计算机科学与探索
*Journal of Frontiers of Computer Science and Technology*
ISSN 1673-9418,CN 11-5602/TP

# 《计算机科学与探索》网络首发论文

题目：多角度语义轨迹相似度计算模型
作者：蔡明昕，孙晶，王斌
网络首发日期：2020-10-23
引用格式：蔡明昕，孙晶，王斌．多角度语义轨迹相似度计算模型．计算机科学与探索．https://kns.cnki.net/kcms/detail/11.5602.TP.20201023.1021.011.html

计算机科学与探索 doi: 10.3778/j.issn.1673-9418.2008095

# 多角度语义轨迹相似度计算模型

蔡明昕+，孙晶，王斌
东北大学计算机科学与工程学院，沈阳 110169
+ 通信作者 E-mail:1801700@stu.neu.edu.cn

**摘要：**移动设备的发展使得轨迹数据可以记录更多有用的信息，比如签到信息、活动信息，构成了语义轨迹数据。

快速有效的轨迹相似度计算会为分析问题带来巨大好处，已有学者对轨迹相似性及语义轨迹相似性做出研究，并提出了一些有效的方法。但是现有轨迹相似性计算方法无法应用于语义轨迹数据，而目前的语义轨迹相似性计算方法又在轨迹采样频率低的情况下效果不佳。因此本文在解决轨迹相似性计算对低采样频率敏感的基础上，结合了语义轨迹的附加访问地点信息提出了一种新的轨迹相似性计算模型，叫做多角度语义轨迹（Multi-Aspect Semantic Trajectory，MAST）相似度计算。模型基于 LSTM 并且引入自注意力机制，学习到的轨迹表达为多个关注轨迹不同方面的低维向量，构成了一个矩阵，从而解决了单一向量无法准确表达轨迹的问题。这个矩阵不仅包含轨迹的空间信息同时也包含语义信息，可用于计算语义轨迹相似度。本文提出的模型在两个现实语义轨迹数据集上进行实验，实验数据表明 MAST 的计算结果优于现有方法。

**关键词：**轨迹相似度计算；语义轨迹；自注意力机制；深度表示学习

文献标志码：A 中图分类号：TP399

蔡明昕，孙晶，王斌. 多角度语义轨迹相似度计算模型[J]. 计算机科学与探索
CAI M X, SUN J, WANG B. Multi-Aspect Semantic Trajectory Similarity Computation Model[J]. Journal of Frontiers of Computer Science and Technology

## Multi-Aspect Semantic Trajectory Similarity Computation Model

CAI Mingxin+, SUN Jing, WANG Bin
College of Computer Science and Engineering, Northeastern University, Shenyang 110169, China

**Abstract:**The development of mobile devices enables trajectory data to record more useful information, such as semantic information and activity information, constituting semantic trajectory data. Fast and effective trajectory similarity computation will bring great benefits to the analysis of problems. Scholars have studied trajectory similarity and semantic trajectory similarity and proposed some effective methods. However, existing trajectory similarity computation methods cannot be applied to semantic trajectory data, and the current semantic trajectory similarity computation methods do not work well under the condition of low trajectory sampling frequency.In this paper, a new trajectory similarity computation model, called Multi-Aspect Semantic Trajectory is proposed based on solving the sensitivity of trajectory similarity computation to low sampling frequency. The model is based on LSTM and introduces the self-attention mechanism. The learned trajectory is expressed as multiple low-dimensional vectors of different aspects of the trajectory, forming a matrix,

*The National Key Research and Development Program of China (No. 2018YFB1700404) (国家重点研究发展计划项目);National Natural Science Foundation of China (No. U1736104) (国家自然科学基金项目); the Fundamental Research Funds for the Central Universities (No. N171602003) (中央高校基础研究基金项目)。

thereby solving the problem that a single vector cannot accurately express the trajectory. This matrix contains not only the spatial information of the trajectory but also the semantic information, which can be used to calculate the similarity of the semantic trajectory. MAST is tested on two realistic semantic trajectory datasets. Experimental data shows that MAST is superior to existing methods.
**Key words:** trajectory similarity computation; semantictrajectory; self-attention mechanism; deep representation learning

## 1 引言

GPS 设备的普及以及传感器技术的发展使得规律性地记录位置信息成为可能，轨迹数据也因此产生。近年来，针对轨迹数据的研究得到了快速的发展。T 时刻坐标为(x,y)的轨迹点可以表示为(x,y,t)，轨迹数据就是这样时空点的序列。在对轨迹数据的研究中，轨迹相似度计算是一个热门话题，它是许多其他轨迹研究的基础，例如轨迹聚类[1]，轨迹异常点检测[2]，运动模式分析[3]，用户行为分析[4]等。

快速有效的轨迹相似度计算会为分析问题带来巨大好处，目前学者们提出了一些轨迹相似度计算方法，包括动态时间包（DTW）[5],最长公共子序列（LCSS）[6],实际序列编辑距离（EDR）[7]，点较准（APM）[8]，带投影的编辑距离（EDwP）[9]等等。其中[5]-[7]基于两条轨迹对应轨迹点的两两配对，通过动态编程获得最佳的匹配策略，这类方法的计算复杂度较高并且在轨迹点采样率不均匀的情况下效果不佳。为了解决轨迹相似度计算对采样率变化敏感的问题，APM 和 EDwP 被提出。同样为了解决此问题，李秀成[10]创新性地首次采用深度学习的方法将轨迹用向量表示，使用轨迹及其子轨迹作为一对训练数据以获得轨迹在低采样率条件下的最佳表示。

随着微博、脸书、推特等移动应用的普及，使得记录的轨迹附加额外的信息，如签到地点、活动信息等等，这样的轨迹称为语义轨迹[11][12]，在文章[13][14]中有对语义轨迹的定义。为了表达简洁，本文采用论文[15]中的定义，即语义轨迹由访问点的序列组成，每个轨迹点记录的信息包含坐标、时间和访问点的名称。

