计算机科学
COMPUTER SCIENCE

融合多类时空轨迹特征的跨网络用户身份识别

刘红，朱焱，李春平

引用本文

刘红，朱焱，李春平.融合多类时空轨迹特征的跨网络用户身份识别[J]．计算机科学, 2023, 50(3): 114-120.

LIU Hong, ZHU Yan, LI Chunping. Cross-network User Identification Based on Multiple Spatio-Temporal Trajectory Features [J]. Computer Science, 2023, 50(3): 114-120.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于双向长短时记忆网络的企业弹性能力预测模型

Prediction Model of Enterprise Resilience Based on Bi-directional Long Short-term Memory Network

计算机科学, 2022, 49(11): 197-205. https://doi.org/10.11896/jsjkx.210900195

基于异构网络表征学习的作者学术行为预测

Author's Academic Behavior Prediction Based on Heterogeneous Network Representation Learning

计算机科学, 2022, 49(9): 76-82. https://doi.org/10.11896/jsjkx.210900078

基于时空注意力克里金的边坡形变数据插值方法

Spatio-Temporal Attention-based Kriging for Land Deformation Data Interpolation

计算机科学, 2022, 49(8): 33-39. https://doi.org/10.11896/jsjkx.210600161

融合RACNN和BiLSTM的金融领域事件隐式因果关系抽取

Implicit Causality Extraction of Financial Events Integrating RACNN and BiLSTM

计算机科学, 2022, 49(7): 179-186. https://doi.org/10.11896/jsjkx.210500190

基于CNN-LSTM的卫星云图云分类方法研究

Study on Cloud Classification Method of Satellite Cloud Images Based on CNN-LSTM

计算机科学, 2022, 49(6A): 675-679. https://doi.org/10.11896/jsjkx.210300177

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200287

融合多类时空轨迹特征的跨网络用户身份识别

刘 红[1] 朱 焱[1] 李春平[2]
1 西南交通大学计算机与人工智能学院 成都 611756
2 清华大学软件学院 北京 100091
(leuh1997@my.swjtu.edu.cn)

摘 要 随着位置社交网络的蓬勃发展,用户移动行为数据得到极大丰富,推动了基于时空数据的身份识别问题的相关研究。跨位置社交网络的用户身份识别,强调学习不同平台时空序列间的相关性,旨在发现同一用户在不同平台的注册账号。为解决现有研究面临的数据稀疏、低质量和时空不匹配问题,提出了一种融合双向时空依赖和时空分布的识别算法 UI-STDD。该算法主要包含 3 个模块:时空序列模块通过结合成对注意力的双向长短时记忆网络来刻画用户移动模式;时间偏好模块从粗、细两个粒度定义用户个性化模式;空间位置模块挖掘位置点的局部和全局信息,量化空间邻近性。基于上述模块得到的用户轨迹对特征,UI-STDD 利用多层前馈网络判断跨网络的两个账户是否对应于现实中的同一个人。为验证 UI-STDD 的可行性和有效性,在 3 组公开的数据集上进行了实验。实验结果表明,所提算法能够提高基于时空数据的用户身份识别率,F1 值平均高于最优对比方法 10%以上。

关键词:用户身份识别;时空数据;移动模式;时间偏好;长短时记忆网络
中图法分类号 TP301

Cross-network User Identification Based on Multiple Spatio-Temporal Trajectory Features

LIU Hong[1],ZHU Yan[1] and LI Chunping[2]
1 School of Computing and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
2 School of Software,Tsinghua University,Beijing 100091,China

Abstract With the flourishing of location-based social networks,users' mobile behavior data has been greatly enriched,which promotes the research on user identification based on spatio-temporal data. User identification in cross-location social networks emphasizes learning the correlation between time and space sequences of different platforms,aiming at discovering the accounts registered by the same user on different platforms. In order to solve the problems of data sparsity,low quality and spatio-temporal mismatch faced by existing researches,a recognition algorithm UI-STDD combining bidirectional spatio-temporal dependence and spatio-temporal distribution is proposed. The algorithm mainly consists of three modules:the space-time sequence module is combined with the bidirectional long short-term memory network of paired attention to describe user movement patterns;the time preference module defines the user personalized mode from coarse and fine granularity;the spatial location module mines local and global information of location points to quantify spatial proximity. Based on the user trajectory pair features obtained by the above modules,a multi-layer feedforward network is used in UI-STDD to distinguish whether two accounts across the network correspond to the same person in real life. To verify the feasibility and effectiveness of UI-STDD,experiments are carried out on three publicly available datasets. Experimental results show that the proposed algorithm can improve the user identification rate based on spatio-temporal data,and the average F1 value is more than 10% higher than the optimal comparison method.

