doi:10.11887/j.cn.202203008 http://journal.nudt.edu.cn

融合空间偏好和语义的个体活动识别方法*

郭茂祖[1],陈加栋[1],张 彬[1],赵玲玲[2],李 阳[1]
(1. 北京建筑大学 电气与信息工程学院,北京 100044;2. 哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

摘 要:个体活动识别对用户画像、个性化推荐、异常行为检测、群体行为分析和基于活动的资源配置优化具有重要价值。提出了一种基于稀疏的社交媒体签到数据的个体活动语义识别方法,从签到数据中提取活动行为的时间周期性和趋势性特征,并采用空间偏好量化算法,从个体与群体活动的空间关联中提取群体和个体的空间访问偏好,使用自然语言嵌入工具BERT模型提取访问兴趣点的语义。时间特征、空间偏好特征和访问兴趣点名称语义特征共同构成表征群体、个体偏好的时空联合特征,通过极限梯度提升分类器对其进行分类,得到活动语义识别结果。在Foursquare数据集上的对比实验和消融实验中验证了所提活动语义识别模型可以有效提升活动语义识别的准确性。

关键词:活动语义识别;空间偏好;兴趣点语义;极限梯度提升树;BERT
中图分类号:TP391 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2022)03-057-10

听语音 与作者互动 聊科研

Method for individual activities recognition incorporating spatial preference and semantics

GUO Maozu[1], *CHEN Jiadong*[1], *ZHANG Bin*[1], *ZHAO Lingling*[2], *LI Yang*[1]
(1. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China;
2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: The recognition of individual activities helps in the realisation of functions such as user profiling, personalized recommendations, abnormal behaviour detection, city-wide group behaviour analysis and resource allocation optimisation. A recognition method for the semantics of individual activities based on sparse social media check-in data was proposed. The temporal periodicity and tendency features of activity behaviors were extracted from the check-in data, and a spatial preference quantification algorithm was utilized to extract the preferences of groups and individuals from the spatial relevance between individual and group activities. The natural language embedding model BERT was used to extract the semantics of POIs (point of interest). The temporal features, spatial preference features and text features of POI′s names constituted the joint spatio-temporal features characterizing group and individual preferences, and the joint features were classified by the extreme gradient boosting classifier to obtain the activity semantic recognition results. With the results of comparison experiments and ablation experiments on the Foursquare dataset, it was validated that the model proposed can effectively improve the accuracy of activity semantics recognition.

Keywords: semantic recognition of activities; spatial preference; semantics of point of interest; extreme gradient boosting tree; BERT

移动网络的普及与移动终端设备的性能提升促进了基于位置的社交网络(location-based social-networks, LBSNs)快速发展[1]。社交网络用户通过文字、图片、评分等形式分享日常动态,而社交关系、时空轨迹、活动出行等信息隐含在用户上传信息中。因此,基于社交网络中用户信息的挖掘能够获取用户的出行和活动的模式与偏好等特征,为用户画像、目的地推荐及出行规划[2-5]、个性化产品广告投放等实际应用提供支持。同时,其也有助于进一步了解城市范围内的群体行为模式,对城市规划[6-8]、资源配置优化、异常检测等领域具有重要价值。

在地理信息系统中,兴趣点(point of interest, POI)的地标用于标示某一地点所代表的设施、景点、地标等场所,包含地点对应的名称、类别、经纬度、海拔等信息。在用户通过社交媒体记录分享自身活动时,其上传的地点名称及兴趣信息能够使对应位置构成POI。由于社交签到数据包含了

* **收稿日期**:2021-06-15
基金项目:国家自然科学基金面上资助项目(61871020,62101022);北京市属高校高水平创新团队建设计划资助项目(IDHT20190506);国家重点研发计划子课题资助项目(2020YFF0305501);北京市教委科技计划重点资助项目(KZ201810016019)
作者简介:郭茂祖(1966—),男,山东德州人,教授,博士,博士生导师,E-mail:guomaozu@bucea.edu.cn;
李阳(通信作者),女,讲师,博士,E-mail:liyang1@bucea.edu.cn

签到时间、POI等信息,其隐含的信息能够描述用户在签到时进行的目的性活动所属类别,即活动语义信息。社交媒体包含的大量签到与活动数据,为挖掘用户活动偏好、识别活动语义提供了基础。

在活动语义识别任务中,当前研究聚焦于挖掘时间信息中的活动行为的周期性和趋势性特征,对空间经纬度信息的使用主要在于计算签到点与周围POI之间的距离,作为POI推荐依据[3,9],或利用用户对访问点的历史信息进行统计获得访问频率,用于个体活动行为偏好分析[10]。数据中所包含的签到地点的名称文本信息对于活动语义的识别同样具有重要意义,并且其在一定程度上可以直接反映活动语义。许多现有研究采用潜在狄利克雷分配(latent Dirichlet allocation, LDA)主题模型来对POI名称进行提取,通过生成的主题分布对个体进行活动行为识别,或进行旅游景点推荐[3-4]等;另有研究采用聚类[2]的方法对POI进行分析,从而为用户推荐兴趣点。

活动语义识别的难点不仅体现在不同个体在同一地点的活动语义差别,同时也体现在不同场所经纬度、高度信息相似所造成的识别困难。现有活动语义识别研究对时间信息中的周期性和趋势性特征考虑较为充分,但并未结合用户的活动空间偏好特点和POI语义信息。

针对上述问题,本文提出了一种基于空间偏好和POI语义的个体活动语义识别方法,通过签到数据中的时间信息、空间信息和POI文本信息对个体活动语义进行识别。将活动语义识别视为多分类问题,通过特征工程挖掘时间、空间、文本信息中的关键特征,构建反映群体和个体偏好的时空联合特征向量,并采用极限梯度提升(extreme gradient boosting,XGBoost)算法建立分类器,对特征进行编码并构建活动语义识别模型。