然而上述提到的方法仅仅考虑了轨迹数据的空间特性，对于语义轨迹并不能完全适用，而轨迹数据的语义信息往往反应了用户的喜好或者行为习惯，是轨迹数据挖掘的重要帮助信息。如图 1(a)所示，轨迹 A、B 的路径相同，轨迹 A 的一个轨迹点$a_2$的访问点类型为商店，轨迹 B 中与之对应的轨迹点为$b_2$，其访问点类型为饭店。$a_2$与$b_2$在空间上十分临近，但访问点却不同（可以理解为商店与饭店位置临近），因此我们不能认为轨迹 A、B 十分相似。如图 1(b)，轨迹 C、D 路径不完全相同，但$c_3$与对应的点$d_2$有相同类型的访问点，我们认为在某种程度上两条轨迹有相似性。

目前已经有学者对语义轨迹相似性做出了研究，并且提出了一些方法。然而现有的计算方法在采样频率低和采样频率不均的情况下无法获得令人满意的结果。如图 1(c)所示，从轨迹 F 的三个采样点我们无法确定它的实际路径为实线还是虚线，从而无法准确计算语义轨迹相似度。

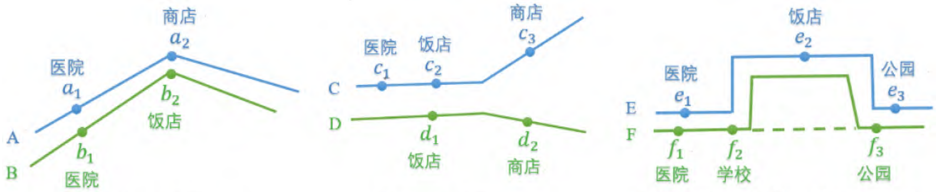



Fig.1Several cases of trajectory similarity computation

**图 1 轨迹相似性计算的几种情况**

不仅如此，之前深度表示学习轨迹大多采用编码器-解码器（encoder-decoder）框架，简单地将最后一层的隐层状态作为轨迹的低维度表达。这样做法的缺点是如果轨迹过长或者特征较多，用一个固定长度的向量无法准确表达轨迹信息。

本文在解决轨迹相似性计算对低采样频率敏感的基础上，结合了语义轨迹的附加访问地点信息提出了一种新的轨迹相似性计算模型，叫做多角度语义轨迹（Multi-Aspect Semantic Trajectory，MAST）相似度计算。学习到的轨迹低维表达矩阵不仅包含轨迹的空间信息同时也包含语义信息，可用于计算前文所述的语义轨迹相似度问题。

首先提取语义轨迹各个轨迹点的特征信息，将每个轨迹点转化为一个包含该点空间特征和语义特征的向量。为了获得整条轨迹更准确的低维表达以计算轨迹间的相似度，我们受到自然语言处理中处理句子方式[16]的启发设计了多角度提取特征的模型。

总结来说本文做出以下贡献：

1、我们分别设计了提取轨迹点空间特征和语义特征的方法，得到轨迹点的嵌入。这样每条轨迹被表示一个向量的序列。

2、为了获得轨迹更准确的低维表示，我们提出了一个基于双向 LSTM 的模型，并且引入自注意力机制来提取轨迹多方面的信息。同时，相似度计算对低采样率具有鲁棒性。

3、我们在从 Gowalla（一个基于位置的社交网络）爬取的语义轨迹数据集上进行了广泛的实验，实验结果表明本文提出的模型在语义轨迹相似度计算的准确性和有效性上比现存的方法有更好的表现。

## 2 相关工作

论文[17]中提出了一种语义轨迹相似性计算方法，首先根据轨迹的转向和速度的变化将轨迹拆分为几段子轨迹，然后计算轨迹之间距离相似度评分和语义相似度评分。其中距离相似度评分综合考虑到轨迹中心距离，轨迹长度以及轨迹起点到终点的位移；而语义相似度计算则是基于最长公共子序列算法找出轨迹之

间最长相同的访问点序列。最终综合这两方面评分得到语义轨迹相似度。该方法存在一定缺陷，例如，如果两个轨迹之间访问点不连续相似，那么此方法计算出的相似度会很低，因为它是基于 LCSS 算法的。

研究[18]中提出的方法将轨迹转化为语义的序列，通过识别轨迹的停留点划分停留区域。考虑到GPS数据的误差，文中把停留区域中包含的所有POI均考虑在内，每个停留区域用一个特征向量表示。随后根据特征向量建立轨迹的语义-位置序列，并且提出最大移动匹配(MTM)算法来比较轨迹之间的相似性。文中引入了语义重叠的概念，即如果两个用户的语义访问顺序具有相似的模式，那这两个用户会有相似的兴趣点，本文的语义相似度计算也是基于此。

而[19]中提出了一种特别的方法，分别利用不同的相似度度量函数计算轨迹不同维度的相似性，例如空间，语义，时间。然而这篇论文的研究重点不在于此，而是提出多维度相似性度量（MSM）算法。不同于传统的LCSS、EDR算法，该算法将至少有一维相似的点匹配，而不是只匹配所有维度相同的轨迹点。

然而，上述方法存在不足之处，即在GPS数据误差和采样频率低的情况下无法准确得到轨迹的真实路径，从而对相似度计算产生负面影响。受到[10]中解决轨迹采样频率影响的启发，我们开始思考深度表示学习方法提取轨迹的空间和语义特征，把学习到的特征用于轨迹相似度比较取得了较好的效果。

## 3 相关概念和问题定义

### 3.1 相关概念

**定义 1**（访问点）一个访问点记为$v = < l, t, m >$，其中$l$是访问点的坐标信息，$t$是时间信息，$m$代表访问点类型。

**定义 2**（语义轨迹）语义轨迹由访问点序列组成，记为$T$，一条语义轨迹可以表示为$T = \{v_1, v_2, \dots v_n\}$。

一条语义轨迹的示意图如图 2 所示，每个轨迹点包含坐标、时间和访问点类型信息。

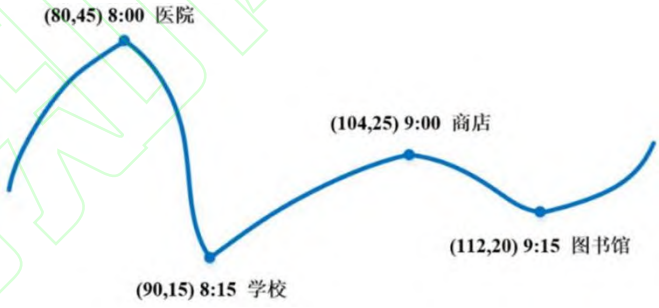



Fig.2An example of a semantic trajectory

**图 2 一条语义轨迹记录示例**

### 3.2 问题定义

给定一组语义轨迹，我们的目标是学习语义轨迹 T 矩阵形式的低维表达 M，M 能全面地包含轨迹的空间信息和语义信息。基于此表示的语义轨迹相似性计算对采样不均匀、低采样频率的轨迹具有鲁棒性。