Keywords User identification,Spatio-Temporal data,Mobile mode,Time preference,Long short-term memory

1 引言

随着移动设备的普及和 GPS 技术的发展,人们时常在各类网站、平台分享实时位置。位置社交网络,如大众点评、Instagram 和 Twitter,在此背景下产生了大量的时空数据。基于时空数据分析用户轨迹模式,极大地丰富了移动社交网络挖掘应用,如地点推荐[1]、交通方式识别[2]、异常轨迹检测[3]以及用户身份识别[4]等。

用户身份识别,也被称为用户对齐、用户匹配,旨在发现注册在不同网络的同一用户[5]。基于时空数据的用户身份

到稿日期:2021-12-27 返修日期:2022-04-14
基金项目:四川省科技计划(2019YFSY0032)
This work was supported by the Sichuan Science and Technology Project(2019YFSY0032).
通信作者:朱焱(yzhu@swjtu.edu.cn)

识别注重时空序列与用户间的相关性。通常，同一用户在不同平台的轨迹分布具有一定的区域性和重复性。例如，用户的路径在办公地点和居住地点间重复。而不同用户轨迹具有一定的独特性，是用户个性化的体现。因此，研究以时空数据为切入点进行身份识别具有可行性。基于时空数据的身份识别可以关联不同平台的同一用户数据，构建更全面的用户画像，提升平台用户体验；可以整合时空序列数据，辅助执法机构识别罪犯活动轨迹，管制和规划城市交通，优化政府管理效能[6]；同时，对于用户加深对隐私暴露风险的认识具有一定的推动作用。

传统的研究计算轨迹中两两位置点的相似度，并将其作为身份识别的依据，这样的方法计算量大且识别效果不佳。Farid 等[7]证明了用户的统计信息中包含了大量可用信息，通过匹配跨两个平台的用户数据直方图来识别身份。该算法的准确率取决于数据的统计特征随时间推移而保留的程度。Riederer 等[8]注意到不同平台位置点成对计算时基于时间对齐的重要性，该算法将时间和位置划分为精密的单元格，用户同时同地出现的概率越大，则越可能为同一用户。由于数据的稀疏和不一致，不同平台的轨迹之间存在显著的时空不匹配，因此这类算法的识别率不高。Wang 等[9]采用马尔可夫链对用户移动轨迹建模，推断缺失的位置记录，但无法学习到长轨迹的序列依赖关系。最近，Feng 等[10]利用循环神经网络挖掘移动轨迹的隐含语义，捕获两个不同数据源中轨迹的相关性，但却忽略了用户的个性化偏好对身份识别的作用。

为解决数据稀疏、数据质量不高、时空不匹配，以及局部匹配的问题，本文提出了 UI-STDD 算法（User Identification algorithm based on bidirectional Spatio-Temporal Dependence and spatio-temporal Distribution）。该方法从时空序列、时间偏好和位置点分布 3 方面充分提炼并整合轨迹对特征，以提升用户身份识别性能。

本文的主要贡献总结如下：

（1）设计实现基于注意力的双向长短时记忆循环神经网络（BiLSTM）模块，捕获轨迹与轨迹对的高阶和复杂的序列信息，模拟序列转换模式和多周期迁移规律，缓解数据稀疏性问题。

（2）从不同粒度探索用户偏好，建立与时间分布关联的用户个性化模式，提高特征质量。

（3）设计随机时间感知的位置点匹配，校正对齐时空特征，同时与基于任意熵的全局区域分布散度相结合，实现从局部和全局深度挖掘轨迹对特征。

2　相关工作

基于时空数据的身份识别工作主要分为两类：基于数理统计和概率论的识别方法和基于深度学习的识别方法。

基于数理统计和概率论的识别方法主要是通过不同平台用户共同出现的地点进行匹配。Farid 等[7]认为大多数在线服务的用户都有独特的行为或使用模式，提出了 HIST 模型，通过匹配用户轨迹的直方图来识别用户。该算法的准确率取决于，随着时间的推移，数据的统计特征保留了多少。Luca 等[11]以用户在某地点出现的频率作为该地点可能被访问的概率估计，考虑了不同地点对身份识别任务的贡献权重。Alket 等[12]的研究发现，少量的共同地点足以重识别大量的用户。该算法将用户同时同地出现的次数作为匹配的依据，但忽略了时空序列的连续性特征。Riederer 等[8]将时间和空间划分为细小的格子，该算法假定用户在某个时间访问某个地点的次数服从泊松分布，通过二部图匹配识别同一用户。值得注意的是，格子划分的不同会直接影响计算量和识别的精度，并且严格的划分可能造成不同平台数据的时空不匹配。Ding 等[13]构造关联规则，从位置点序列中挖掘频繁项集，进而补充相对稀疏的数据集，解决数据稀疏性问题。但该算法是基于推断隐含位置足够准确，否则会有产生噪声数据的风险。Wang 等[9]利用马尔可夫链建模用户移动模型，同时聚合时间上下文来推断缺失的位置记录，强调去匿名化攻击和位置隐私保护机制。但基于马尔可夫链的轨迹建模方法以与历史状态无关为假设约束，同时，马尔可夫模型无法学习到长轨迹的序列依赖关系。

最近，一些研究将深度学习引入时空数据挖掘任务。Li 等[14]采用 seq2seq 模型度量从密集轨迹中提取的子轨迹间的相似性，但无法对来自不同平台的轨迹关系建模。Feng 等[10]提出了一种端到端的深度学习框架 DPLink，它由特征提取器和比较器组成。特征提取器通过循环神经网络提取轨迹特征，比较器判断两条轨迹是否对应于同一个用户。但该算法忽略了空间邻近性和用户个性化偏好对身份识别的作用。

针对上述研究存在的问题，本文提出了一种更充分利用时空数据的身份识别算法 UI-STDD。该算法能够刻画用户移动模式和个人偏好，以捕捉不同平台用户账号之间的关联性，增强用户轨迹的表征能力。

