一种基于用户轨迹的跨社交网络用户身份识别算法

陈鸿昶 徐 乾* 黄瑞阳 程晓涛 吴 铮
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要：针对现有的基于用户轨迹的跨社交网络用户身份识别算法未考虑用户轨迹中的位置访问顺序特征的缺点，该文提出一种基于Paragraph2vec的跨社交网络用户轨迹匹配算法(CDTraj2vec)。首先将用户轨迹转化为易于处理的网格化表示，并按照一定的时间粒度、距离尺度对原始的用户轨迹进行划分，使用户轨迹中的位置访问顺序特征易于抽取；然后利用Paragraph2vec算法中PV-DM模型抽取轨迹序列中位置访问顺序特征，得到用户轨迹的向量表示。最后通过用户轨迹向量判定轨迹是否匹配。在社交网络BrightKite上的实验结果表明，与基于位置访问频率或者基于轨迹间距离的方法相比，*F*值提高了2%～4%个百分点，所提算法能够有效地抽取出用户轨迹中的位置访问顺序特征，更加准确地实现了基于用户轨迹的跨社交网络用户身份识别。

关键词：社交网络；用户身份识别；轨迹相似度；Paragraph2vec
中图分类号：TP393; TP391.4 **文献标识码**：A **文章编号**：1009-5896(2018)11-2758-07
DOI: 10.11999/JEIT180130

User Identification Across Social Networks Based on User Trajectory

CHEN Hongchang XU Qian HUANG Ruiyang CHENG Xiaotao WU Zheng
(*National Digital Switching Engineering & Technological Research Center*, *Zhengzhou* 450002, *China*)

Abstract: The performance of trajectory based user identification is poor since the existing methods ignore the order feature of location sequence. To solve this problem, a Cross Domain Trajectory matching algorithm based on Paragraph2vec (CDTraj2vec) is proposed. Firstly, the user trajectory is transformed to the grid representation which is easy to handle. The PV-DM model in the Paragraph2vec algorithm is utilized for extracting order feature of location sequence in trajectory. Then the original user trajectories are divided by a certain time size and distance scale to construct a training sample suitable for training PV-DM model. The PV-DM model is trained by different types of training samples, and the vector representation of the user trajectories is obtained. Finally, the matching of the trajectory is determined by the user trajectory vector. Experimental results on BrightKite shows that the *F*-measure is improved by 2%～4% compared with the existing frequency based and distance based algorithm. The proposed algorithm can effectively extract the order feature of location sequence, and realize the trajectory based user identification across social networks.

Key words: Social networks; User identification; Trajectory similarity; Paragraph2vec

1 引言

在互联网时代，社交网络已成为人们生活中不可分割的一部分。社交网站Aboutme表明，一个用户通常在多个社交网站上注册账号来与不同的社交网络中的朋友进行交互，产生了丰富的社交用户信息。然而，各个社交网络账号间是孤立的没有联系的，因此用户的社交行为分散在多个社交网络中。跨网络用户身份识别，指的是将不同社交网络中属于同一个真实用户的账号关联起来，这一技术的解决能够为跨网络推荐、跨网络用户建模以及跨网络用户行为分析等应用提供全面的用户数据，实现对多源社交网络大数据的充分挖掘[1]。

近年来逐渐发展的地理位置采集和无线通信技术使得社交网络用户可以轻易地使用移动设备在发布的内容中加入地理位置标签。因为用户轨迹具有不容易模仿伪造的特点，文献[2]已经证明了用户移动轨迹的独特性，因此用户轨迹数据为跨社交网络用户身份的准确识别带来了全新技术途径。文献[3]基于两条用户轨迹中位置的共现频率，提出了一种处理多源位置数据的身份识别方法，缺点是参数过多，调整参数的过程非常繁琐。文献[4]将整个地图

收稿日期：2018-01-30；改回日期：2018-06-11；网络出版：2018-06-30
*通信作者： 徐乾 549529376@qq.com
基金项目：国家自然科学基金(61521003)
Foundation Item: The National Natural Science Foundation of China (61521003)

分成许多个网格，然后将每个用户的轨迹表示成若干小网格组成的序列，使用TF-IDF模型将每个用户的轨迹转化成向量，通过计算向量间余弦相似性得到用户轨迹间的相似性。文献[5]使用将每一个地理位置表示为该位置的语义位置对应的词，每条用户轨迹组成一篇由语义位置组成的文章，然后用LDA模型表示出每个用户的主题分布，最后计算分布间的KL散度得到轨迹相似性。文献[6]假设用户在一段时间内访问某个地点的次数服从泊松分布，进而得到两个账号属于同一个真实用户的概率形式的函数表示，最终通过优化目标函数得到最佳的匹配结果。文献[7]综合考虑轨迹的空间信息和时间信息，将原始的用户轨迹转化为三部图的来表示，最后通过求取三部图的最优划分，来得到最优的匹配方案。