## 4 多角度语义轨迹（Multi-Aspect Semantic Trajectory，MAST）相似度计算模型

下面我们将介绍 MAST 模型的细节。在第一节中，我们介绍如何提取轨迹的特征，即把轨迹嵌入到空间中，考虑轨迹数据的空间和语义信息。针对这两方面信息，我们设计了不同的嵌入方式。第二节详细介绍提出的模型结构以及使计算结果对低采样频率具有鲁棒性的方法。

### 4.1 轨迹特征提取

（1）空间特征嵌入

我们提出的 MAST 模型基于 LSTM，期待的输入为离散的标记序列。因此，我们应该将连续的空间信息映射为离散的标记。空间数据分析中常用的一种策略是将整个区域分成许多相等大小的单元格，这样每个点都可以表示为一个标记。为了更好地捕捉轨迹的空间临近性，我们使用跳跃语法（skip-gram）算法，通过最大化以下函数来初始化每个单元格的表示：

$$\frac{1}{N}\sum_{t=1}^{N}\sum_{-c\le j\le c, j\neq 0}\ln P(w_{t+j}|w_t) \quad (1)$$

如果给定当前单元格表示为$\boldsymbol{w_t}$，则$\boldsymbol{w_{t+j}}$代表$\boldsymbol{w_t}$的相邻单元格。概率用以下 softmax 函数计算：

$$P(w_O|w_I)=\frac{exp({v_{w_O}}^T\cdot v_{w_I})}{\sum_{w=1}^{W}exp({v_w}^T\cdot v_{w_I})} \quad (2)$$

其中$\boldsymbol{w_I}$是当前单元格，$\boldsymbol{w_O}$是此单元格的上下文，$\boldsymbol{v}$是单元格的向量表示。因此原始轨迹$T$可以转换为空间嵌入序列$T_S$：

$$T_S=\{S_1,S_2,S_3\cdots\} \quad (3)$$

其中$\boldsymbol{S_n}$为学习到的新单元格空间特征向量。

（2）语义特征嵌入

我们同样要把语义信息转化为向量表达，在本文的问题定义中，轨迹的语义信息相似也会提高轨迹相似计算的值。通过调查现实生活中公共场所的种类及分布，我们一共归纳出 35 种访问点类型，包括餐厅、宾馆、住宅等等，完整清单见附录。直观来说，对于这 35 种访问点的向量表达有两种方式：独热（one-hot）表达和分布（distributed）表达。独热表达是一种最简单的方式，它把访问点类型编码为互相独立的向量，即两个访问点类型只有相同和不同两种关系，这对于相似度计算并不是一个好的表达方式，而分布表达则可以避免这种情况。所以为了更准确地计算语义之间的相似性，我们选择分布表达。

根据现实生活中人们在不同类型访问点的活动，我们归纳出访问点的 10 个属性，分别为公共场地、消费、娱乐、学习、观光、运动、健康、交通、办公、停留时间，这 10 个属性分别对应语义特征向量的 10 个维度。对每种访问点，我们在以上的 10 个属性打分，将得到一个 10 维度的向量，这个向量即为该类型访问点的语义特征向量。例如电影院的语义特征向量为（0.7，0.8，0.8，0.2，0.3，0.3，0.3，0.4，0.1，0.5），依照此种方法得到 35 个访问点的语义特征向量。全部访问点的语义特征向量也可以在附录中找到。

原始轨迹$T$的语义信息可以转换为嵌入序列$T_C$：

$$T_c=\{C_1,C_2,C_3\cdots\} \quad (4)$$

### 4.2 MAST 模型

在得到以上两种特征向量以后，我们将详细介绍本文提出的 MAST 模型。

（1）MAST 结构

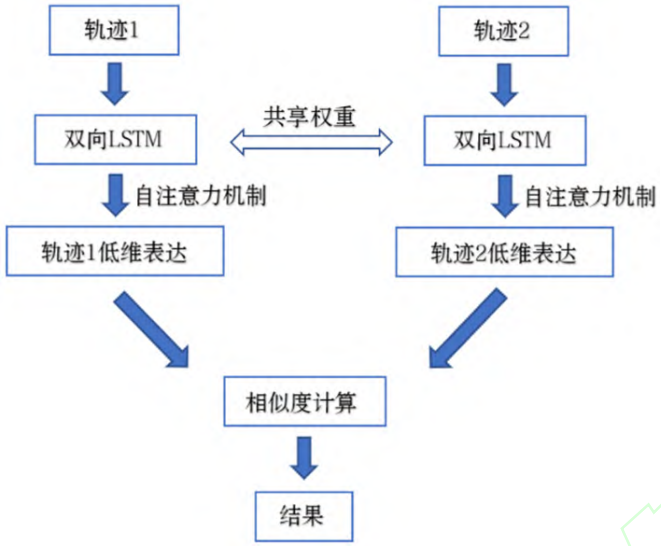
如图 3 所示，MAST 是由两个共享权重的双向 LSTM 组成，并且引入自注意力机制，两部分分别得到两条轨迹的低维表达用于计算轨迹相似性。我们以其中一部分为例介绍模型结构。