3　问题描述和符号定义

定义 1（位置点记录）　位置点记录指用户在基于位置的社交网络（LBSN）产生的一次时空位置数据，也被称作签到记录[6]。位置点记为 $l=(x,y,t)$，其中 x,y 代表当前位置的经纬度，t 为发布该位置时对应的时间戳。

定义 2（轨迹）　轨迹是由一系列按照时间顺序排列的位置点组成的序列[10]，可表示为 $R=\{l_1,l_2,\cdots,l_{|R|}\}$，其中 $|R|$ 为位置点的个数，即轨迹长度。

定义 3（用户身份识别）　给定两个不同平台的 LBSN，符号定义为 L^s 和 L^d，它们分别包含一系列用户账号 $U=\{u_1,u_2,\cdots,u_{|U|}\}$，$V=\{v_1,v_2,\cdots,v_{|V|}\}$，其中每个用户 $u\in U(v\in V)$，$|U|(|V|)$ 表示 $U(V)$ 中用户数。用户生成的轨迹记为 $R_u=\{l_{u1},l_{u2},\cdots,l_{u|R_u|}\}$，$R_v=\{l_{v1},l_{v2},\cdots,l_{v|R_v|}\}$。两个来自不同平台的用户轨迹组合 (R_u,R_v) 被称为轨迹对。用户身份识别任务旨在根据已知的轨迹对 (R_u,R_v) 识别用户账户 (u,v) 在现实生活中是否隶属于同一个人[5]，即 (u,v) 是否为对齐用户。

4　基于时空轨迹的跨网络用户身份识别算法

基于时空轨迹的跨网络用户身份识别算法框架如图 1 所示，其核心主要由 3 个模块组成：时空序列模块、时间偏好

模块和空间位置模块。以不同平台的用户生成轨迹为输入，时空序列模块实现融合注意力的双向依赖建模，提取轨迹连续性特征和模式；时间偏好模块将发布时间从细、粗两个粒度量化为 24 维和 7 维向量，集成个性化特征；空间位置模块从位置点着手，从局部对齐到全局分布刻画空间邻近性。最终，所提算法基于融合的多类时空轨迹特征利用多层前馈网络识别不同平台的两条轨迹是否为同一用户。下文以来自不同平台的用户(u,v)产生的轨迹对(R_u,R_v)为例展开讨论。

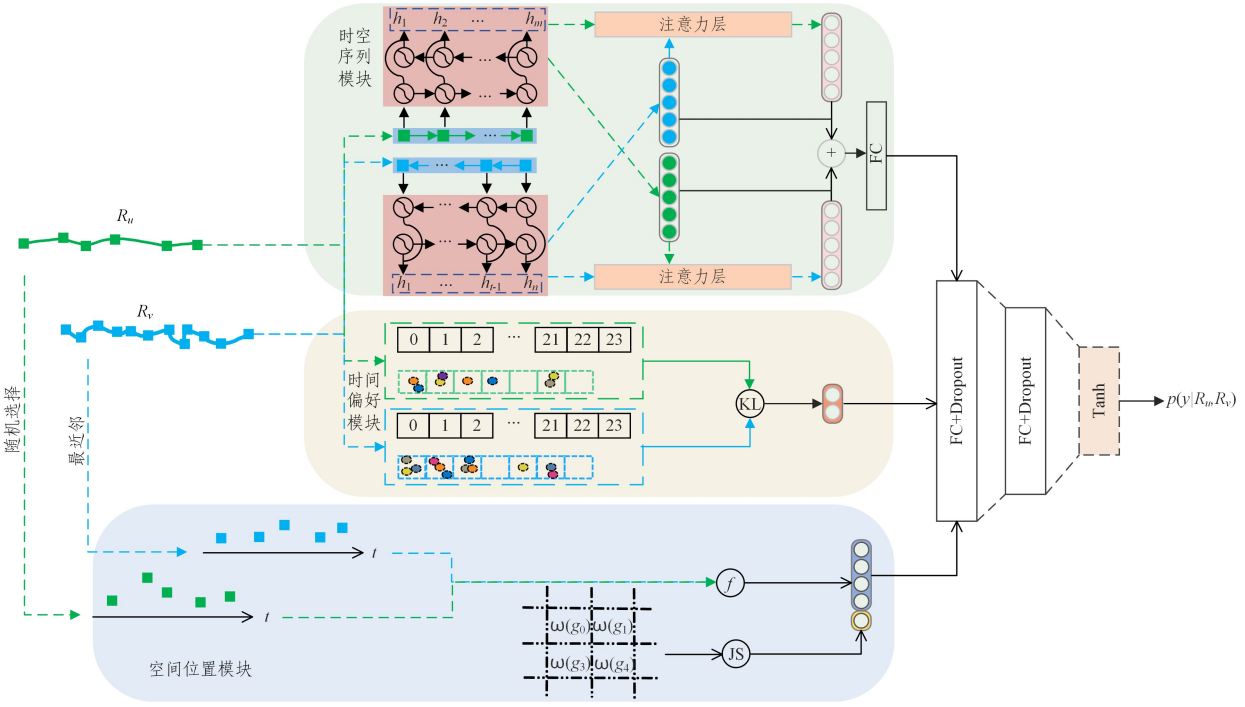


图 1 基于时空轨迹的跨网络用户身份识别框架

Fig. 1 Cross-network user identification framework based on spatio-temporal trajectory