1 相关工作

个体的活动行为在时间和空间中通常具有较强的规律性,分析时空数据能够实现用户的活动类别识别,对于城市规划、交通规划、针对个体兴趣偏好的推荐系统有着重要的价值。

用户的时空轨迹或签到数据在个体的活动偏好和推荐领域存在许多研究应用。Zhu等[2]设计了一种个性化的旅游景点推荐算法,向用户推荐旅游景点。通过基于密度的聚类算法(density-based spatial clustering of applications with noise, DBSCAN)对空间地理数据进行聚类,获得旅游景点POI对应签到数据,再通过主题提取模型LDA分别为用户及区域生成潜在主题分布,并与用户的社交关系数据相结合,构建模型获取个体用户对应不同活动语义的得分。之后将模型与旅游景点的位置、评分相结合,对旅游景点进行排名,实现对用户的景点推荐。Qiao等[3]通过学习用户与POI的潜在表示形式,将地理位置信息、用户关系信息和时间信息进行综合,提出联合表示学习框架,将上述因素纳入计算得出各个POI间的转移概率,完成POI的推荐和社交链接的预测。Cao等[4]采用了基于社交网络的矩阵分解框架,根据用户与地区POI的交互数据分析其活动偏好,采用频谱聚类对POI进行聚类,结合地理信息为用户推荐POI。在用户活动行为的建模中,挖掘相似用户的活动行为,对建模用户个体的活动偏好和移动模式有重要作用。Rizwan等[10]通过核密度估计(kernel density estimation,KDE)方法帮助观察分析活动行为和目的地的稀疏分布,以及识别活动事件的精细密度,并使用标准偏差椭圆(standard deviational ellipse,SDE)方法分析签到行为的空间分布区域。研究结果表明男性和女性在活动行为偏好上有很大区别,活动时间选择上也有不同。Araújo等[11]将随机森林模型和马尔可夫模型结合构建一个集成学习器,用于预测用户的下一活动位置。

基于社交位置的社交网络数据存在很多问题,比如数据稀疏、数据量低、可信度差、缺乏权威认可、隐私问题等。Martí等[12]分析了LBSN数据的主要问题,并提供了相关使用方法。Kim等[13]采用模拟用户数据解决上述问题,方法以大量社交网络数据为基础构建框架,结合人类实际生活行为模式生成模拟用户数据,提取模拟用户的时间、空间、社交关系信息。文献[14]通过类似方法构建了地理模拟系统,生成人群需求模拟数据进行研究。

Zhong等[15]挖掘用户移动位置和时间的上下文相关性,分析用户移动模式,获取兴趣偏好。通过位置信息对用户活动相似性进行建模,将用户频繁活动的区域视为兴趣中心,提出多中心聚类算法衡量用户的相似性。Wang等[16]提出基于图嵌入的半监督学习框架为位置场所注释,挖掘场所位置相关性和访问相同地点的用户相似性。Zhang等[17]将时空特征和活动轨迹的额外活动信息相融合,用以解决在轨迹补全问题中的不足。

Shi 等[18]在人流流动性研究中,关注人类在移动中的时空特性和移动动机,探讨活动轨迹研究中活动目的对移动的影响。

社交媒体数据能够辅助决策城市规划和城市资源配置,以数据来驱动规划和发展。González 等[5]挖掘个体时空活动轨迹,不完全重合的活动轨迹遵循简单的可复制模式。出行方式的固有相似性影响到城市规划、城市资源配置以及流行病的预防和响应。Van Weerdenburg 等[6]基于休闲活动和旅游类数据,对比三种有监督多标签机器学习方法,探究这些理论在城市休闲和旅游业研究以及相关城市政策和规划中的潜力,为城市休闲和旅游研究提供了新视角。Cai 等[7]通过人类活动的时空模式,提取城市空间动态语义,并揭示了北京城市动态的五个小时模式、四个每日模式和六个空间模式。Huang 等[8]通过市民与交通系统的互动解决交通拥堵,利用社交媒体数据分析城市交通和城市动态,探索人类活动对日常交通拥堵影响。

个体社会活动行为与个体社会关系网络有重要关系。基于此,Pan 等[19]提出发现社交关系中有影响力的朋友算法,通过签到数据中的用户语义信息计算不同用户之间的影响。Papangelis 等[20]将地理区域中的地域性概念用于理解个体的空间活动和社交行为。这种地域性特点会影响到个体与个体间的交互以及他们所处的环境。本文受个体活动语义识别研究[2-4,10,15]启发,针对个体数据中隐含的兴趣偏好对识别准确率的影响进行了探究。为获取用户空间访问偏好,文献[2,4,15]采用聚类方法由地理信息提取热点访问区域,但基于聚类方法获取的空间偏好区域[2]可能将不属于热点区域的边缘点包含到聚类簇中,对识别效果造成不利影响。本文基于用户对各区域的访问频数进行统计,提出了更加直观的个体空间访问偏好表征方法。

为了利用签到点名称来挖掘文本信息,文献[2]采用 LDA 聚类对热点区域内的旅游景点进行主题生成,从而获取用户的潜在访问主题偏好以及各区域的潜在主题分布。在个体活动语义识别任务中,不同个体在同一地点可以拥有不同的活动语义,而聚类所得的空间区域潜在主题分布固定,不利于活动语义识别。本文通过词向量嵌入模型 BERT(bidirectional encoder representations from transformers)将签到地点名称转换为带有语义的向量,由于包含语义的特征向量和活动语义间不存在固定对应关系,在活动语义识别准确性上能够获得更好的表现。