在双向 LSTM 上引入自注意机制，它为 LSTM 的隐层状态提供了一组权向量。这些权向量点乘 LSTM

Fig.3MAST structure

**图 3MAST 结构**

的隐层状态后求和，得到的加权 LSTM 隐层状态即是轨迹的低维表达。图 4 所示为获得某条轨迹低维表达过程的示例。



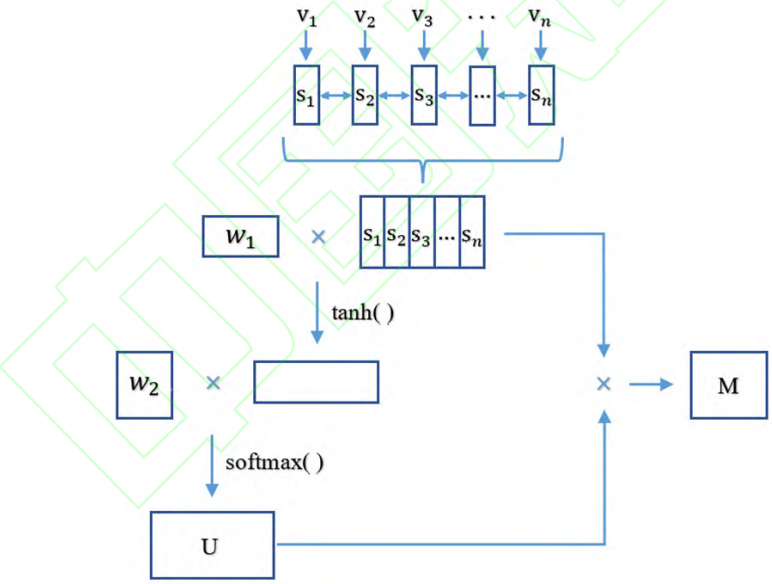




Fig.4Model computation process

**图 4 模型计算过程**

假设有一条轨迹 $T$，它有 $n$ 个访问点记录，即:

$$T = (v_1, v_2, \cdots v_n) \quad (5)$$

其中$\boldsymbol{v_i}$是通过上一节得到的 $k$ 维访问点空间和语义特征嵌入。因此，$T$ 为 $n$ 个 $k$ 维向量的序列，也就是一个 $n \times k$ 的二维矩阵。

现以 $T$ 作为双向 LSTM 的输入，获得轨迹中相邻访点之间的依赖性，可以得到：

$$\overrightarrow{s_t} = f(v_t, \overrightarrow{s_{t-1}}) \quad (6)$$

$$\overleftarrow{s_t} = f(v_t, \overleftarrow{s_{t+1}}) \quad (7)$$

将每个$\overrightarrow{s_t}$和$\overleftarrow{s_t}$连接，获得一个隐层状态向量$s_t$。设每个单向 LSTM 的隐层状态向量为 $a$ 维，把所有的 $n$ 个$s_t$记为 $\boldsymbol{S}$，则 $\boldsymbol{S}$ 是大小为 $2a \times n$ 的矩阵。

$$S = (s_1, s_2, \cdots s_n) \quad (8)$$

我们的目标是将一个可变长度的轨迹编码成一个固定大小的向量，由 $\boldsymbol{S}$ 中 $n$ 个 LSTM 隐层状态向量的线性组合来实现。线性组合的权重通过引入自注意机制而得到，自注意机制以整个 LSTM 隐层状态 $\boldsymbol{S}$ 为输入，输出权向量 $\boldsymbol{u}$:

$$u = softmax(w_2 tan\, h(w_1 S))(9)$$

其中$\boldsymbol{w_1}$是$b \times 2a$ 的权重矩阵，$\boldsymbol{w_2}$是$b$维度的参数向量，这里的$b$是一个可以任意设置的超参数，softmax( )确保所有计算的权重之和为 1。根据 $\boldsymbol{u}$ 提供的权值对 LSTM 隐层状态 $\boldsymbol{S}$ 加权求和，得到输入语义轨迹的向量表示形式 $\boldsymbol{m}$。

这种向量表示的信息集中在轨迹的特定部分，比如一组访问点类型相关联的轨迹点或者空间临近的轨迹点，它能够反映一条轨迹中的局部信息。因此，为了表示完整语义轨迹信息，我们设计了多角度的注意力机制，关注轨迹的不同部分。假设从轨迹中提取 $q$ 个不同的部分，需要将$\boldsymbol{w_2}$扩展成一个 $q \times b$的矩阵，得到权重矩阵 $\boldsymbol{U}$:

$$U = softmax(w_2 tanh(w_1 S))(10)$$

我们将权重矩阵 $\boldsymbol{U}$ 与 LSTM 隐层状态 $\boldsymbol{S}$ 相乘，得到的矩阵为完整轨迹的低维表达，于是轨迹的低维表达 $\boldsymbol{m}$ 变成一个 $q \times 2a$ 的矩阵 $\boldsymbol{M}$:

$$M = US(11)$$

同理，另一部分也是同样的方法得到另一条输入语义轨迹的低维表达矩阵，矩阵的每一列为轨迹某一部分的特征向量。这里我们计算两个矩阵对应列向量的余弦相似度值，并求和。两条轨迹越相似计算值越小。

（2）模型训练

当轨迹点采样频率低的情况下，为了让我们得到的轨迹表达更接近于轨迹的实际路径，训练模型时两个双向 LSTM 的输入分别为一条轨迹及它的子轨迹。用$T_a$表示某轨迹，$T_b$表示相对于这条轨迹采样率较低的子轨迹。这两条轨迹的实际路径和语义信息相似，所以标签设置为较小值（相似的轨迹计算结果小）。

根据以上分析，给定语义轨迹集合$\{T_{a1}, T_{a2}, \cdots T_{an}\}$。我们为集合中的每条轨迹创建多组轨迹对$(T_a, T_b)$，其中$T_a$为原始轨迹，$T_b$是从$T_a$中以采样率 r 随机采样轨迹点得到的，以此来模拟采样率不均匀且采样率低的轨迹。$T_a$的起点和终点在$T_b$中保留，以避免改变降采样轨迹的基本路径。此种方法训练的模型能最小化轨迹低维表达与实际轨迹的误差，使得模型对于轨迹采样频率的变化具有鲁棒性。

## 5 实验结果与分析

在本节中，我们利用多种轨迹相似度评价标准进行实验，证明所提出方法的有效性，并分析了各种参数对实验结果的影响。所有方法都在一台使用 GTX 1050Ti GPU 和 Intel i7 7700K CPU 的计算机上以 PyTorch 实现。

### 5.1 实验数据集

我们使用从 Gowalla 上爬取的两个语义轨迹数据集。第一个数据集是在美国洛杉矶市收集的，历时 5 个月，包含 100 万条轨迹。我们移除长度小于 5 的轨迹，得到 80 万条轨迹。第二个数据集包含在纽约市 7 个月内收集到的轨迹。我们选择长度至少为 5 的轨迹，共 100 万条。根据轨迹的起始时间戳将每个数据集划分为训练数据和测试数据，其中前 60 万条轨迹为训练数据，其余轨迹为测试数据。详情见下表 1。