4.1 融合注意力的双向时空依赖建模

为解决数据稀疏的问题，本文引入 BiLSTM 捕获序列转换模式，学习用户移动轨迹的潜在表示。同时模块融合注意力机制聚焦轨迹对的相似部分，缓解不同设备精度不一致或人为因素造成的噪声数据的影响。图 1 的时空序列模块中展示了本节所述的主要技术。

在单轨迹特征表示部分，R_u和R_v分别输入双向长短时循环网络。首先借鉴文献[14]的单元格编号法，将轨迹中二维的位置点离散化。接着，BiLSTM 将一系列按时间排序的格子编号编码得到多个隐层输出，记为$\{h_1,h_2,\cdots,h_n\}$。BiLSTM 由前向和后向两个 LSTM 网络组合而成，其考虑了历史位置和将来位置序列信息，能更好地表征时空序列，表示为：

$$h_i=\overrightarrow{h}_{Ri}+\overleftarrow{h}_{Li} \tag{1}$$

在循环网络后引入最大池化操作，得到固定长度的轨迹表征 y，即单轨迹表征。

在提取轨迹对特征部分，引入注意力机制学习轨迹对R_u和R_v的相关性。将轨迹R_u的编码隐层状态$\{h_{u1},h_{u2},\cdots,h_{um}\}$作为候选状态，轨迹$R_v$的表征向量$\boldsymbol{y}_v$为查询状态，输出轨迹对表征向量$\boldsymbol{z}_v$，计算式为：

$$\alpha_i=\mathrm{softmax}(h_{ui}^T y_v) \tag{2}$$

$$\boldsymbol{z}_v=\sum_{i=1}^{m}\alpha_i h_{ui} \tag{3}$$

同理，以轨迹R_u的表征向量$\boldsymbol{y}_u$为查询状态，输出轨迹对表征向量$\boldsymbol{z}_u$。最后，单轨迹表征向量$\boldsymbol{y}_u$，$\boldsymbol{y}_v$与轨迹对表征向量$\boldsymbol{z}_u$，$\boldsymbol{z}_v$融合，记为f_{tp}，作为身份识别模块的输入。

4.2 发布时间偏好建模

文献[15-16]指出，在不同社交平台上，用户也通常在相似的时间段签到或发布位置信息，反映了用户的在线活动模式。例如，有人喜欢在工作日的早上去咖啡店打卡，而另一些人更倾向于周日的午后打卡。集成上述潜在的发布时间偏好能扩充特征信息，更准确地判别用户身份，有利于解决数据质量低的问题。

UI-STDD 在两种粒度上量化轨迹中的时间分布，并结合 Kullback-Leibler 散度(KL)度量用户间的相关性，如图 1 的时间偏好模块所示。对于一个用户 u 的轨迹R_u，其时间序列$\{t_{u1},t_{u2},\cdots,t_{u|R_u|}\}$由一系列时间戳组成。一方面将每个时间戳以小时为单位量化到细粒度的 24 维区间中，例如，量化 2021 年 12 月 11 日 17 时 23 分 31 秒，数组下标 17 的值为 1，其余为 0。另一方面，将时间戳区分工作日和周末，同时映射到 5 个以不同时段为单位的粗粒度区间中，用 7 维向量表示 5 个时段，分别为：早上[8:00,11:30)、中午[11:30,14:00]、下午[14:00,17:30)、晚上[17:30,22:00)和其他时间段。例如，2021 年 12 月 11 日(星期六)17 时 23 分 31 秒，表示为$[0,1,0,0,1,0,0]^{\mathrm{T}}$。其中，第 2 维为 1 代表周末，第 5 维为 1 代表在[14:00,17:30)时间段。针对不同粒度区间统计并归一化处理后，此时用户细、粗粒度时间偏好概率P_g和P_c分别对应为 24 维和 7 维向量。

输入一对用户$(u,v|u\in U,v\in V)$轨迹，算法通过 KL 度量两个时间偏好表征向量的相关性，计算式如下：

$$T_{r1}=-\sum_{i=0}^{23}P_{ug}(i)\log\frac{P_{ug}(i)}{P_{vg}(i)} \tag{4}$$

$$T_{r2}=-\sum_{i=0}^{6}P_{uc}(i)\log\frac{P_{ugc}(i)}{P_{vc}(i)} \tag{5}$$

其中，log 以 2 为底数。此时，(T_{r1},T_{r2})作为用户对(u,v)的时间偏好特征，记为f_{td}。

4.3　随机时间感知的局部点对齐

计算用户对(u,v)空间位置相似度的一种直观方法是计算两两位置点对的相似度，即计算次数为轨迹对(R_u,R_v)的笛卡尔积$R_u\times R_v=\{(l_u,l_v)\mid l_u\in R_u,l_v\in R_v\}$。这导致计算量很大，而且由于位置点之间没有基于时间对齐，计算的相似度误差也较大。为解决此问题，本文提出随机时间感知的位置点匹配，在对齐时间的同时放松严格对齐区间的限制。以稀疏数据的平台用户 u 为基准，随机选取 K 个时间($K\ll|R_v|$，$|R_v|$为密集数据平台用户 v 轨迹的位置点个数})对应的位置点，与 v 签到记录中最近邻时间的位置点组合为基于时间对齐的 K 个位置点对$\{(l_{ui},l_{vi})\mid i=1,2,\cdots,k\}$。位置点对的相似度采用高斯核函数计算，公式为：

$$f=\frac{1}{2\pi h}\exp\left(-\frac{(l_{vi}-l_{ui})^2}{2h^2}\right) \tag{6}$$

其中，h 是带宽参数，$l_{vi}-l_{ui}$是位置点l_{vi}和l_{ui}的欧几里得距离。K 个局部位置点距离估计描述用户对签到活动的空间邻近性，根据空间模式区分用户。