2　个体活动的时空联合特征表示

本文在进行个体活动语义识别过程中同时考虑了时间、空间以及 POI 名称文本中潜在的语义信息,将活动语义识别作为一个多分类问题处理。在访问时间特征方面,提取签到时间特征,以表示活动行为的时间周期性和趋势性。

空间特征方面,由于用户在不同访问位置的访问频率差异反映了一定的偏好,因此,本文提出基于访问热度的群体和个体的空间偏好特征。空间偏好带有群体性和个体性两个层面:群体偏好的产生表现为签到数据空间中存在若干热点访问区域,这些热点访问区域往往关联典型的活动类型,如旅游胜地、热门餐厅等;个体偏好则是针对个体签到数据而言,由于用户的出行习惯或工作要求,也可能出现热点访问区域,且个体在这些区域倾向于进行相同的活动。

POI 语义方面,采用 BERT 模型提取 POI 名称隐含的语义信息。综合上述三个维度提取的特征构建联合特征向量,使用 XGBoost 算法建立分类器。整体的活动语义识别模型框架如图 1 所示。

2.1　空间偏好特征表示

不同个体具有不同的出行和活动习惯偏好,在签到行为数据的空间分布上也有所反映,具体表现为个体或群体对不同子区域的访问频率差异,并随签到行为进行逐渐形成用户个体和群体的热点访问区域。由于个体在特定空间区域的活动类别与其日常习惯偏好存在紧密联系,相关空间特征的挖掘对个体活动识别具有重要的意义。本文基于用户签到数据对不同子空间的访问频率提出空间偏好的估计方法,分别对个体、群体的空间偏好进行了度量,并获取相应的空间热点访问区域,作为签到行为的空间特征表示个体的潜在活动模式。

2.2　用户空间偏好的度量方法

本文采用基于子空间访问频率统计的度量方法,分别提取用户个体、群体的热点访问区域及相应访问量作为空间偏好特征,具体提取方法描述如下。

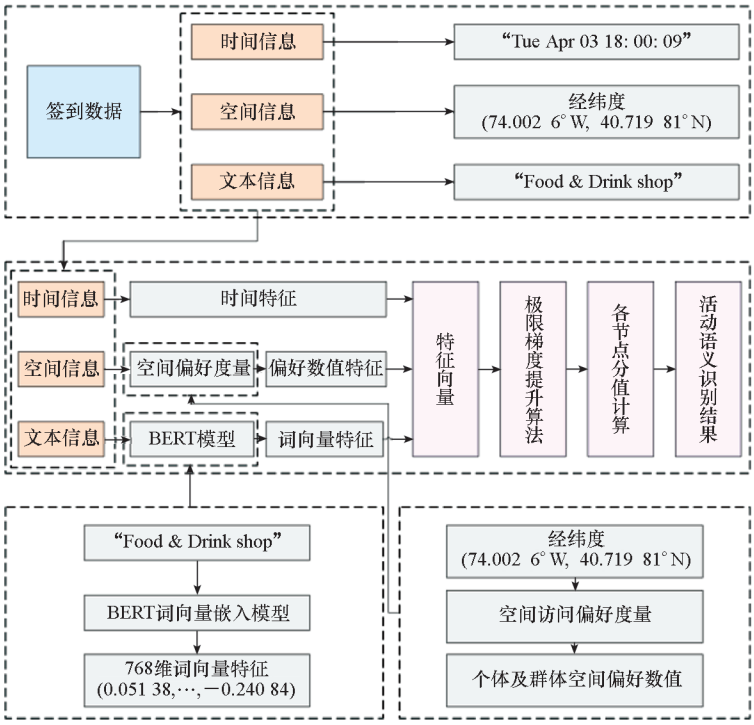

图 1 基于空间偏好和 POI 语义的活动语义识别框架

Fig. 1 Activity semantic recognition framework based on spatial preference and POI semantic

1)首先定义子空间区域的经纬度阈值 ε,将签到数据集空间范围 S 划分为网格子区域 S_i,子空间起始经度 a_0、纬度 b_0 分别取签到点分布范围的经纬坐标下界:

$$\forall (a, b) \in S_i \begin{cases} a \in (a_0 + (i-1)\varepsilon, a_0 + i\varepsilon) \\ b \in (b_0 + (i-1)\varepsilon, b_0 + i\varepsilon) \end{cases} \tag{1}$$

将用户签到空间位置坐标集合表示为 L,第 i 条数据的签到点经纬度位置表示为 l_i。

$$L = \{l_1, l_2, \cdots, l_n\} \tag{2}$$

$$l_i = (a_i, b_i) \tag{3}$$

2)采用如下方法估计某个区域的空间偏好:首先统计 L 中所有签到点坐标,获取落在子区域 S_i 内的签到位置坐标及个数 n_i,构成子区域签到点数量集合 N。然后分别计算子区域内任意点 l_i 作为质心 C_i 时的均方误差 E,选取计算所得均方误差最小的签到点作为该子区域质心:

$$E = \sum_{l_i \in S_i} \| l_i - C_i \|_2^2 \tag{4}$$

$$C_i = \frac{1}{S_i} \sum_{l_i \in S_i} l_i \tag{5}$$

3)考虑到签到点有可能落在划分子区域边缘上,两类边缘点分别类似于图 2 的 A、B 两点,对两类点分别计算坐标与邻近各子区域质心的欧式距离 d_i,即对 A 点类型,计算 $d_i(i = 1, 3)$;对 B 点类型,计算 $d_i(i = 1, 2, 3, 4)$。取最小 d_i 对应的子区域作为边缘点所属:

$$d_i = \| l_i - C_i \|_2^2 \tag{6}$$

图 2 展示了签到数据在空间中的分布,每条横线表示纬度,竖线表示经度。通过上述方法将整个签到空间划分为一个个的签到子区域,各个子区域中的一个点就表示有一次签到行为发生于此空间内,模拟了签到点在签到空间内的分布情况。