纵观现有方法，在计算用户轨迹间的相似性时，主要是将轨迹看作地理位置的集合或者时空域中点的集合，然后使用基于频率或者基于共现次数的方法来计算位置集合之间的相似度，如果这个相似度超过一定阈值则认为这两个用户对应现实世界中同一个人。这些方法虽然取得了一定的效果，但是仍然存在以下问题：将用户访问的各个地理位置看作坐标点的集合进行处理，忽略了各个地理位置之间的相关性。例如用户在访问位置l之后，很有可能会访问另一个附近的位置l'。现有方法缺乏对这种轨迹序列信息的建模，没有充分挖掘用户访问地理位置顺序的潜在规律，导致算法精度下降。

针对上述问题，本文提出一种考虑位置访问顺序特征的跨社交网络用户轨迹匹配方法CDTraj2vec (Cross Domain Trajectory to vector algorithm)。考虑到基于深度学习的段落向量算法Paragraph2vec[8]在对段落进行向量化时考虑词序特征的良好表现，尝试使用Paragraph2vec方法抽取用户轨迹中的位置访问顺序特征；通过对用户轨迹按照一定的时间粒度、距离尺度的划分，构建出适用于训练PV-DM模型的轨迹序列，然后通过训练PV-DM模型得到用户轨迹在多个类型的训练样本上的向量化表示，进而得到蕴含用户位置访问顺序信息的轨迹向量。最终通过向量之间的相似性度量得到不同社交网络中用户轨迹间的相似性，从而进行用户身份识别。本文在真实社交网络中的时空数据上进行实验，验证了所提算法的有效性。

2　相关定义及数据预处理

2.1　相关定义

定义 1　用户轨迹：用户轨迹定义为随时间变化的GPS坐标点序列组成的集合，用$T=\left[p_1, p_2, \cdots, p_{|T|}\right]$表示，每一个坐标点$p_i$都包含3个属性$x, y, t$，$(x, y)$是坐标点$p_i$的GPS坐标，$t$是$(x, y)$被记录下来的时间。一个用户的在一个社交网络中产生的所有位置信息都记录在一条轨迹中。

定义 2　用户轨迹匹配：$T_A \in D_A$和$T_B \in D_B$是来自两个不同的时空数据集(D_A和D_B)的两条轨迹。如果T_A和T_B是同一个真实用户产生的，则称T_A和T_B互相匹配。本文的最终目的是尽可能多且准确地识别出匹配的轨迹对。

2.2　数据预处理

2.2.1 原始数据的网格表示　原始的时空数据集中存储的是2维地理空间上的GPS坐标点，直接处理这些数据点会遇到样本稀疏的问题，一种简单的解决方法是将原始GPS坐标点转化为网格表示。首先根据时空数据集中GPS坐标点的地理位置在地图上定义一个矩形的经纬度边界，这个矩形包含了所有时空数据集中的GPS坐标点。矩形边界对应的纬度范围是$(\mathrm{lat}_1, \mathrm{lat}_2)$，经度范围是$(\mathrm{lon}_1, \mathrm{lon}_2)$。然后根据需要的精度定义小网格的行数$r$和列数$c$(行数和列数越大对应的精度越高)。任意一个GPS坐标点$(\mathrm{lat}, \mathrm{lon})$都可以按照式(1)～式(3)转化为小网格编号ci来表示。

$$\mathrm{cx}=\left\lfloor \frac{(\mathrm{lat}-\mathrm{lat}_1)\, r}{\mathrm{lat}_2-\mathrm{lat}_1} \right\rfloor \tag{1}$$

$$\mathrm{cy}=\left\lfloor \frac{(\mathrm{lon}-\mathrm{lon}_1)\, c}{\mathrm{lon}_2-\mathrm{lon}_1} \right\rfloor \tag{2}$$

$$\mathrm{ci}=((c(\mathrm{cx}-1))+\mathrm{cy}) \tag{3}$$

其中，$\lfloor\cdot\rfloor$是向下取整函数，cx是小网格所在的行号，cy是小网格所在的列号。

2.2.2 训练样本构建方法　真实社交网络中用户产生的轨迹通常是相对稀疏的，一些用户轨迹中前后相邻的小网格之间并不存在很强的关联，例如某些用户轨迹片段中前后相邻的小网格相距很远或者相隔时间很长，这使得位置访问顺序特征难以显现，直接对这种前后位置相关性较弱的轨迹序列进行建模会导致用户轨迹向量表示精度的下降。因此需要对用户轨迹进行进一步处理，筛选出能够体现用户访问顺序特征的轨迹片段，选取相邻小网格之间相关性较强的轨迹序列作为下一步模型的训练样本。本文采用3种不同的训练样本构建方法：第1种构建方法是按时间划分轨迹，这种划分方法假设用户在Δt时间内访问的位置之间具有一定的关联。一段有效的位置序列定义如下：用户轨迹为$T=[c_{t_1}, c_{t_2}, \cdots, c_{t_m}]$，称$T$的子序列$T_s=[c_{t_i}, c_{t_{i+1}}, \cdots, c_{t_{i+l}}]$为有效序列如果$T_s$满足：(1)$t_{i+l}-t_i \le \Delta t$。(2)不存在$T$的一个子序