4.4　全局区域分布

从全局空间来看，轨迹的区域分布具有区分用户的能力，因为虽然一个用户在不同的社交平台可能发布不同的签到记录，但该用户在各平台签到记录的空间分布往往是相似的[17]。本文在全局空间划分区域，从频繁访问的区域差异区分用户，通过量化用户 u 与用户 v 区域分布的近似度，并将其作为用户对(u,v)位置点信息的补充。全局空间由覆盖位置点的最高和最低经纬度包围组成，通过经纬区间为 100×100，10×10 甚至更大的网格将整个空间划分为多个矩形区。结合数据集分析，本文选取经纬区间为 100×100 的网格，共划分为 8 个矩形区，如图 2 所示。同时，本文设计基于任意熵的方法为每个矩形区赋予权重 $\omega(g_j)$，解决热门区域会出现多人访问的情况。计算式如下：

$$H(g_j)=\frac{3}{2}\log\sum\left(\frac{N_u(g_j)}{|R_u|}\right)^{\frac{2}{3}} \tag{7}$$

$$\omega(g_j)=\exp(-H(g_j))=-\frac{3}{2}\sum\left(\frac{N_u(g_j)}{|R_u|}\right)^{\frac{2}{3}} \tag{8}$$

其中，g_j为第 j 区域，$N_u(g_j)$为 u 访问g_j的次数。故 u 的区域分布e_u表示为$\left[\frac{N_u(g_1)}{|R_u|}\omega(g_1),\frac{N_u(g_2)}{|R_u|}\omega(g_2),\cdots,\frac{N_u(g_8)}{|R_u|}\omega(g_8)\right]$。考虑到用户访问区域不完全重叠，$u,v$ 间的全局区域分布近似度采用 Jensen-Shannon 散度(JS)计算。计算式如下：

$$D_{KL}(p,q)=-\sum_{i=1}^{8}p(i)\log\left(\frac{p(i)}{q(i)}\right) \tag{9}$$

$$D_{JS}=\frac{1}{2}D_{KL}\left(e_u,\frac{e_u+e_v}{2}\right)+\frac{1}{2}D_{KL}\left(e_v,\frac{e_u+e_v}{2}\right) \tag{10}$$

分布近似度D_{JS}与局部位置点估计向量拼接，记为f_{ld}，表示用户轨迹对的空间特征，如图 1 中的空间位置模块。

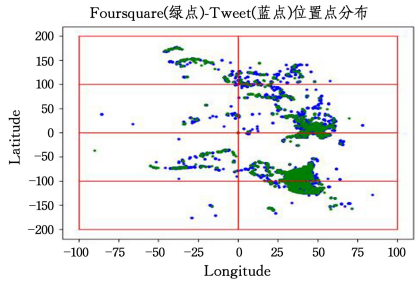


图 2　FS-TW 区域划分示意图(电子版为彩图)

Fig. 2　Diagram of FS-TW region division

4.5　身份识别

上述工作从轨迹f_{tp}、时间偏好f_{td}、空间f_{ld}得到轨迹对的矢量表示，3 类向量分别用于捕获时空特征、深入挖掘轨迹潜在语义和轨迹间的内在联系、判定两个用户的相关性。基于特征表示，身份识别模块采用多层前馈网络识别跨网络用户身份，将识别任务归类为二元分类任务。定义该轨迹对是否由同一用户产生(两条轨迹对应的用户账号是否属于同一个人)的概率为：

$$p(1\mid(u,v))=\text{sigmoid}((f_{tp}+f_{td}+f_{ld})\cdot W) \tag{11}$$

其中，W 为可学习参数。模型通过二元交叉熵损失函数进行优化，函数定义如下：

$$Loss=-\sum q_i\log p_i+(1-\boldsymbol{q}_i)\log(1-p_i) \tag{12}$$

其中，p_i是识别模块的预测输出，$\boldsymbol{q}_i$是真实分类结果所对应的二值化向量。在训练阶段，本文采用 Adam 算法优化模型，借助 L2 正则化、随机失活(Dropout)避免过拟合问题。

5　验证实验与结果分析

5.1　实验数据集

本文使用 Foursquare-Twitter，Instagram-Twitter，Brightkite1-Brightkite2 这 3 个真实的用户签到数据集进行实验。为了评估算法性能，本文只使用签到记录中的位置信息和时间戳，数据集的统计信息如表 1 所列。

表 1　实验数据统计信息

Table 1　Statistical information of experimental data

Dataset	Network	users	Records	Data range
FS-TW	FS	5392	76972	2008.10−2012.11
	TW	5223	164919	
IT-TW	IT	2505	337934	2010.09−2015.04
	TW	1727	447366	
BK1-BK2	BK1	8133	566350	2008.04−2010.10
	BK2	7418	1579470	