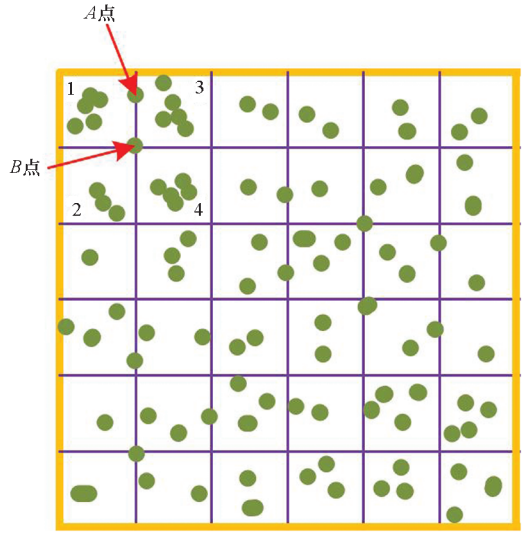

图 2 签到点空间分布示意

Fig. 2 Schematic diagram of check-in points

图 3 和图 4 分别展示了 Foursquare 纽约市公开数据集的群体、个体签到记录空间分布,并显示了通过本文空间偏好度量方法获取的特征值。特

征值越大,表示个体或群体对该子区域的访问偏好程度越高。

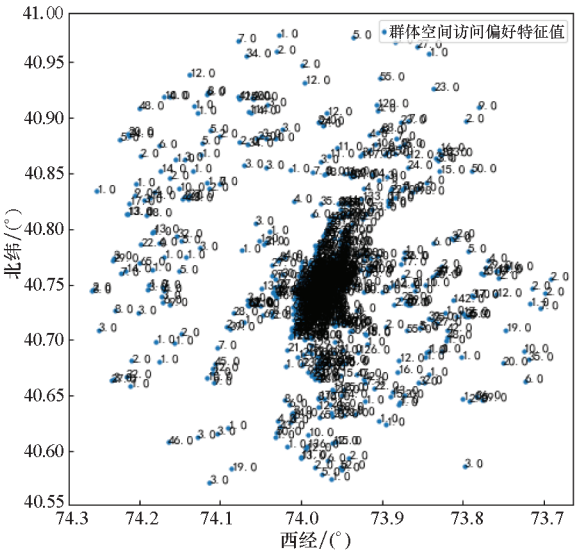

图 3　群体空间访问分布及偏好特征

Fig. 3　Group spatial preference and numerical feature

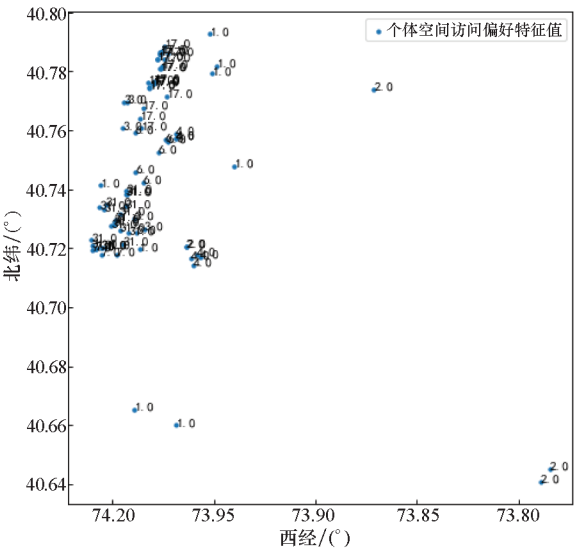

图 4　1 号用户的空间访问分布及偏好特征

Fig. 4　Spatial preference feature of the No. 1 user

2.3　POI 语义信息提取

用户在签到时刻进行的活动往往与其所在位置的类别、功能具有紧密的联系,签到地点信息对个体活动的识别具有重要意义。因此,本文通过提取签到点 POI 名称中的文本信息特征,获取其隐含的活动语义信息。

词嵌入方法在自然语言处理领域得到了广泛的应用,通过将词语映射到一个数学空间里,能够获取文本对应的反映其特点的向量表征。本文使用自然语言处理中的 BERT 模型将签到地点名称转换为带有语义的词向量。

BERT 模型以 Transformer 模型架构为基础,能够在左右两侧上下文的联合条件下,从无标注的文本中预训练出词的深层双向表征。

本文方法使用谷歌公司提供的 Uncased BERT-Base 模型对签到地点名称进行词嵌入,将文本信息映射至特征空间,以获取对应的词向量特征。变换流程如图 5 所示,具体详细如下:

1)基于个体用户签到数据,获取签到 POI 名称文本序列$\{P_1,P_2,\cdots,P_n\}$,分别对位置名称 P_i 进行词嵌入、位置编码、片段编码得到对应的向量表示 $\boldsymbol{X}$。

2)将 $\boldsymbol{X}$ 输入自注意力层,计算其对应的 Query 矩阵、Key 矩阵、Value 矩阵:

$$\boldsymbol{Q}_i=\boldsymbol{X}\cdot\boldsymbol{W}^{\mathrm{Q}} \tag{7}$$

$$\boldsymbol{K}_i=\boldsymbol{X}\cdot\boldsymbol{W}^{\mathrm{K}} \tag{8}$$

$$\boldsymbol{V}_i=\boldsymbol{X}\cdot\boldsymbol{W}^{\mathrm{V}} \tag{9}$$

其中,$\boldsymbol{W}^{\mathrm{Q}}$、$\boldsymbol{W}^{\mathrm{K}}$、$\boldsymbol{W}^{\mathrm{V}}$ 分别代表对应权重矩阵。