列，它真包含T_s且满足条件(1)。图1描述了一个按时间划分轨迹从而得到有效的位置序列的例子，用户轨迹$T=[c_{t_1}, c_{t_2}, c_{t_3}, c_{t_4}, c_{t_5}]$, $\Delta t=5$ h，从中可以得到1段有效的位置序列$\mathrm{ET}=[T_{S_1}, T_{S_2}, T_{S_3}]$。

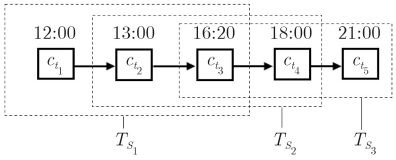

图 1 按时间划分轨迹

第2种构建方法是按照距离进行轨迹划分。一段轨迹序列中任意两个前后相邻的位置的距离小于Δd时，则认为这个位置序列是有效的。图2描述了一个按距离划分轨迹从而得到有效的轨迹序列的例子，用户轨迹为$T=[c_{t_1}, c_{t_2}, c_{t_3}, c_{t_4}, c_{t_5}]$, $\Delta d=7$ km，从中可以得到2段有效的轨迹序列$\mathrm{ET}=[T_{S_1}, T_{S_2}]$。

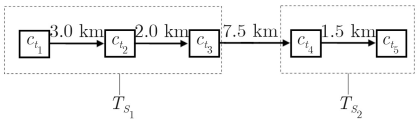

图 2 按距离划分轨迹

第3种构建方法是在时空域构建训练样本。时空域的结构如图3所示，纵轴表示网格id，横轴表示一天的4个时间段，$\Delta t_1, \Delta t_2, \Delta t_3, \Delta t_4$分别代表：0～6点，6～12点，12～18点，18～24点。图3举例描述了在时空域上构建训练样本的过程，用户轨迹为$T=\left[c_{t_1}^{\mathrm{id}_1}, c_{t_2}^{\mathrm{id}_2}, c_{t_3}^{\mathrm{id}_3}, c_{t_4}^{\mathrm{id}_4}\right]$，上标$\mathrm{id}_1, \mathrm{id}_2, \mathrm{id}_3, \mathrm{id}_4$为小网格id，下标$t_1, t_2, t_3, t_4$为它们被记录的时间。与之前两种方法一样，每次选取尽可能长的时空域序列作为训练样本。在时空域上构建的轨迹序列描述了用户访问各个位置的时间偏好，是一种将时间和空间信息结合起来的模型，最大限度地降低了原始轨迹信息的损耗。

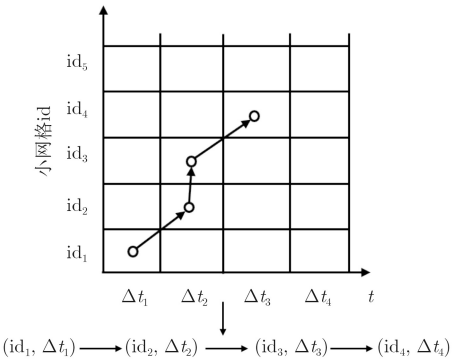

图 3 在时空域上构建位置序列

通过上述3种方法可以得到3种不同类型的训练样本，训练样本包含的轨迹序列中前后相邻的小网格具有一定的相关性，这使得用户轨迹中的位置访问顺序特征能够很容易地被抽取出来，为下一步轨迹建模打下基础。

3 基于用户轨迹的身份识别算法

3.1 基于Paragraph2vec的轨迹模型构建

Paragraph2vec是Le等人[8]提出的段落向量模型，其目标是得到段落的向量表示。Paragraph2vec算法中提出了一个重要的模型PV-DM(Distributed Memory model of Paragraph Vectors)，PV-DM模型通过段落pg_i以及词wd的上下文ct (wd)来预测词wd出现的概率，具体地，给定一个段落pg_i，以及其中的部分单词序列$W_1^n=(\mathrm{wd}_0, \mathrm{wd}_1, \cdots, \mathrm{wd}_n)$，其中$\mathrm{w}_i \in V$($V$是词典)，将$\mathrm{pg}_i$以及所有单词都采用向量表示，然后通过段落向量以及词向量构造目标函数$P(\mathrm{wd}_n|\mathrm{wd}_0, \mathrm{wd}_1, \cdots, \mathrm{wd}_{n-1}, \mathrm{pg}_i)$，最后通过训练样本训练模型并求解得到使得目标函数最大化的段落向量。