Table 1 Dataset statistics

**表 1 数据集统计**

| 数据集 | 点数量 | 轨迹数量 | 平均长度 |
|---|---|---|---|
| 纽约 | 9721940 | 810162 | 12 |
| 洛杉矶 | 10334413 | 1003341 | 10.3 |

为了创建第 4 章 2 节所述的训练轨迹对，我们对数据集中用于训练的轨迹进行采样。对于每个轨迹$T_a$，分别以采样率 0.2，0.4，0.6，0.8 来创建它的 4 个子轨迹$T_b$。这样为每个原始轨迹$T_a$创建了 4 个训练对$(T_a, T_b)$。

### 5.2 超参数设置

（1）空间单元格大小：将原始空间信息转化为单元格表示，每个单元格的边长为 200 米。对于洛杉矶数据集，我们得到 12489 个单元格，对于纽约数据集得到 9856 个。

（2）网络参数：表 2 列出了模型的参数设置。

Table 2 Parameter setting

**表 2 参数设置**

| 参数 | 值 |
|---|---|
| 学习率 | 0.01 |
| 隐层维度 | 256 |
| 丢弃率 | 0.1 |
| 批量大小 | 64 |

### 5.3 评价标准

实际上我们不能直接判断两条轨迹是否相似，因为获取轨迹真实的相似性是一项复杂的任务。文献[8][9]中提出了以下三种分析模型有效性的方法:

（1）自相似度（Self-similarity measure, SSM）：从同一轨迹截取两个子轨迹，分别将子轨迹放入两个数据库中。假设两个子轨迹 A 和 B 从同一个原始轨迹截取，使用 A、B 等轨迹构成数据库$D_a$和$D_b$，其中 A 在$D_a$中，B 在$D_b$中。在最优情况下，当检索$D_b$中与 A 最相似的前 $k$ 条轨迹时，B 应该排在最前面。在我们的实验中，

将 A 分别与$D_b$中的每条轨迹组成一对输入轨迹数据，构成一组，我们把 SSM 的值定义为多组输入数据的 B 排名均值。

（2）交叉距离（Cross-distance, CSD）：这是一种新的距离评价准则，其中 V()表示低采样率或其他轨迹变化。该评价准则希望模型能够保持轨迹的原始相似性，而不受轨迹变化的影响。也就是说，交叉距离越小，模型对低采样的鲁棒性越好:

$$dev(T_a, T_b) = \frac{|d(V(T_a),V(T_b)) - d(T_a,T_b)|}{d(T_a,T_b)} \quad (12)$$

（3）KNN 查询:在我们的实验中，10000 个轨迹构成目标数据库。查询轨迹分别与目标数据库中的每个轨迹组成一对输入数据，根据计算出的相似度值在数据库中找到查询轨迹的 $k$ 个最相似的轨迹作为标注值。与 CSD 方法类似，我们获取轨迹变量并在目标数据库中计算其 $k$ 个最相似的轨迹。最后统计该方法求得的 $k$ 个轨迹在标注值的占比。

### 5.4 比较模型

为了说明我们的模型的有效性，我们将其与三个现有效果较好的语义轨迹相似度计算模型进行了比较：

（1）SMMO[17]：SMMO 根据轨迹的转向和速度的变化将轨迹拆分为几段子轨迹，然后计算轨迹之间距离相似度评分和语义相似度评分。

（2）MTM[18]：将轨迹转化为语义的序列，通过识别轨迹的停留点划分停留区域，每个停留区域用一个特征向量表示。提出最大移动匹配算法来比较轨迹之间的相似性。

（3）MSM[19]：分别利用不同的相似度度量函数计算轨迹不同维度的相似性，不同于传统的 LCSS、EDR 算法，该算法将至少有一维相似的点匹配，而不是只匹配所有维度相同的轨迹点。

