一个相对均衡的权重，并且得到了最佳的相似性度量结果。

表3　不同权重组合下的度量值比较

Table 3　Comparison of similarity measure value under different weight combinations

组合	α	β	γ	度量值
1	0.5	0.3	0.2	0.967 3
2	0.5	0.4	0.1	0.966 0
3	0.6	0.3	0.1	0.963 5
4	0.7	0.2	0.1	0.959 6

2.4　模型评估

2.4.1　评价指标　针对本文的应用场景，随机挑选轨迹点大于 10 的 500 条轨迹作为实验对象。并根据轨迹层次聚类的方式(Liang et al.，2021)，利用本文相似性度量方法，找到类间距离 BC 和类内距离 WC 的聚合系数 $\mathrm{AC}=\dfrac{\mathrm{WC}}{\mathrm{WC}+\mathrm{BC}}\in(0,1)$ 来评估相似性度量方法的有效性。其中，AC 越小则表示轨迹相似性度量的有效性越好。

2.4.2　对比模型　本文采用两种对比模型：(1)仅基于改进 Word2Vec 模型，利用空间序列关联关系度量轨迹相似性。(2)基于耦合传统度量方法和动态语义地图的模型，从空间、时间及出行行为三个维度上度量轨迹相似性。

Hausdorff 距离能够衡量两条轨迹之间的不匹配程度，适用于时间戳不规则的卡口序列轨迹。双向 Hausdorff 距离定义为

$$H(T_m,T_n)=\max\left\{h(T_m,T_n)=\max_{g_i^{(m)}\in T_m}\left(\min_{g_j^{(n)}\in T_n}\left\|g_i^{(m)}-g_j^{(n)}\right\|\right),\ h(T_n,T_m)=\max_{g_j^{(n)}\in T_n}\left(\min_{g_i^{(m)}\in T_m}\left\|g_i^{(m)}-g_j^{(n)}\right\|\right)\right\},$$

式中 $\min_{g_i^{(m)}\in T_m}\left\|g_i^{(m)}-g_j^{(n)}\right\|$ 为轨迹 T_m 中 $g_i^{(m)}$ 到轨迹 T_n 中 $g_j^{(n)}$ 的最小欧氏距离；$\min_{g_j^{(n)}\in T_n}\left\|g_i^{(m)}-g_j^{(n)}\right\|$ 同理；$h(T_m,T_n)$ 为最小欧式距离集合的最大值；$h(T_n,T_m)$ 同理。

对 Hausdorff 距离进行归一化。首先，找到轨迹 T_m 和 T_n 中所有点之间的最大欧式距离 $\max_{g_i^{(m)}\in T_m,g_j^{(n)}\in T_n}\left(\left\|g_i^{(m)}-g_j^{(n)}\right\|\right)$，得到 $H(T_m,T_n)$ 与其的比值 $\mathrm{NH}(T_m,T_n)\in[0,1]$。

$$\mathrm{NH}(T_m,T_n)=\frac{H(T_m,T_n)}{\max_{g_i^{(m)}\in T_m,g_j^{(n)}\in T_n}\left(\left\|g_i^{(m)}-g_j^{(n)}\right\|\right)}.$$

$\mathrm{NH}(T_m,T_n)$ 越大则表示轨迹间距离尺度越大。最后，用 $\mathrm{NH}(T_m,T_n)$ 替换 Word2Vec 模型的空间相似性 $\mathrm{SSim}_{m,n}$，并耦合动态语义地图，得到新的 Haus-MAP 模型。

2.4.3　试验结果　选择 5~15 个聚类进行聚类效果评价。如表 4 所示，本文研究方法在 11 次实验(K=5~15)均表现出最低的 AC 值，与其他两种方法存在显著差距，这表明本文研究方法能够在较少的簇数下实现更优的聚合效果。相较之下，Haus-map 方法的 AC 值在整个 K 值范围内变化较小，始终维持在较高水平，说明其对不同轨迹之间的结构差异敏感性较弱，难以有效区分不同簇。而改进的 Word2Vec 方法尽管在较高聚类簇下 AC 值有所降低，但其鲁棒性和聚类效果仍明显不如本文方法。

进一步比较后发现，本研究方法的平均 AC 值较改进 Word2Vec 方法低 0.20，较 Haus-map 方法低 0.51。这是因为改进 Word2Vec 方法缺乏对时间、车辆出行行为维度的刻画能力，因而难以展现全面的方法优势。而 Haus-map 方法在空间维度基于传统度量方法，更适用于 GPS 密集轨迹场景。因此，针对稀疏轨迹问题及稀疏的卡口节点时，本文方法轨迹相似性度量效果更好。