3)计算得分,即 $\boldsymbol{Q}_i$ 与 $\boldsymbol{K}_i$ 的点积,通过 softmax 函数对结果进行归一化处理,并乘以 $\boldsymbol{V}_i$ 矩阵得到注意力矩阵 $\boldsymbol{Z}_i$:

$$\boldsymbol{Z}_i=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{K}_i}{\sqrt{d_k}}\right)\boldsymbol{V}_i \tag{10}$$

4)由于多头注意力机制,需要将多个 $\boldsymbol{Z}_i$ 矩阵相连,并与权重矩阵 $\boldsymbol{W}^{\mathrm{O}}$ 进行点乘,$\boldsymbol{Z}$ 即为对应文本的向量表征:

$$\boldsymbol{Z}=Concat(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_m)\cdot\boldsymbol{W}^{\mathrm{O}} \tag{11}$$

5)将结果 $\boldsymbol{Z}$ 保存为对应的嵌入结果,并循环执行完成所有文本到向量的转换,对序列中其他 POI 名称文本循环此过程,获取全部 POI 文本对应的向量表征。

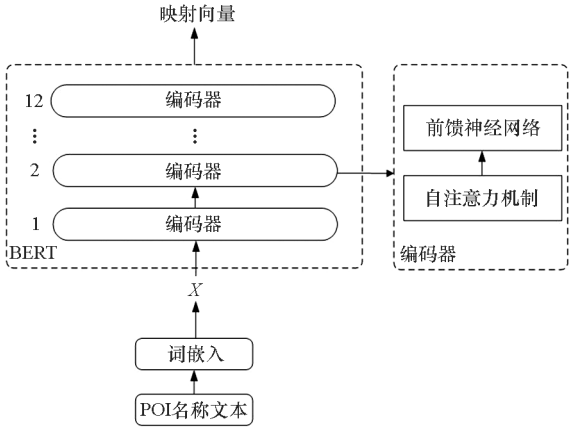

图 5　BERT 模型流程

Fig. 5　Flow chart of BERT model

3　活动语义识别算法

本文将活动语义识别作为一个多分类任务进

行处理,提取了时间、空间、文本三个维度的特征,组合成特征向量,采用 XGBoost 模型作为分类器,构建一种融合多特征的活动识别(multi-feature activities recognition,MFAR)算法模型。模型的输入包括签到数据的用户 ID、签到地点空间经纬度坐标、签到时间戳、签到地点 POI 名称。算法从时间信息中提取月、日、星期、工作日、签到时刻,并计算签到点的空间访问偏好程度;文本上对于签到地点名称,利用 BERT 模型对其进行编码,将字符信息转换为具有文本语义的向量。组合三个维度中的所有特征构成特征向量,利用 XGBoost 模型完成活动语义识别,MFAR 算法具体流程如算法 1 所示。

算法 1 MFAR 算法

Alg. 1 MFAR algorithm

输入:签到事件数据集 D,其中每条签到数据包含用户编号 U、签到地点经纬度坐标 L、签到地点名称文本 P、签到时间戳 T

输出:活动语义识别结果 $\hat{y}$

1. 由签到时间 T 读取月、日、星期特征 T_{month}、T_{day}、T_{week}、T_{weekday},并提取工作日特征,若星期标签属于工作日,则标签 $T_{\text{weekday}}=1$,否则 $T_{\text{weekday}}=0$,将提取所得全部时间特征记为 X_{tem};
2. 读取签到地点经纬度信息 L,提取签到地点的个体、群体空间偏好特征 X_{spa};
3. 使用 BERT 模型将签到地点名称 P 转换为对应词向量特征 X_{tex};
4. 以 $\{X_{\text{tem}}, X_{\text{spa}}, X_{\text{tex}}\}$ 构成联合特征,训练 XGBoost 模型;
5. 调用模型,由输入特征得到识别结果 $\hat{y}$, 其中 $\hat{y}_i = \sum_{i=1}^{k} f_i(x_i)$

与其他个体活动识别方法不同,MFAR 不仅考虑常见形式的时间、空间特征,还提取了空间的个体、群体偏好特征和地点名称文本特征。根据 MFAR 算法的组成,其时间复杂度主要由联合特征提取与 XGBoost 分类器两部分构成。其中,在联合特征提取过程中,个体空间访问偏好特征、时间特征、文本特征提取部分的时间复杂度均为 $O(n)$。XGBoost 分类器的时间复杂度主要由分裂点查找过程、建树过程决定,由于采用了贪婪算法查找决策树分裂点,每次查找都需要进行排序,其时间复杂度为 $O(\|x\|\log n)$;建树过程对已排序的特征列进行线性查找,其时间复杂度为 $O(kd\|x\|)$,二者相加得到分类器时间复杂度为 $O(\|x\|\log n + kd\|x\|)$,其中 d 与 k 分别表示树的总棵数与最大深度,$\|x\|$表示训练数据中所有非缺失项,n 表示数据样本数量。

4 实验与结果

MFAR 的主要任务是识别 LBSN 用户在签到位置的活动语义,实验采用 Foursquare 社交平台公开签到数据集来验证本文方法的有效性,选用的数据集中包括来自纽约的 227 428 条及东京的 573 703条用户签到数据。每条签到记录主要包含匿名的用户 ID、签到位置 ID、位置所属类别 ID、签到位置名称、经纬度坐标、UTC 时差、世界标准时间。

具体活动语义的标签类别及描述如表 1 所示,数据集包含 12 种活动语义标签,描述了用户在签到地点的活动语义,例如:对于一条记录了某用户在健身房锻炼的签到数据,其 POI 名称为“Gym/Fitness Center”,对应活动语义标签为“Sports”。