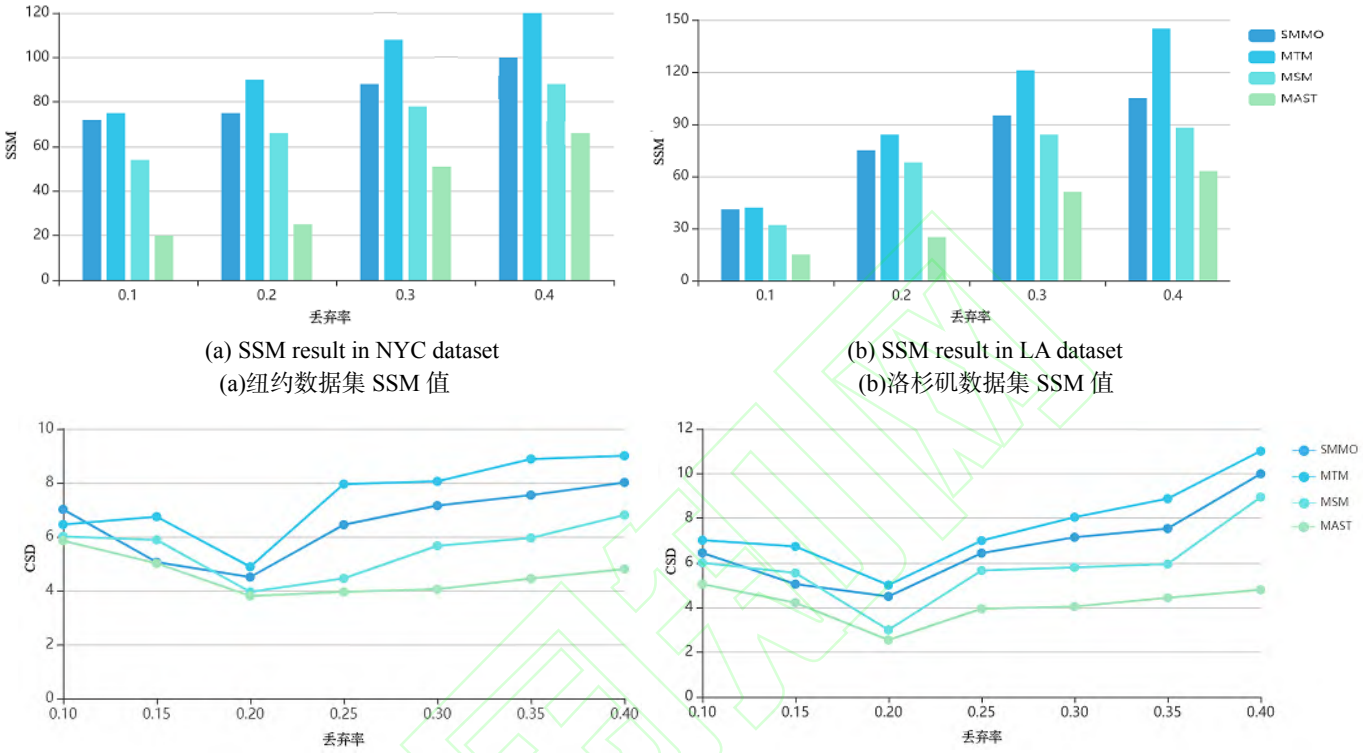
### 5.5 实验结果

接下来，我们展示了两个语义轨迹数据集在本文提出的 MAST 模型下的自相似度、交叉距离和 KNN 查询结果，并与目前表现较好的语义轨迹相似度计算模型（5.4 节列出）比较，而后测试了训练数据集大小对 MAST 模型计算结果的影响。由于计算结果与相关参数的选取有关，我们接着又评估了参数对轨迹相似度计算的影响。

（1）不同模型结果对比

图 5 为不同模型的实验结果，其中横坐标为对测试轨迹数据的随机丢弃率（制造低采样频率和采样频率不均匀的轨迹），纵坐标为 5.3 节中提出的评价标准的数值。从图 5(a)和图 5(b)可以看出，在所有的对照组中，MSM 的表现最好，SMMO 和 MTM 的性能比较差。而我们的模型在两个数据集上的性能都优于 MSM。随着轨迹点丢弃率的增加，本文提出的方法有效性更加明显。

如图 5(c)、5(d)，我们使用 CSD 数值的百分比作为纵坐标值。从图中可以看出，我们的模型要优于现有的方法。




(a) SSM result in NYC dataset
(a)纽约数据集 SSM 值

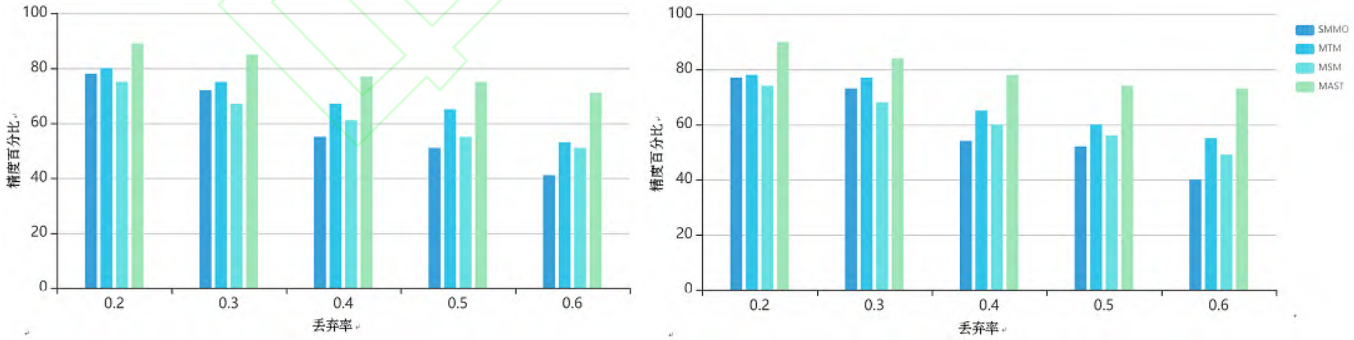
(b) SSM result in LA dataset
(b)洛杉矶数据集 SSM 值

(c) CSD result in NYC dataset
(c)纽约数据集 CSD 值

(d)CSD result in LA dataset
(d)洛杉矶数据集 CSD 值

Fig.5 SSM and CSD results of various methods

**图 5 不同方法的 SSM 和 CSD 计算结果**

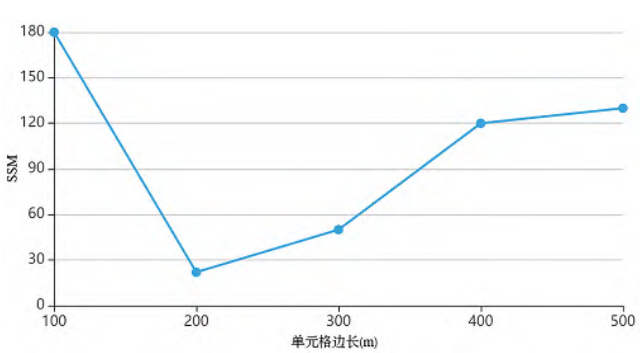


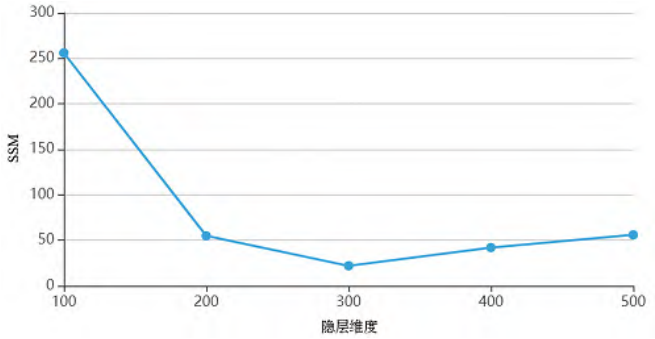

(a) k=20

(b) k=40

Fig.6 KNN results of various methods

**图 6 不同方法的 KNN 结果**




(a) Spatial cell
(a)空间单元格

(b) Hidden layer
(b)隐层

Fig.7 SSM and CSD results of various methods

**图 7 不同方法的 SSM 和 CSD 计算结果**

如图 6 所示，对于 KNN 查询，我们只展示洛杉矶市数据集上的实验结果（横纵坐标的含义同图 5）。本文提出的模型效果优于 MTM，SMMO 和 MSM 的实验效果不佳。随着轨迹点丢弃率的增加，所有方法的准确率都在下降。

（2）参数影响

下面我们以 SSM 为例分析空间单元格大小与隐层维数对计算结果的影响。对于空间单元格大小做了几组对比实验，实验结果如图 7(a)所示。从图中可以看出，当时空间单元尺寸为 200 m 时，训练效果最好。