表4　不同度量方法下的 AC 值

Table 4　AC value under different measurement methods

K	AC 值		
	改进 Word2Vec	Haus-map	本文方法
5	0.45	0.68	0.13
6	0.44	0.68	0.13
7	0.44	0.68	0.11
8	0.42	0.68	0.11
9	0.42	0.66	0.10
10	0.25	0.66	0.10
11	0.24	0.66	0.09
12	0.24	0.66	0.09
13	0.18	0.66	0.08
14	0.10	0.42	0.08
15	0.09	0.25	0.06

表 5 展示了轨迹降采样率为 20%、30%、40% 和 50% 四种情况下三种算法的轨迹聚类效果。可以看出，随着聚类簇的增加，本文研究方法的 AC 值平稳下降。在簇数 K<9 时，改进 Word2Vec 方法的 AC 值明显高于本文研究方法；在簇数 $K\geqslant 9$ 之后，其数值骤然下降，原因是其仅依赖邻近卡口的共现关系进行嵌入学习，难以在粗粒度聚类中保持类内一

表5 不同降采样率下的AC值

Table 5 AC values under different down sampling rates

K	改进Word2Vec方法				Haus-map方法				本文方法			
	降采样20%	降采样30%	降采样40%	降采样50%	降采样20%	降采样30%	降采样40%	降采样50%	降采样20%	降采样30%	降采样40%	降采样50%
5	0.45	0.45	0.45	0.45	0.68	0.68	0.68	0.68	0.13	0.12	0.13	0.13
6	0.45	0.45	0.45	0.45	0.68	0.68	0.68	0.68	0.13	0.12	0.13	0.13
7	0.44	0.44	0.44	0.44	0.68	0.68	0.68	0.68	0.12	0.12	0.11	0.12
8	0.44	0.43	0.44	0.43	0.68	0.68	0.68	0.68	0.11	0.11	0.11	0.11
9	0.19	0.42	0.26	0.26	0.66	0.66	0.66	0.66	0.10	0.10	0.10	0.10
10	0.17	0.24	0.24	0.24	0.66	0.66	0.66	0.66	0.10	0.09	0.10	0.10
11	0.17	0.24	0.24	0.24	0.66	0.66	0.66	0.66	0.09	0.09	0.09	0.09
12	0.17	0.18	0.19	0.18	0.66	0.66	0.66	0.66	0.09	0.08	0.09	0.09
13	0.12	0.10	0.10	0.10	0.66	0.66	0.66	0.66	0.08	0.08	0.08	0.08
14	0.10	0.06	0.06	0.06	0.42	0.42	0.42	0.42	0.06	0.07	0.08	0.07
15	0.05	0.06	0.06	0.06	0.25	0.25	0.25	0.25	0.05	0.07	0.07	0.07

致性和类间区分性，表现出较弱的聚类稳健性。Haus-MAP方法在所有降采样率下AC值均较高，在簇数K > 13后下降迅速，原因是Hausdorff距离缺乏对稀疏轨迹结构扰动的鲁棒性，导致轨迹聚类效果不稳定。

同时，不同降采样率下，本文方法在9次实验（K = 5~13）中均表现出最低的AC值，比其他两种方法平均低0.34。基于降采样结果，本文方法展现了更强的鲁棒性，适用于基于稀疏卡口数据的目的地预测场景。进一步分析，三种方法的轨迹相似性度量效果受降采样率的影响较小。这是因为Haus-MAP方法中的Hausdorff距离为基于形态的轨迹相似性度量方法，对采样率变化不敏感；另外两种方法中的Word2Vec模型更关注卡口序列之间的上下文关系，尽管采样率发生变化，两种方法仍能够把握轨迹整体结构特征。

3 结 论

针对目的地预测任务中轨迹样本稀疏的问题，本文提出了一种耦合改进Word2Vec和动态语义地图的轨迹相似性度量方法。该方法从空间相似性、时间相似性和出行行为相似性三个维度进行轨迹综合度量，能够在稀疏的卡口数据中发挥有效的识别能力，适用于目的地预测的场景。同时，本文基于层次聚类效果研究了方法的有效性，验证了方法在不同降采样率下具有稳健性。主要结论如下：

1）城市功能区在一天之中随时间窗变化的功能区有46个，证明了功能区的动态变化特性。

2）本文方法的层次聚类AC值相较于Haus-map方法平均低0.51，表明了考虑序列关联性的自然语言处理模型在轨迹聚类中的有效性。

3）本文方法的层次聚类AC值相较于仅基于改进Word2Vec的方法平均低0.20，验证了动态语义地图通过对时间、出行行为维度的轨迹刻画，提升了相似性轨迹的识别能力。

4）对比其他两种方法，本文方法在不同降采样率下AC值平均低0.34，并且随着聚类簇增加，本文方法的AC值平稳下降，验证了其对车辆轨迹相似性度量的稳健识别能力。

然而，本文仅考虑了一天之中的语义地图变化，且POI的开放时间依据经验值设定，未能获取更为精准的时间数据。因此，动态语义地图在交通流、目的地预测等方面的应用还有待进一步探索。

参考文献：

池娇，焦利民，董婷，等，2016. 基于 POI 数据的城市功能区定量识别及其可视化[J]. 测绘地理信息，41(2)：68-73.