Foursquare-Twitter(FS-TW)由文献[17]提供，并被最近的工作[4,8,10,17-18]广泛采用。数据集涉及从 2008 年 10 月到 2012 年 2 月的签到数据，共包含 2410 对对齐用户。

Instagram-Twitter(IT_TW)由 Riederer 等[8]采集，共包含 1717 对对齐用户。Instagram 的位置点涵盖(−53.16，−170.27)到(71.03，177.43)区间，Twitter 的位置点涵盖(−74.05，−159.76)到(71.03，175.81)区间。

Brightkite1-Brightkite2 (BK1-BK2) Brightkite 由 Cho 等[19]收集，原始数据集包含 58228 个用户在 2008 年 4 月到 2010 年 10 月产生的位置数据。实验数据集为从原始数据集

中导出的子集，本文随机选择 10000 个用户产生的轨迹数据，以 0.2 的概率保留在 BK1 中，以 0.6 的概率保留在 BK2 中，其余的直接丢弃。数据预处理包括去除少于 5 次的签到记录等，共计 6637 个用户对，得到跨网络数据集 BK1-BK2。

5.2 对比方法

实验选取以下 8 种基准方法进行对比。

(1)ME：Alket 等[17]注意到时间对齐对识别用户的重要性，并且认为两条轨迹中同时同地出现的次数越多，越可能对应于同一用户。

(2)WYCI：Luca 等[11]提出了基于概率论的算法，以用户在某地点出现的频率作为该地点可能被访问的概率估计，考虑了不同地点对身份识别任务的贡献权重。

(3)POIS：在一定的精度下，将时间和空间划分为细小的格子，Riederer 等[8]假定用户在某个时间段访问某个地点的次数服从泊松分布，基于移动模型识别同一用户。

(4)HIST：Farid 等[7]通过统计用户访问每个位置的次数来构造频率直方图，将两个直方图的相似度得分定义为两用户的匹配度。

(5)GM：Wang 等[9]采用马尔可夫链学习用户移动模型，同时聚合时间上下文减少数据中噪声的影响。

(6)GKR-KDE：Chen 等[4]提出了一种基于核密度估计的方法，用于直接测量两个账号产生的位置点间的相似度，并基于网格结构精简搜索空间。

(7)DeepTUL：Miao 等[20]将单平台轨迹分为历史轨迹和当前查询轨迹，利用循环神经网络编码轨迹并链接到对应的用户。DeepTUL 不能直接应用于跨平台身份识别任务，因此，本文将两类单平台轨迹扩展为跨平台用户产生的轨迹，模型目标为分类任务，识别输入轨迹是否为同一用户产生。

(8)DpLink：Zhang 等[18]通过位置编码器和轨迹编码器对单轨迹特征建模，用基于共注意力的选择器进行判定。

5.3 参数设置和评价指标

考虑到不同平台移动轨迹的异质性等情况，本文对模型进行了预训练。对数据相对密集的单平台按两平台数据量平均比采样子轨迹，通过双向时空依赖模块学习单平台上两条轨迹移动模式，判断两条子轨迹是否属于同一用户。然后将预训练的网络迁移到跨网络学习任务中，帮助学习模型参数和轨迹模式。

本文对比了不同参数对应方法在测试集上的表现，经过选择，UI-STDD 的关键超参数设置如表 2 所列。

表 2 超参数设置

Table 2 Hyperparameter setting

Hyperparameter	Value
Batch size	32
Embedding size	120
Number of random times/K	10
Divided region	100 * 100
Learning rate	10^{-3}
Dropout	0.3

为了更好地进行对比，本文使用与文献[4,8-10]相同的评价指标，包括 Precision，Recall 和 F1，定义如下：

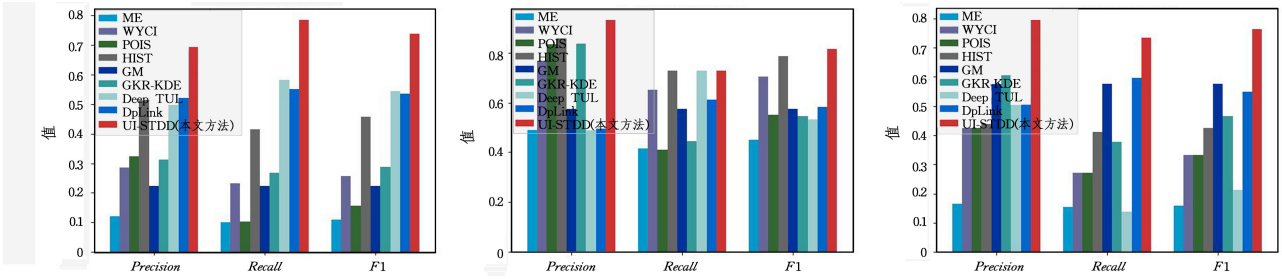
$$Precision=\frac{\beta}{\gamma},Recall=\frac{\beta}{\delta} \tag{13}$$