隐层的维数将决定学习的质量。通常隐层维度越高效率更高，但对训练数据量的需求更大。从图 7(b)中可以看出，维度为 256 时为最优，之后逐渐减小。

（3）训练数据集大小影响

在本实验中，我们评估了训练数据集大小对轨迹相似度搜索精度的影响。与上一实验设置相似，我们只改变了训练数据集的大小，并将丢弃率 r 固定在 0.6。

图 8 显示了训练数据集大小对纽约数据集 SSM 计算结果的影响。当我们将训练数据从 10 万增加到 30 万时，SSM 值下降得很快，继续扩大训练数据的规模，下降的速度减慢。然而，当我们进一步增加训练数据的大小时，使用更大训练数据的效益就不那么明显了。

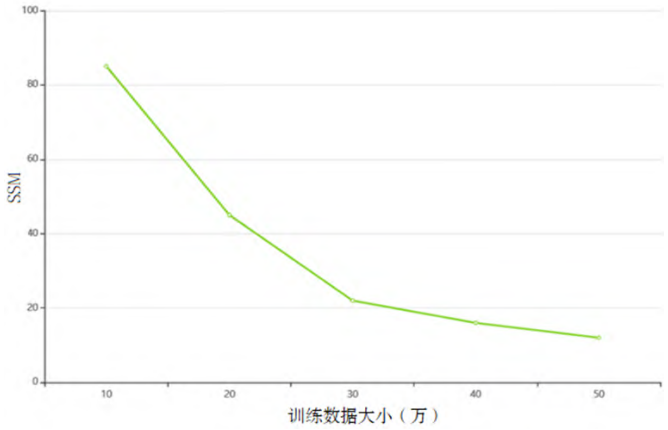



Fig.8 The impact of training dataset size on model computation results

**图 8 训练数据集大小对模型计算结果的影响**

## 6 结束语

本文在解决轨迹相似性计算对低采样频率敏感的基础上，结合了语义轨迹的附加访问地点信息提出了一种新的语义轨迹相似性计算模型（MAST）。模型基于 LSTM 并且引入自注意力机制，学习到的轨迹表达为多个关注轨迹不同方面的低维向量，从而解决了单一向量无法准确表达轨迹的问题。学习到的矩阵低维表达不仅包含轨迹的空间信息同时也包含语义信息，可用于计算语义轨迹相似度。本文提出的模型在两个现实语义轨迹数据集上进行实验，均得到了比现有语义轨迹相似性计算方法更优的实验结果。实验过程中发现模型对轨迹噪声点的鲁棒性还需要进一步提升，这是未来继续研究的方向。

## References:


[1] Hung C C, Peng W C, Lee W C. Clustering and aggregating clues of trajectories for mining trajectory patterns and routes[J]. The VLDB Journal, 2015, 24(2): 169-192.

[2] Zhu J, Jiang W, Liu A, et al. Effective and efficient trajectory outlier detection based on time-dependent popular route[J]. World Wide Web, 2017, 20(1): 111-134.

[3] Li Z, Han J, Ji M, et al. Movemine: Mining moving object data for discovery of animal movement patterns[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(4): 1-32.

[4] Bichicchi A, Belaroussi R, Simone A, et al. Analysis of Road-User Interaction by Extraction of Driver Behavior Features Using Deep Learning[J]. IEEE Access, 2020, 8: 19638-19645.

[5] Yi B K, Jagadish H V, Faloutsos C. Efficient retrieval of similar time sequences under time warping[C]//Proceedings 14th International Conference on Data Engineering. IEEE, 1998: 201-208.

[6] Vlachos M, Kollios G, Gunopulos D. Discovering similar multidimensional trajectories[C]//Proceedings 18th international conference on data engineering. IEEE, 2002: 673-684.

[7] Chen L, Özsu M T, Oria V. Robust and fast similarity search for moving object trajectories[C]//Proceedings of the 2005 ACM SIGMOD international conference on Management of data. 2005: 491-502.

[8] Su H, Zheng K, Wang H, et al. Calibrating trajectory data for similarity-based analysis[C]//Proceedings of the 2013 ACM SIGMOD international conference on management of data. 2013: 833-844.

[9] Ranu S, Deepak P, Telang A D, et al. Indexing and matching trajectories under inconsistent sampling rates[C]//2015 IEEE 31st International Conference on Data Engineering. IEEE, 2015: 999-1010.

[10] Li X, Zhao K, Cong G, et al. Deep representation learning for trajectory similarity computation[C]//2018 IEEE 34th International Conference on Data Engineering (ICDE). IEEE, 2018: 617-628.

[11] Alvares L O, Bogorny V, Kuijpers B, et al. A model for enriching trajectories with semantic geographical information[C]//Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems. 2007: 1-8.

[12] Parent C, Spaccapietra S, Renso C, et al. Semantic trajectories modeling and analysis[J]. ACM Computing Surveys (CSUR), 2013, 45(4): 1-32.

[13] Ying J J C, Lu E H C, Lee W C, et al. Mining user similarity from semantic trajectories[C]//Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks. 2010: 19-26.

[14] Bogorny V, Renso C, de Aquino A R, et al. Constant–a conceptual data model for semantic trajectories of moving objects[J]. Transactions in GIS, 2014, 18(1): 66-88.

[15] Spaccapietra S, Parent C, Damiani M L, et al. A conceptual view on trajectories[J]. Data & knowledge engineering, 2008, 65(1): 126-146.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[17] Liu H, Schneider M. Similarity measurement of moving object trajectories[C]//Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming. 2012: 19-22.

[18] Xiao X, Zheng Y, Luo Q, et al. Inferring social ties between users with human location history[J]. Journal of Ambient Intelligence and Humanized Computing, 2014, 5(1): 3-19.

[19] Furtado A S, Kopanaki D, Alvares L O, et al. Multidimensional similarity measuring for semantic trajectories[J].

Transactions in GIS, 2016, 20(2): 280-298.