用户运动轨迹与文本句式具有高度的结构相似性，均为由离散状态的基本元构成的连续序列[9]。利用Paragraph2vec中对一段文本进行向量化的类似过程，可将每一个小网格视作“单词”，将每一条用户轨迹对应的小网格序列视作“一个段落”，从而将Paragraph2vec算法中的PV-DM模型应用到用户轨迹特征的抽取(图4)。用于轨迹特征抽取PV-DM模型的结构由输入层、投影层和输出层组成，下面以一个滑动窗口内的轨迹序列($[c_{t_{i-w+1}}, c_{t_{i-w+2}}, \cdots, c_{t_{i-1}}], c_{t_i}$)作为输入进行说明，窗口大小为$w$，轨迹序列产生于用户轨迹$T$。

输入层：出现在小网格c_{t_i}之前的$w-1$个小网格$c_{t_{i-w+1}}, c_{t_{i-w+2}}, \cdots, c_{t_{i-1}}$对应的向量序列$\mathbf{vc}(c_{t_{i-w+1}})$, $\mathbf{vc}(c_{t_{i-w+2}}), \cdots, \mathbf{vc}(c_{t_{i-1}}) \in \mathbb{R}^d$，以及用户轨迹$T$对应的向量$\mathbf{vc}(\mathrm{id}(T)) \in \mathbb{R}^d$, $\mathrm{id}(T)$为产生轨迹T的用户id，向量的初始值通过随机初始化得到。

投影层：将输入层的w个小网格对应的向量以及1个用户轨迹向量进行求和操作，生成一个d维向量，即$\boldsymbol{S}=\mathbf{vc}(\mathrm{id}(T))+\sum_{l=i-w+1}^{i-1}\mathbf{vc}(c_{t_l}) \in \mathbb{R}^d$。

输出层：采用与Paragraph2vec输出层相同的hierarchical softmax方法[10]，其结构为一颗二叉Huffman树，将所有用户轨迹中出现过的小网格作为叶子节点，以各个小网格在训练集中出现的次数作为权值构造Huffman树，在每个非叶节点上设置一个参数待定的二分类器，从根节点到叶子节点可看作多个二分类拟合多分类的过程。通过Huff-

man树可实现概率$p(c_{t_i}|\boldsymbol{S})$的构造，即已知某一用户及其产生的小网格序列，预测下一个小网格为c_{t_i}的概率。

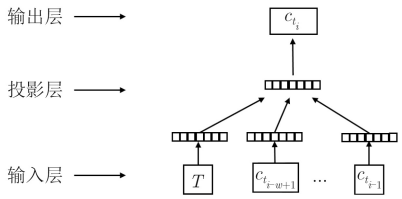

图 4 用于轨迹特征抽取的PV-DM模型结构图

上述结构中，每一条用户轨迹被映射为唯一的一个向量$\mathbf{vc}(\mathrm{id}(T))\in\mathbb{R}^d$，每一个小网格同样也被映射为唯一的一个向量。模型的输入通过在用户轨迹上构建长度为w的滑动窗口得到，同一用户轨迹生成的多组模型输入共享着同一轨迹向量，不同用户轨迹之间采用不同用户轨迹向量。小网格向量与之不同，它在不同的用户轨迹之间是共享的，例如，最终得到的小网格c对应的向量$\mathbf{vc}(c)$对于所有的用户轨迹都是相同的。

基于上述模型，可假设一段轨迹序列T包含n个小网格$c_{t_1}, c_{t_2}, \cdots, c_{t_n}$，则构建式(4)所示的概率函数：

$$f=\prod_{j=w}^{n} p\left(c_{t_j}\middle|\, c_{t_{j-w+1}}, c_{t_{j-w+2}}, \cdots, c_{t_{j-1}}, \mathrm{id}(T)\right) \tag{4}$$

假设某种训练样本中有m条用户轨迹，每条用户轨迹中包含n_i个小网格，则构建模型的目标函数为

$$\mathrm{ft}=\prod_{i=1}^{m}\left(\prod_{j=w}^{n_i}\lg p\left(c_{t_j}\middle|\, c_{t_{j-w+1}}, c_{t_{j-w+2}}, \cdots, c_{t_{j-1}}, \mathrm{id}(T_i)\right)\right) \tag{5}$$

其对数似然为

$$F=\lg \mathrm{ft}=\sum_{i=1}^{m}\left(\sum_{j=w}^{n_i}\lg p\left(c_{t_j}\middle|\, c_{t_{j-w+1}}, c_{t_{j-w+2}}, \cdots, c_{t_{j-1}}, \mathrm{id}(T_i)\right)\right) \tag{6}$$

模型求解与优化采用随机梯度法迭代求取目标函数最优解[11]。

得到的用户轨迹向量可以用来预测用户下一个要访问的小网格，例如：对于用户轨迹T，通过训练得到其轨迹向量$\mathbf{vc}(\mathrm{id}(T))$，已知当前该用户已经访问了$i$个小网格$[c_{t_1}, c_{t_2}, \cdots, c_{t_i}]$，则该用户下一步访问小网格$c_{t_{i+1}}$的概率为$p(c_{t_{i+1}}|\, c_{t_{i-w+2}}, c_{t_{i-w+3}}, \cdots, c_{t_i}, \mathrm{id}(T))$。用户轨迹向量可以用来准确地求取用户访问各个位置的转移概率，而位置转移概率是对用户访问小网格顺序的合理刻画，因此得到的用户轨迹向量中蕴含了表示用户访问小网格顺序的位置访问顺序特征。