表 1 活动语义类别及描述

Tab. 1 Tags and descriptions of user activities

活动语义标签	活动语义类别描述	常见 POI
Art	美术、音乐等相关活动	Gallery
Education	教育培训相关活动	School
Entertainment	日常娱乐相关活动	Arcade
Medical	就医、购药等医疗相关活动	Medical Center
Meeting	会议、宗教等集体活动	Church
Rest	休息	Home
Restaurant	外出就餐	Burger Joint
Service	接受社会服务	Bank
Shopping	购物相关活动	Mall
Sports	参加体育锻炼	Gym/Fitness Center
Travel	出行、旅行活动	Subway
Work	工作相关活动	Office

XGBoost 算法中决策树数量 d 与最大深度 k 的超参数对模型性能影响较大。经参数调优,本文将模型学习率设置为 0.3,决策树数量为1 000,树的深度为 6。

此外,本文采用控制变量思想分别调整不同

参数值,计算不同实验设置下模型多分类结果的准确率 S_{acc}、精确率 S_{pre}、召回率 S_{rec}、F1 值 S_{F1}。为避免实验结果偶然性的影响,在对比实验中采用十折交叉验证,取评价指标平均值进行对比。同时,采用混淆矩阵进一步获取模型对不同活动类别的识别情况,并对主要误识别原因进行分析。四种主要评价标准计算公式如下:

$$S_{acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$S_{pre} = \frac{TP}{TP + FP} \tag{13}$$

$$S_{rec} = \frac{TN}{TN + FP} \tag{14}$$

$$S_{F1} = \frac{2 \cdot S_{pre} \cdot S_{rec}}{S_{pre} + S_{rec}} \tag{15}$$

根据模型对样本的分类结果与其实际类别的匹配情况,TP、TN 分别代表模型分类正确的正例与反例样本,即真正例与真反例;FP、FN 分别代表分类错误的正例与反例样本,即假正例与假反例。

本文围绕空间偏好特征、POI 文本特征和分类器进行了一系列对比实验,并对模型性能进行对比分析,以证明 MFAR 提取特征的有效性与分类器的性能。

图 6 展示了引入空间偏好特征对模型性能产生的影响,对比了仅采用时间特征以及同时采用时间与空间偏好特征训练所得分类器的性能指标。对比实验结果,相比较于单一的时间特征,空间特征的引入为模型的活动语义识别性能带来了明显提升。空间访问频率反映了群体和个体的空间访问偏好,由于在群体访问所形成的热点区域(如知名餐厅、名胜古迹等)中大多数个体所进行的活动行为相同,多数个体用户在相应位置进行的活动行为具有较高相重合度。个体访问数据形成的热点访问区域同样会反映个体活动的空间访问习惯偏好,在签到数据中常体现为个体在进行特定活动时多次重复访问某一地点。实验结果证明,空间访问偏好特征对活动语义识别有重要作用。

图 7 展示了文本特征对模型性能的影响,对比了仅采用时空特征、分别采用 LDA 主题模型与 BERT 模型进行词向量嵌入的模型性能,三种特征选取方式分别对应三种不同颜色柱状图。对比时间空间相结合的特征,在加入了文本作为识别特征后,识别准确率又有了较高的提升。在活动语义识别中文本特征是一个重要的特征,活动点的名称与个体在活动点进行的活动行为有重要联系。对于文本特征,本文通过两种不同的思路来挖掘:①LDA 主题模型,从词到主题的分布来对文本进行向量的表征;②BERT 模型,从词的语义来进行向量表征。在通过 LDA 模型进行词向量的转换时,首先以所有签到点的名称进行训练生成主题,之后对所有签到地点名称进行主题归属预测,得到主题归属预测向量,完成词到向量的编码。BERT 模型将签到地点经过一个 12 层的编码器结构来完成语义表征。图 7 的实验验证了基于 BERT 的语义表征优于基于 LDA 的主题表征。在活动语义识别中,同一地点的活动行为不尽相同,LDA 主题模型对文本以主题语义的形式进行词向量表征,但主题语义并不等同于活动语义。

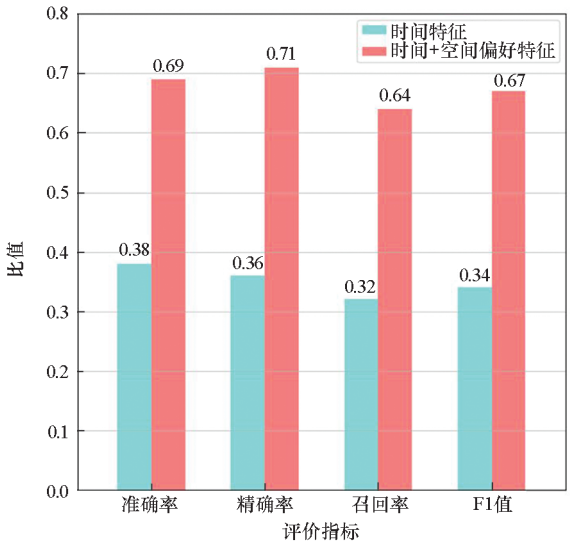

图 6　关于空间偏好特征的模型性能对比

Fig. 6　Comparison of the model performance in view of spatial preference features