## 附件

访问点类型：

（1）宾馆（0.4，0.8，0.3，0.2，0.2，0.2，0.3，0.2，0.2，0.7）

（2）饭店（0.8，0.7，0.2，0.2，0.4，0.3，0.3，0.3，0.3，0.4）

（3）咖啡馆（0.7，0.6，0.3，0.4，0.3，0.3，0.3，0.2，0.5，0.6）

（4）酒吧（0.8，0.8，0.8，0.1，0.5，0.4，0.3，0.2，0.2，0.5）

（5）茶座（0.7，0.6，0.6，0.3，0.2，0.3，0.3，0.2，0.4，0.4）

（6）公寓（0.2，0.6，0.2，0.2，0.2，0.2，0.3，0.2，0.5，0.7）

（7）住宅（0.1，0.2，0.1，0.2，0.1，0.1，0.3，0.1，0.5，0.9）

（8）公共浴室（0.8，0.7，0.6，0.1，0.2，0.3，0.3，0.3，0.2，0.4）

（9）理发店（0.6，0.8，0.3，0.3，0.3，0.2，0.3，0.3，0.2，0.3）

（10）美容院（0.5，0.9，0.2，0.1，0.3，0.2，0.3，0.3，0.2，0.4）

（11）电影院（0.7，0.8，0.8，0.2，0.3，0.3，0.3，0.4，0.1，0.5）

（12）游戏厅（0.7，0.7，0.9，0.2，0.5，0.4，0.3，0.3，0.1，0.4）

（13）KTV （0.8，0.8，0.9，0.2，0.5，0.4，0.3，0.3，0.1，0.5）

（14）景点（0.9，0.9，0.8，0.3，0.9，0.6，0.3，0.5，0.2，0.6）

（15）网吧（0.7，0.7，0.9，0.2，0.6，0.2，0.3，0.4，0.3，0.7）

（16）游乐园（0.8，0.9，0.9，0.2，0.9，0.6，0.3，0.5，0.2，0.7）

（17）体育馆（0.7，0.3，0.7，0.4，0.6，0.9，0.3，0.4，0.2，0.6）

（18）游泳馆（0.7，0.7，0.7，0.4，0.6，0.9，0.3，0.4，0.1，0.5）

（19）公园（0.8，0.1，0.6，0.3，0.8，0.8，0.3，0.5，0.2，0.5）

（20）健身房（0.6，0.8，0.5，0.4，0.6，0.9，0.3，0.4，0.3，0.4）

（21）广场（0.8，0.2，0.5，0.2，0.8，0.7，0.3，0.5，0.2，0.4）

（22）展览馆（0.7，0.3，0.6，0.5，0.7，0.5，0.3，0.4，0.4，0.5）

（23）博物馆（0.7，0.5，0.6，0.5，0.7，0.5，0.3，0.4，0.4，0.5）

（24）美术馆（0.7，0.5，0.6，0.6，0.7，0.3，0.3，0.4，0.5，0.5）

（25）图书馆（0.6，0.1，0.5，0.8，0.2，0.2，0.3，0.3，0.7，0.6）

（26）学校（0.8，0.1，0.1，0.9，0.3，0.3，0.3，0.2，0.8，0.8）

（27）幼儿园（0.7，0.1，0.1，0.8，0.3，0.4，0.3，0.2，0.5，0.7）

（28）商场（0.9，0.9，0.7，0.3，0.8，0.5，0.3，0.6，0.4，0.5）

（29）书店（0.6，0.6，0.4，0.7，0.6，0.3，0.3，0.4，0.5，0.6）

（30）超市（0.8，0.9，0.3，0.2，0.7，0.4，0.3，0.5，0.3，0.3）

（31）药店（0.6，0.8，0.1，0.2，0.4，0.2，0.3，0.4，0.2，0.2）

（32）车站（0.7，0.3，0.2，0.3，0.6，0.5，0.3，0.9，0.2，0.3）

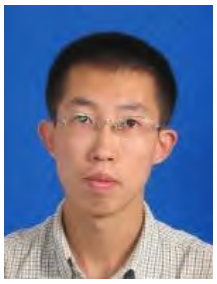
（33）医院（0.9，0.6，0.1，0.2，0.2，0.3，0.3，0.5，0.4，0.5）

（34）工厂（0.7，0.2，0.2，0.2，0.2，0.2，0.3，0.3，0.4，0.7）

（35）政府部门（0.5，0.1，0.1，0.4，0.1，0.2，0.3，0.2，0.9，0.8）

CAI Mingxin was born in 1996 in Shenyang, Liaoning Province. She received the bachelor's degree from Northeastern University, in 2018, where she is currently pursuing the master's degree with the Computer System Institute. Her research interests include time series data processing and machine learning.

蔡明昕(1996-)，女，辽宁沈阳市人。2018 年在东北大学获得学士学位，目前正在东北大学计算机科学与工程学院攻读硕士学位。主要研究领域为时序数据处理和机器学习。



SUN Jing was born in 1989. He received the bachelor degree in computer science from Northeastern University in 2012. He is currently a PhD candidate student in Northeastern University. His research interests include indoor localization, indoor spatial query and graph query.

孙晶(1989-)，辽宁沈阳人，2012 年与东北大学计算科学与技术专业取得学士学位，现在是东北大学博士研究生。主要研究方向包括室内定位、室内空间查询和图查询。

WANG Bin was born in 1972 in Shenyang, Liaoning Province. He obtained the doctor's degree from Northeastern University in 2008 and is currently an associate professor of Northeastern University. His main research fields are data management and data quality. As the project leader, he presided over 4 general projects of the National Natural Science Foundation, 1 general project of the National Natural Science Foundation Joint Fund, 1 sub-project of the National Key Basic Research and Development Plan (973 Plan), 1 key project funded by the Ministry of Education for Basic Research, 1 pre-research project and more than 10 other participating projects. More than 80 papers have been published, including 25 SCI searches, 67 EI searches, 34 first authors (including correspondence authors) and 20 CCF A English papers.

王斌(1972-)，男，辽宁沈阳市人，2008 年于东北大学获得博士学历，现任东北大学副教授，主要研究领域为数据管理，数据质量。作为项目负责人主持国家自然科学基金面上项目 4 项，国家自然科学基金联合基金面上项目 1 项，国家重点基础研究发展计划（973 计划）子课题项目 1 项，教育部基础科研资助重点项目 1 项，预研项目 1 项，其它参与项目 10 余项。发表论文 80 余篇，包括 SCI 检索 25 篇、EI 检索 67 篇，第一作者（含通信作者）34 篇， CCF A 类英文论文长文 20 篇。