$$F1=\frac{2*Precision*Recall}{Precision+Recall} \tag{14}$$

其中，β 是算法正确对齐的用户对数量，γ 是算法对齐的总用户对数量，δ 是数据集中实际匹配的用户对数量。

5.4 实验结果分析

将本文方法 UI-STDD 和基准算法进行比较，图 3(a)—图 3(c)分别对应跨 FS-TW，IT-TW 和 BK1-BK2 网络平台数据上的用户对齐性能。



(a)跨 FS-TW 网络平台 (b)跨 IT-TW 网络平台 (c)跨 BK1-BK2 网络平台

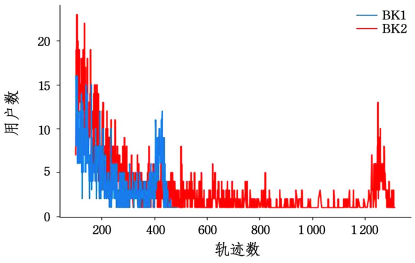
图 3 实验结果

Fig. 3 Experimental results

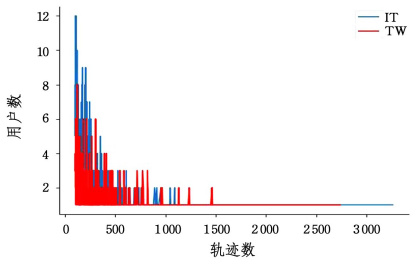
在 FS-TW 和 BK1-BK2 中，Foursquare 和 BK1 上的数据稀疏，而 Twitter 和 BK2 上的数据相对密集，两平台数据分布存在较大的差异性。如图 4(a)所示，x 轴表示轨迹长度，y 轴表示对应的用户数量。实验结果表明，DeepLink 和 UI-STDD(本文算法)比前 6 类基于概率论和数理统计的算法对齐用户的能力更强。原因在于，前 6 类方法忽略了运动轨迹的连续性特征和隐含语义，而基于深度学习的算法通过对不同平台轨迹隐含语义的表征，在一定程度上能缓解数据稀疏和数据量级不一致的问题。其中，本文方法识别效果最好，这得益于双向时空依赖建模和时空分布融合，与 DeepTUL，DeepLink 相比，增强了模型对用户移动模式的提取能力。此外，UI-STDD 在轨迹表征的同时融入成对注意力，考虑了隐层输出对识别用户的贡献程度，更注重两轨迹间相关性的捕捉。

对于 IT-TW，两平台的数据量级比较一致，如图 4(b)所示。各算法识别准确率均有所提升，且基于统计的算法 HIST 比基于深度模型的 DeepTUL 和 DeepLink 有更好的表现。这是因为用户在两平台的签到位置数量量级差异不大以及签到点较重合时，基于位置点的匹配和统计的算法更能

发挥优势。融合局部和全局位置模块的 UI-STDD 在 Precision 和 F1 上仍保持最优，Precision 提高了 7%以上。具体地，该模块通过计算局部点匹配和跨平台空间分布相似性，解决了时空不匹配的问题，有效提升了用户识别率。



(a)跨 BK1-BK2 网络平台



(b)跨 IT-TW 网络平台

图 4　用户轨迹数量分布

Fig. 4　Number distribution of user trajectory

为判断 UI-STDD 的不同特征提取模块对用户身份识别的贡献程度，设计实现了仅时空依赖模块(UI-FTP)、本文算法去空间位置模块(UI-FTP-T)、本文算法去时间偏好模块(UI-FTP-L)3 组消融实验，结果如表 3－表 5 所列。在仅考虑时空依赖的情况下，通过 BiLSTM 结合成对注意力捕获轨迹的移动模式，在一定程度上对齐用户。在 UI-FTP-T 实验中，增加用户发表位置点的时间分布特征，在原有轨迹编码的基础上考虑用户在线活动模式，较前一实验对识别身份更有效。而 UI-FTP-L 融合位置点对齐与空间位置分布，从局部和全局描述空间邻近性，有效增补了代表性特征，用户对齐效果更好。总之，增强特征并融合多类特征，大大提升了用户对齐性能，F1 值比仅使用双向时空特征模式至少提高了 20%。

表 3　本文算法变体对比(跨 FS-TW 网络平台)

Table 3　Comparison of the proposed algorithm variants (across FS-TW)

Algorithm	*Precision*	*Recall*	*F1*
UI-FTP	0.521	0.550	0.535
UI-FTP-T	0.732	0.723	0.727
UI-FTP-L	**0.754**	0.578	0.654
UI-STDD	0.695	**0.784**	**0.737**

表 4　本文算法变体对比(跨 IT-TW 网络平台)

Table 4　Comparison of the proposed algorithm variants (across IT-TW)

Algorithm	*Precision*	*Recall*	*F1*
UI-FTP	0.420	0.729	0.533
UI-FTP-T	0.493	0.560	0.524
UI-FTP-L	0.847	**0.775**	0.809
UI-STDD	**0.933**	0.729	**0.818**

表 5　本文算法变体对比(跨 BK1-BK2 网络平台)

Table 5　Comparison of the proposed algorithm variants (across BK1-BK2)

Algorithm	*Precision*	*Recall*	*F1*
UI-FTP	0.521	0.550	0.535
UI-FTP-T	0.579	0.418	0.485
UI-FTP-L	0.744	0.587	0.656
UI-STDD	**0.794**	**0.734**	**0.762**