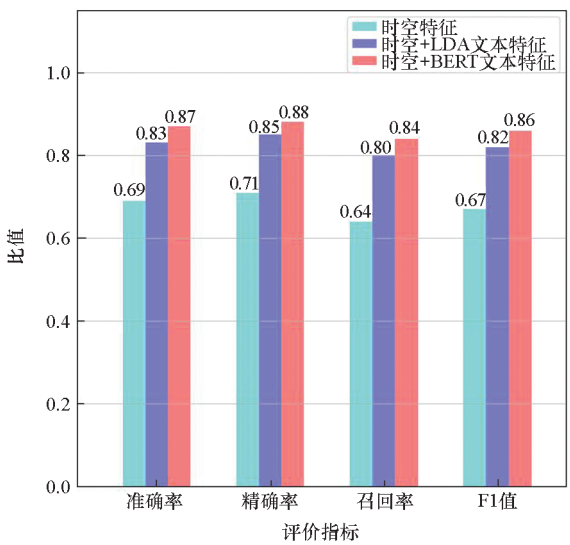

图 7　关于文本特征提取的模型性能对比

Fig. 7　Comparison of model performance in view of text feature extraction

BERT 模型对文本以自然语言语义的形式进行词向量表征，相同字符在不同语境中存在不同含义，而主题则更加确定，所以对于相同地点的不同活动行为经由 BERT 模型转化来的向量有着更好的表达效果，因此该模型更有利于活动语义识别。

将 MFAR 算法与只采用时间、空间、文本单一特征进行训练得到的识别准确率进行对比。其中，单一采用时间特征所得准确率为 0.38，空间特征准确率为 0.48，文本特征准确率为 0.78，MFAR 算法准确率为 0.87，结果显示 MFAR 算法表现优于其中任一种特征。此外，为探究算法中各特征的作用，基于单一特征的识别准确率定义了特征重要性度量，计算方法如式(16) ~ (17)所示。

$$I_i = \frac{S_i}{S_{\text{total}}} \cdot S_{\text{MFAR}} \tag{16}$$

$$S_{\text{total}} = S_{\text{spa}} + S_{\text{tem}} + S_{\text{tex}} \tag{17}$$

其中，I_i 代表不同特征类别的重要度分值，S_{tem}、S_{spa}、S_{tex} 分别表示基于时间、空间、文本特征单独训练所得模型的准确率，S_{MFAR} 表示同时采用三种特征的 MFAR 算法模型准确率。

根据式(16)进行计算，时间特征重要度为 0.20，在三种特征中最低；空间特征重要度为 0.26，略高于时间特征；文本特征重要度最高，为 0.41。根据上述结果，签到位置文本信息在活动识别中发挥了更重要的作用。进行打卡签到时用户更倾向于记录新颖活动，对日常生活频繁活动记录较少，整个打卡签到行为随机性大，导致时间信息中周期性和序列性行为较难挖掘，因此单一时间维度特征识别效果较差。空间偏好信息相比较于时间信息识别较好的原因是，对于空间中的热点访问区域，多数个体在这些地区都进行相同的活动行为。签到地点名称文本反映了签到地点的固有属性，这些固有属性决定了当前地点能够提供何种活动行为，因此识别效果最好。

如图 8 所示，实验对比了 XGBoost 分类器、基于随机森林的分类器、K 近邻分类算法以及支持向量机在采用相同特征前提下的模型性能，结果表明 MFAR 采用的 XGBoost 分类器的各项性能指标均显著优于其他分类器。

图 9 为活动语义识别模型的混淆矩阵，混淆矩阵展示了算法对不同活动语义的预测情况，其中横轴代表 MFAR 算法根据签到数据特征所得的活动语义预测值，纵轴代表该签到数据包含的活动语义真实值，位于矩阵对角线上的数字代表活动语义预测值与真实值相符的签到数据样本数

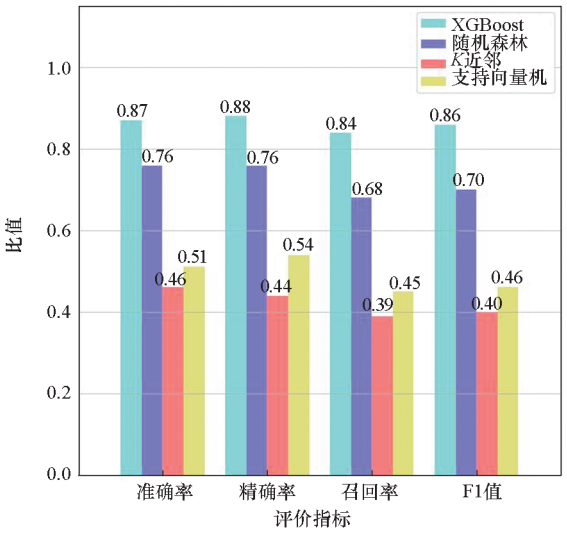

图 8 采用不同分类器的模型性能对比

Fig. 8 Comparison of model performance with different classifiers

量。例如：在活动语义真实值为 Shopping 的样本中，有 7 972 条签到数据预测正确，其余预测值与真实值不符，其中 273 条被预测为 Entertainment；在对活动语义真实值为 Entertainment 样本的预测结果中，有 250 条被预测为 Shopping。

观察混淆矩阵不难发现，模型对 Education、Medical、Service 等活动类别识别较为准确，此类标签在数据集中通常对应具有特殊性的签到地点，例如教学楼、诊所、药店、政府大楼、银行等。由于活动语义、签到地点间的关联性及地点本身的特殊性，通过 POI 名称文本隐含的语义信息即可实现一定程度的活动判断，并结合签到时空特征进一步提升识别准确性。同时，模型在 Entertainment 类、Restaurant 类和 Shopping 类活动样本间出现了较多误识别情况。通过观察数据集相关样本，能够发现出现误识别的活动类别间存在时空重叠，即用户进行几种活动的时间、空间信息相似度较高。另外，Entertainment 类活动语义范围广泛且模糊，部分样本的时空特征与 Restaurant 类、Shopping 类活动并不存在明显界线，相关活动类别的区分主要根据签到地点名称文本进行，在文本信息不足以体现区别时易造成混淆。