江婧，张怀峰，皮德常，2019. 基于卷积神经网络的移动对象目的地预测[J]. 小型微型计算机系统，40(12)：2519-2525.

晋广印，赵旭俊，龚艺璇，2024. 基于长短期记忆网络的移动轨迹目的地预测[J]. 计算机工程与科学，46(3)：525-534.

李威，2018. 基于历史轨迹的车辆类别预测[D]. 山东：山东大学.

罗月童，汪涛，杨梦男，等，2021. 基于历史行车轨迹集的车辆行为可视分析方法[J]. 计算机科学，48(9)：86-94.

吴晨昊，向隆刚，张叶廷，等，2023. 基于地理空间感知型表征学习的轨迹相似度计算[J]. 测绘学报，52(4)：670-678.

余丹青，邬群勇，姚江涛，等，2023. 融合卷积、注意力和 MLP 的出租车目的地预测[J]. 计算机工程与应用，59(11)：302-311.

余志，黄敏，何兆成，2022. 道路交通信息物理系统(TCPS)概述[EB/OL]. (2022-05-18)[2025-06-03]. https://www.openits.cn/news/821.jhtml.

郑国强，乔宇昊，孙思民，2023. 基于 POI 数据的城市功能区识别研究[J]. 地理空间信息，21(10)：58-61.

周杭，樊红，2022. 基于众源地理数据的城市功能区及其热点的识别研究[J]. 武汉大学学报(工学版)，55(4)：417-426.

CAO S，WU L，WU J，et al，2022. A spatio-temporal sequence-to-sequence network for traffic flow prediction [J]. Inf Sci，610：185-203.

DOROSTI A，ALESHEIKH A A，SHARIF M，2024. Measuring trajectory similarity based on the spatio-temporal properties of moving objects in road networks [J]. Information，15(1)：51.

DU S，ZHANG H，XU H，et al，2019. To make the travel healthier：A new tourism personalized route recommendation algorithm [J]. J Ambient Intell Humaniz Comput，10(9)：3551-3562.

FURTADO A S，KOPANAKI D，ALVARES L O，et al，2016. Multidimensional similarity measuring for semantic trajectories[J]. Trans GIS，20：280-298.

GAO S，JANOWICZ K，COUCLELIS H，2017. Extracting urban functional regions from points of interest and human activities on location-based social networks [J]. Trans GIS，21(3)：446-467.

GUI Z，SUN Y，YANG L，et al，2021. LSI-LSTM：An attention-aware LSTM for real-time driving destination prediction by considering location semantics and location importance of trajectory points[J]. Neurocomputing，440：72-88.

KANG J，MA H，DUAN Z，et al，2021. Vehicle trajectory clustering in urban road network environment based on Doc2Vec model [C]// International Joint Conference on Neural Networks. Shenzhen，China：1-8.

LI X，ZHAO K，CONG G，et al，2018. Deep representation learning for trajectory similarity computation [C]// IEEE 34th International Conference on Data Engineering. Paris，France：617-628.

LI Y，LIU C，LI Y，2022. Identification of urban functional areas and their mixing degree using point of interest analyses[J]. Land，11(7)：996.

LIANG M，LIU R W，LI S，et al，2021. An unsupervised learning method with convolutional auto-encoder for vessel trajectory similarity computation[J]. Ocean Eng，225：108803.

MIKOLOV T，CHEN K，CORRADO G，et al，2013. Efficient estimation of word representations in vector space [C]// International Conference on Learning Representations.Chicago，America：1301-1313.

SCHLOSSER F，BROCKMANN D，2021. Finding disease outbreak locations from human mobility data [J]. EPJ Data Sci，10(1)：1-17.

SHANG S，CHEN L，WEI Z，et al，2018. Parallel trajectory similarity joins in spatial networks[J]. VLDB J，27(3)：395-420.

XIA Y，WANG G Y，ZHANG X，et al，2011. Spatio-temporal similarity measure for network constrained trajectory data[J]. Int J Comput Intell Sys，4(5)：1070-1079 .

XUE A Y，ZHANG R，ZHENG Y，et al，2013. DesTeller：A system for destination prediction based on trajectories with privacy protection [J]. Proc VLDB Endow，6(12)：1198-1201.

（责任编辑　王海蓉）