结束语　本文提出了一种整合多类时空特征的跨平台社交网络用户身份识别方法 UI-STDD。该方法提出融合注意力的双向时空建模捕获轨迹序列移动模式，解决了数据稀疏性问题；从不同粒度量化时间分布，集成个性化模式特征，提升了数据质量；提取位置点的局部和全局特征，计算跨平台空间分布相似性，解决时空不匹配问题，有效优化了跨网络识别用户的能力。对比实验结果表明，UI-STDD 跨网络识别相同用户的准确率远优于对比算法。下一步工作将研究用户发布的文本内容，探索其对用户身份识别的作用，以及考虑跨多个平台实现识别任务，使研究结果具有更广泛的应用。

参 考 文 献

[1] LUO Y T, LIU Q, LIU Z C. STAN: Spatio-Temporal Attention Network for Next Location Recommendation[C]// Proceedings of the Web Conference. New York: ACM, 2021: 2177-2185.

[2] SINA D, CHANG T L, KEVIN H, et al. Semi-Supervised Deep Learning Approach for Transportation Mode Identification Using GPS Trajectory Data[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(5): 1010-1023.

[3] GUO Y S, LIU M D. Anomaly detection based on spatio-temporal trajectory data[J]. Computer Science, 2021, 48(S1): 213-219.

[4] CHEN W, YIN H Z, WANG W Q, et al. Effective And Efficient User Account Linkage Across Location Based Social Networks [C]// IEEE 34th International Conference on Data Engineering. New York: IEEE Press, 2018: 1085-1096.

[5] ZHOU X P, LIANG X, ZHAO J C, et al. A review of related user mining methods for social network convergence[J]. Journal of Software, 2017, 28(6): 1565-1583.

[6] LI H, CAO S Y, CHEN Y Z, et al. User Trajectory Identification Model via Attention Mechanism[J]. Computer Science, 2021, 49(3): 308-312.

[7] FARID M N, JAVKRICHAN U, PATRICK T, et al. Where You Are Is Who You Are: User Identification by Matching Statistics[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(2): 358-372.

[8] RIEDERER C, KIM Y S, CHAINTREAU A, et al. Linking Users Across Domains with Location Data: Theory and Validation[C]// World Wide Web. New York: ACM, 2016: 707-719.

[9] WANG H D, GAO C, LI Y, et al. De-anonymization of Mobility Trajectories: Dissecting the Gaps between Theory and Practice [J]. IEEE Transactions on Mobile Computing, 2021, 20(3): 796-815.

[10] FENG J, ZHANG M Y, WANG H D, et al. DPLink: User Identity Linkage via Deep Neural Network From Heterogeneous

Mobility Data[C]// World Wide Web. New York: ACM, 2019: 459-469.

[11] LUCA R, MIRCO M. It's the way you check-in: Identifying users in location-based social networks[C]// Proceedings of the Second ACM Conference on Online Social Networks. New York: ACM, 2014: 215-226.

[12] ALKET C, MARCO M, FRANCO Z. Re-identifification and information fusion between anonymized CDR and social network data[J]. Journal Ambient Intelligent Human Computing, 2016, 7(1): 83-96.

[13] DING F X, MA X Q, YANG Y, et al. User Identity Linkage across Location-Based Social Networks with Spatio-Temporal Check-in Patterns[C]// IEEE International Conference on Parallel & Computing & Communications. New York: IEEE Press, 2020: 1278-1285.

[14] LI X L, ZHAO K Q, CONG C, et al. Deep Representation Learning for Trajectory Similarity Computation[C]// IEEE 34th International Conference on Data Engineering. New York: IEEE Press, 2018: 617-628.

[15] XI D B, ZHUANG F Z, LIU Y C, et al. Modelling of Bi-Directional Spatio-Temporal Dependence and Users' Dynamic Preferences for Missing POI Check-In Identification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 5458-5465.

[16] CHEN W, WANG W Q, YIN H Z, et al. User Account Linkage Across Multiple Platforms with Location Data[J]. Journal of Computer Science and Technology, 2020, 35: 751-768.

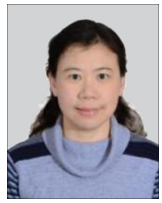
[17] KONG X G, ZHANG J W, PHILIP S Y, Inferring anchor links across multiple heterogeneous social networks[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM, 2013: 179-188.

[18] ZHANG J W, PHILIP S Y. Integrated Anchor and Social Link Predictions across Social Networks[C]// Proceeding of the 24th International Joint Conference on Artificial Intelligence. California: Morgan Kaufmann, 2015: 2125-2132.

[19] CHO E, MYERS S A, LESKOVES J. Friendship and Mobility: Friendship and Mobility: User Movement in Location-Based Social Networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011: 1082-1090.

[20] MIAO C C, WANG J L, YU H, et al. Trajectory-User Linking with Attentive Recurrent Network[C]// Proceeding of the 19th International Conference on Autonomous Agents and Multi Agent Systems. Richland: Springer, 2020: 878-886.

LIU Hong, born in 1995, postgraduate. Her main research interests include cross-network user identification and data mining.



ZHU Yan, born in 1965, Ph.D, professor, Ph.D co-supervisor, is a member of China Computer Federation. Her main research interests include Web data mining, social networking, privacy preserving, deep learning and AI.

(责任编辑:何杨)