3.2 CDTraj2vec算法

基于上述分析，基于Paragraph2vec的用户轨迹匹配算法CDTraj2vec的伪代码如表1的算法1所示。第3行中的$\boldsymbol{\Phi}_{\mathrm{t}}$, $\boldsymbol{\Phi}_{\mathrm{d}}$和$\boldsymbol{\Phi}_{\mathrm{st}}$分别对应3种训练样本训练模型得到的用户轨迹表示；第1行—第4行主要包含的步骤有未知变量的初始化，以及将原始数据转化为小网格，并且按照3种训练样本构建方法得到训练数据。第5行—第9行是利用训练样本训练PV-DM模型来更新用户轨迹向量。第10行—第12行是将3种不同的用户轨迹向量进行拼接，得到多维时空数据下信息更加完整的用户轨迹向量表示。

表 1　CDTranj2vec伪代码

算法1 基于Paragraph2vec的用户轨迹匹配算法
CDTraj2vec(D_A, D_B, w, d)

输入：时空数据集D_A, D_B，窗口大小w，用户轨迹向量的维数d
输出：匹配结果
(1)将原始时空数据集D_A, D_B转化为小网格表示，转化后的用户轨迹为D_A^c, D_B^c
(2)$D_{AB}^c = D_A^c + D_B^c$
(3)随机初始化$\boldsymbol{\Phi}_{\mathrm{t}}, \boldsymbol{\Phi}_{\mathrm{d}}, \boldsymbol{\Phi}_{\mathrm{st}} \in \mathbb{R}^{(|D_A|+|D_B|)\times d}$
(4)按照2.2.2节中的3种训练样本构建方法，由D_{AB}^c构建出3种训练样本集合$\mathrm{Sp_t}$, $\mathrm{Sp_d}$, $\mathrm{Sp_{st}}$
(5) for each i in ['t', 'd', 'st'] do:
(6) for $j=0$ to $|\mathrm{Sp}_{ij}|$ do:
//$\mathrm{Sp}_{ij}=\left\{\text{'ts'}:\left[c_{t_1}, c_{t_2}, \cdots, c_{t_{|\mathrm{Sp}_{ij}|}}\right], \text{'ui'}:\mathrm{id}(T)\right\}$
(7) PV-DM($\boldsymbol{\Phi}_i, \mathrm{Sp}_{ij}, w$)
(8) end for
(9) end for
(10) for $i=0$ to $|D_A|+|D_B|$ do:
(11) $\boldsymbol{\Phi}(i,:)=\mathrm{concatenate}(\boldsymbol{\Phi}_{\mathrm{t}}(i,:), \boldsymbol{\Phi}_{\mathrm{d}}(i,:), \boldsymbol{\Phi}_{\mathrm{st}}(i,:))$
(12) end for
(13) rs = matching_strategy($\boldsymbol{\Phi}$)
(14) return rs

第13行是在得到用户轨迹向量表示之后，利用文献[12]中的匹配策略，得到最终的匹配结果。文献[12]中的匹配策略包括3个步骤：账号选择、账号匹配以及剪枝过滤。在账号选择阶段需要设置筛选条件，本文将条件C设置为：与待匹配轨迹T_A至少在2个不同的小网格中产生共现。在账号匹配阶段，将账号间的相似性度量修改为轨迹向量之间加权的余弦相似度。剪枝过滤部分与文献[12]相同。

PV-DM模型对应的算法如表2的算法2所示，

其中Sp_{ij}表示按照方法i生成的训练样本中的第j个样本，Sp_{ij}.ts表示一段有效的轨迹序列，Sp_{ij}.ui为产生该轨迹序列的用户id，α为学习率。第2行为PV-DM模型的输入，第3行的J为待优化的目标函数，第4行—第7行为用户轨迹向量以及小网格向量的更新过程。

表 2 PV-DM ($\boldsymbol{\Phi}_i, \mathrm{Sp}_{ij}, w$)算法

算法2 PV-DM$(\boldsymbol{\Phi}_i, \mathrm{Sp}_{ij}, w)$

输入：按照方法i生成的训练样本在训练模型的过程中待更新的用户轨迹向量$\boldsymbol{\Phi}_i$，按照方法i生成的训练样本中的第j个样本Sp_{ij}，窗口大小w

输出：无

(1)for $k=w$ to $|\mathrm{Sp}_{ij}|$ do:

(2) $\boldsymbol{S}=\boldsymbol{\Phi}_i(\mathrm{Sp}_{ij}.\mathrm{ui})+\sum_{l=k-w+1}^{w-1}\mathbf{vc}(c_{t_l})$

(3) $J=-\lg(c_{t_k}|\boldsymbol{S})$

(4) for $l=k-w+1$ to $w-1$ do:

(5) $\mathbf{vc}(c_{t_l})=\mathbf{vc}(c_{t_l})-\alpha\times\dfrac{\partial J}{\partial \boldsymbol{S}}$

(6) end for

(7) $\boldsymbol{\Phi}_i(\mathrm{Sp}_{ij}.\mathrm{ui})=\boldsymbol{\Phi}_i(\mathrm{Sp}_{ij}.\mathrm{ui})-\alpha\times\dfrac{\partial J}{\partial \boldsymbol{S}}$

(8)end for

4 实验与分析

4.1 数据集

本文在两个数据集上对所提算法进行验证，第1个数据集来自微软亚洲研究院的GeoLife项目[13]。数据集信息如表3所示。由于真实的多源用户轨迹数据难以获取，本文将单个时空数据集划分为两个部分，然后将这两个部分看作两个社交网络来进行身份识别。在将原始GPS数据转化为停留点之后，将每一条用户轨迹按照停留点的个数平均划分为两个部分D_A和D_B，例如一条用户轨迹中包含100个按时间顺序排列的停留点，那么就将前50个划分到D_A中，后50个划分到D_B中。这个数据集由于采样频率较高(每1～5 s或5～10 m记录一次GPS坐标)，因此在其上进行身份识别较为容易。第2个数据集来自基于位置服务的社交网络BrightKite[14]，数据集信息同样包含在表3中。实验过程中排除了坐标点少于5的用户轨迹。与Geolife数据集一样，采用同样的方法将该时空数据划分为两部分D_C和D_E。此数据集采样频率没有Geolife高，在其上进行身份识别相对困难。

4.2 评价指标

本文采用准确率(pc)、召回率(rc)以及综合评价指标F1作为衡量算法性能的评价标准。具体定义如式(7)～式(9)所示：

$$\mathrm{pc}=\mathrm{tp}/(\mathrm{tp}+\mathrm{fp}) \tag{7}$$

$$\mathrm{rc}=\mathrm{tp}/(\mathrm{tp}+\mathrm{fn}) \tag{8}$$

$$F=2\times\mathrm{pc}\times\mathrm{rc}/(\mathrm{pc}+\mathrm{rc}) \tag{9}$$

式中，tp是指被算法判定为匹配且判断正确的轨迹对数，fp是指被算法判定为匹配但判断错误的轨迹对数，fn是指被算法判定为不匹配但实际上匹配的轨迹对数。F值是准确率和召回率的调和平均数，是算法性能的综合评价指标。

4.3 参数设置

本文所提参数的默认值如表4所示。利用单个训练本进行实验，选取最佳结果对应的Δt和Δd。小网格的行数r和列数c的选取，是通过预先设定小网格的边长sl，结合网格经纬度范围计算得到的。

4.4 对比算法

实验中采用4种基准算法来与CDTraj2vec算法进行对比，具体如下：

(1)k-BCT: k-BCT (k Best Connected Trajectories)[15]是一种有效的不考虑位置访问顺序的轨迹相似性度量方法，对于两条用户轨迹T_A和T_B，k-BCT相似度如式(10)所示，其中c_a,c_b表示转化后的小网格，$\mathrm{Dis}(c_a, c_b)$表示小网格c_a与c_b的中心在地图上的直线距离。

$$\mathrm{k-BCT}(T_A, T_B)=\sum_{c_a\in T_A}\exp\left(-\min_{c_b\in T_B}\mathrm{Dis}(c_a, c_b)\right) \tag{10}$$

表 3 数据集信息

数据来源	Geolife		BrightKite	
划分后的数据集	D_A	D_B	D_C	D_E
用户轨迹数	182	182	2971	2974
坐标点个数	16128	16802	332326	328031
不同的小网格个数	3140	3517	73255	72098
网格范围	纬度：39°～41° 经度：115°～117°	纬度：39°～41° 经度：115°～117°	纬度：25°～49° 经度：–130°～–70°	纬度：25°～49° 经度：–130°～–70°
时间范围	2007-04～2012-08	2007-04～2012-08	2008-03～2010-10	2008-03～2010-10

表 4　文中所提参数的默认值

变量	值	说明
sl	100 m	小网格的边长
$\mathrm{lat}_1, \mathrm{lat}_2$	29°N, 49°N	小网格的纬度范围
$\mathrm{lon}_1, \mathrm{lon}_2$	70°W, 30°W	小网格的经度范围
r	53374	小网格行数
c	120932, 87541	最低纬度(29°N)和最高纬度(49°N)上的小网格列数
Δt	4 h(Geolife), 3 h(BrightKite)	按时间划分轨迹时设定的时间阈值
Δd	1 km	按距离划分轨迹时设定的距离阈值

(2)TF-IDF：文献[4]提出了一种基于TF-IDF(Term Frequency-Inverse Document Frequency)的跨域轨迹相似性计算方法：将原始用户轨迹转化为小网格表示，把小网格看作单词，把每条用户轨迹看作一篇文章，采用TF-IDF方法将每条用户轨迹转化为向量表示，从而得到轨迹相似性。TF-IDF方法综合考虑了用户轨迹中小网格的独特性以及各个小网格出现的次数，但是并未考虑轨迹中小网格之间的顺序特征。

(3)小网格顺序打乱数据集上的CDTraj2vec-sf算法：为了进一步验证本文算法能够抽取出轨迹中的位置访问顺序特征，以及用户位置访问顺序特征被用来进行用户轨迹匹配的合理性，本文将用户轨迹中的小网格打乱：将用户轨迹中的各个小网格的时间固定，将小网格打乱之后在对应到产生的时间上。

(4)不划分训练样本的CDTraj2vec-wcet算法：为了验证本文中样本构建算法的有效性，设置一种不划分训练本的CDTraj2vec算法，称之为CDTraj2vec-wcet(CDTraj2vec Without constructing effective training set)算法，CDTraj2vec-wcet不对用户轨迹进行划分，直接使用用户轨迹对应的小网格序列来训练PV-DM模型。算法的参数与CDTraj2vec保持一致。

4.5　结果分析

实验结果如表5，表6所示，CDTraj2vec相比其他算法在不同数据集上均拥有较好性能。具体原因在于，相比于CDTraj2vec, k-BCT与TF-IDF方法均没有考虑用户访问位置的顺序以及时间信息；CDTraj2vec-sf打乱了轨迹点的顺序，使得轨迹位置访问顺序特征变得不明显甚至不存在，用户潜在的行为模式难以被识别，所以其实验结果并不理想；CDTraj2vec-wcet由于没有对用户轨迹进行有效划分，模型难以从样本中抽取出位置访问顺序特征，导致算法效果不理想。

表 5　Geolife上的实验结果($\Delta t = 4$ h, $\Delta d = 1$ km)

算法	准确率(pc)	召回率(rc)	F值
k-BCT	0.9080	0.9530	0.9299
TF-IDF	0.9651	0.9185	0.9412
CDTraj2vec-sf	0.9241	0.8462	0.8834
CDTraj2vec-wcet	0.9151	0.8694	0.8917
CDTraj2vec (使用训练样本A, B, C)	**0.9888**	**0.9899**	**0.9893**

表 6　BrightKite上的实验结果($\Delta t = 3$ h, $\Delta d = 1$ km)

算法	准确率(pc)	召回率(rc)	F值
k-BCT	0.8802	0.9238	0.9015
TF-IDF	0.8741	0.9164	0.8947
CDTraj2vec-sf	0.8841	0.8457	0.8647
CDTraj2vec-wcet	0.8645	0.8108	0.8367
CDTraj2vec (使用训练样本A, B, C)	**0.9465**	0.9245	0.9354
CDTraj2vec (只使用训练样本A, B)	0.9440	**0.9281**	**0.9360**

本文的CDTraj2vec算法从用户潜在的移动模式出发，对用户位置访问顺序特征进行建模，在原始用户轨迹的基础上构建体现空间信息以及时间信息的训练样本，使得模型更加容易地从中抽取出位置访问顺序特征，从而实验结果更加精确。实验结果显示，在两个数据集上，单一的训练样本的F值均高于CDTraj2vec-wcet，验证了3种样本构建方法的有效性。此外，在Geolife数据集上，同时使用3种训练样本时，准确率、召回率和F值达到最大的0.9888，0.9899和0.9893，比使用单独一种或两种类型的训练样本效果更好，验证了本文算法的有效性。在BrightKite数据集上，只使用训练样本A，B时(分别将按时间划分轨迹得到的训练样本、按距离划分轨迹得到的训练样本以及在时空域上构建的训练样本称为训练样本A, B, C), F值达到最大的0.9360，在此基础上加入训练样本C时，F值从0.9360降到了0.9354，其原因是在时空域上构建的训练样本具有稀疏性，最终导致F值的降低，但与此同时，F值的略微降低换来的是准确率达到最大，这使得本文算法适用于对准确率要求较高的应用场景。

5　结束语

针对现有轨迹匹配算法未考虑用户轨迹中的位置访问顺序特征的缺点，本文提出了基于Paragraph2vec的轨迹匹配算法CDTraj2vec。利用Paragraph2vec算法中的PV-DM模型抽取用户轨迹

中的位置访问顺序特征。通过对原始轨迹按照不同的时间粒度以及距离尺度进行划分，得到3种适用于训练PV-DM模型的训练样本。将各个训练样本训练模型得到的轨迹向量进行拼接作为最终的轨迹向量表示。最终通过向量相似性计算用户轨迹相似性。实验结果表明所提算法在准确率上具有较为明显的优势。由于本文算法在划分小网格时没有考虑其在地图上的语义位置，有时表达同一个语义位置的两个坐标点，对应的可能是不同的小网格，所以后续研究可以集中在：如何在数据预处理部分对原始的时空数据进行准确地网格化，使其网格化过后信息量损失最小。

参 考 文 献

[1] 桑基韬, 路冬媛, 徐常胜. 基于共同用户的跨网络分析: 社交媒体大数据中的多源问题[J]. 科学通报, 2014, 59(36): 3554–3560. doi: 10.1360/n972014-00292.
SANG Jitao, LU Dongyuan, and XU Changsheng. Overlapped user-based cross-network analysis: Exploring variety in big social media data[J]. *Chinese Science Bulletin*, 2014, 59(36): 3554–3560. doi: 10.1360/n972014-00292.

[2] GONZÁLEZ M C, HIDALGO C A, and BARABÁSI A L. Understanding individual human mobility patterns[J]. *Nature*, 2008, 453(7196): 779–782. doi: 10.1038/nature06958.

[3] CAO Wei, WU Zhengwei, WANG Dong, *et al*. Automatic user identification method across heterogeneous mobility data sources[C]. IEEE International Conference on Data Engineering, Helsinki, Finland, 2016: 978–989. doi: 10.1109/ICDE.2016.7498306.

[4] HAO Tianyi, ZHOU Jingbo, CHENG Yunsheng, *et al*. User identification in cyber-physical space: A case study on mobile query logs and trajectories[C]. GIS'16 Proceedings of the 24th ACM International Conference on Advances in Geographic Information Systems, California, USA, 2016: 1–4. doi: 10.1145/2996913.2997017.

[5] HAN Xiaohui, WANG Lianhai, XU Shujiang, *et al*. Linking social network accounts by modeling user spatiotemporal habits[C]. IEEE International Conference on Intelligence and Security Informatics, Beijing, China, 2017: 19–24. doi: 10.1109/ISI.2017.8004868.

[6] RIEDERER C, KIM Y, CHAINTREAU A, *et al*. Linking users across domains with location data: Theory and validation[C]. WWW'16 Proceedings of the 25th International Conference on World Wide Web. Montréal, Canada, 2016: 707–719. doi: 10.1145/2872427.2883002.

[7] HAN Xiaohui, WANG Lianhai, XU Lijuan, *et al*. Social Media account linkage using user-generated geo-location data[C]. Intelligence and Security Informatics, Tucson, USA, 2016: 157–162. doi: 10.1109/ISI.2016.7745460.

[8] LE Q and MIKOLOV T. Distributed representations of sentences and documents[C]. International Conference on International Conference on Machine Learning, Beijing, China, 2014: II-1188.

[9] 殷浩腾, 刘洋. 基于社交属性的时空轨迹语义分析[J]. 中国科学: 信息科学, 2017, 47(8): 1051–1065. doi: 10.1360/N112016-00310.
YIN Haoteng and LIU Yang. Semantic analysis of spatial temporal trajectory in LBSNs[J]. *Scientia Sinica* (*Informationis*), 2017, 47(8): 1051–1065. doi: 10.1360/N112016-00310.

[10] MIKOLOV T, SUTSKEVER I, CHEN K, *et al*. Distributed representations of words and phrases and their compositionality[C]. International Conference on Neural Information Processing Systems, Daegu, SUKO. 2013: 3111–3119.

[11] BOYD S and VANDENBERGHE L. Convex Optimization[M]. Cambridge: Cambridge University Press, 2004: 466–468. doi: 10.1017/CBO9780511804441.

[12] 吴铮, 于洪涛, 刘树新, 等. 基于信息熵的跨社交网络用户身份识别方法[J]. 计算机应用, 2017, 37(8): 2374–2380. doi: 10.11772/j.issn.1001-9081.2017.08.2374.
WU Zheng, YU Hongtao, LIU Shuxin, *et al*. User identification across multiple social networks based on information entropy[J]. *Journal of Computer Applications*, 2017, 37(8): 2374–2380. doi: 10.11772/j.issn.1001-9081.2017.08.2374.

[13] ZHENG Yu, XIE Xing, and MA Weiying. GeoLife: A collaborative social networking service among user, location and trajectory[J]. *Bulletin of the Technical Committee on Data Engineering*, 2010, 33(2): 32–39.

[14] CHAINTREAU A. COMS 6998: Social Networks[EB/OL]. http://socialnetworksfall14.wikischolars.columbia.edu/, 2014-10/2017-12.

[15] CHEN Zaiben, SHEN Hengtao, ZHOU Xiaofang, *et al*. Searching trajectories by locations: An efficiency study[C]. International Conference Proceedings, Association for Computing Machinery, Indianapolis, Indiana, USA, 2010: 255–266. doi: 10.1145/1807167.1807197.

陈鸿昶：男，1964年生，教授，研究方向为通信与信息系统.

徐 乾：男，1993年生，硕士生，研究方向为社交网络挖掘、机器学习.

黄瑞阳：男，1986年生，副研究员，研究方向为网络大数据处理与分析.

程晓涛：男，1986年生，博士生，研究方向为社交网络挖掘、机器学习.

吴 铮：男，1992年生，博士生，研究方向为复杂网络、网络大数据处理与分析.