为了进一步验证模型性能，将 MFAR 算法与位置感知 Dirichlet 分配(location-aware latent Dirichlet allocation，LLDA)算法[2]、相似用户模式(similar user pattern，SUP)算法[21]和多层感知机(multi-layer perception，MLP)基线模型进行对比，通过评价指标及十折交叉验证标准差衡量算法的

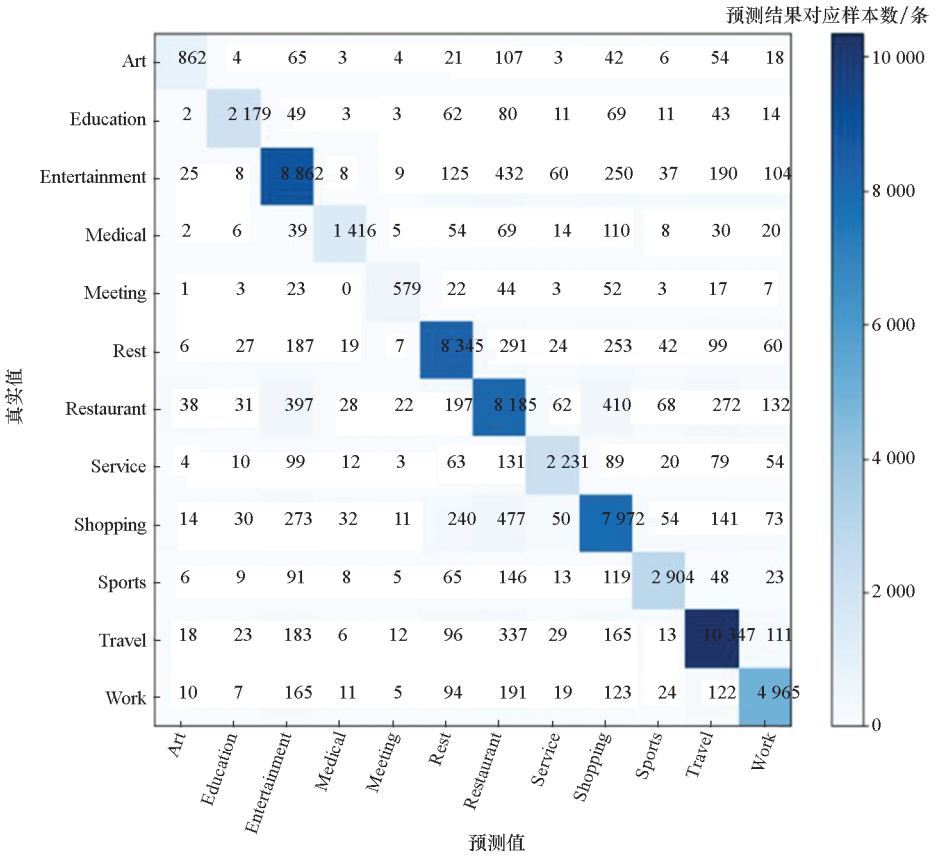

图 9　MFAR 算法识别混淆矩阵

Fig. 9　Identify confusion matrix of MFAR

个体活动识别性能。其中 SUP 算法对相似位置进行分类,获取用户偏好相似度及用户间签到活动的关联,引入了名为相似用户模式的位置特征,通过多个分类器对位置语义标签进行识别。实验采用 SUP 算法提取用户特征,并结合其他特征识别个体活动。LLDA 算法采用 DBSCAN 算法对空间位置进行聚类,并结合 LDA 主题生成模型获取区域及潜在活动的主题分布,建立用户兴趣评分模型,将最高分值对应活动类别作为用户活动语义。此外,作为基线对比实验,本文采用 MLP 构建分类器模型,各隐藏层之间全连接并在各神经元添加 ReLU 激活函数,将联合特征向量输入模型,最终由 softmax 层输出属于不同类别的概率,取概率最大类别作为分类结果。

对比实验结果如表 2 所示,MFAR 算法在识别准确率上相比较于 SUP 算法提高了 42 个百分点,相较于 LLDA 算法提高了 11 个百分点,在和 MLP 基线模型的对比中准确率也提升了 4 个百分点。

基于 Foursquare 东京数据集的活动语义识别算法对比实验结果如表 3 所示,在该数据集上,本文 MFAR 算法的识别准确率相较于 SUP 算法提高了 43 个百分点,相较于 LLDA 算法提高了 10 个百分点,与 MLP 基线算法实验结果相比,准确率提升了 26 个百分点。

表 2　纽约市数据集的活动语义识别算法对比结果

Tab. 2　Comparison of several activity semantics recognition algorithms on New York City dataset

算法	准确率	精确率	召回率	F1 值	标准差
SUP	0.45	0.44	0.39	0.38	0.000 2
LLDA	0.76	0.73	0.68	0.70	0.000 3
MLP	0.83	0.80	0.74	0.75	0.000 2
MFAR	0.87	0.88	0.84	0.86	0.000 2

表 3　东京数据集的活动语义识别算法对比结果

Tab. 3　Comparison of several activity semantics recognition algorithms on Tokyo dataset

算法	准确率	精确率	召回率	F1 值	标准差
SUP	0.53	0.41	0.42	0.41	0.000 3
LLDA	0.86	0.85	0.82	0.83	0.000 2
MLP	0.70	0.68	0.66	0.67	0.000 2
MFAR	0.96	0.95	0.95	0.94	0.000 2

对比两数据集实验结果能够发现,本文算法及 LLDA 的准确率在表 3 所示实验中有明显提升,推测为东京数据集数据量较纽约市数据集更大,使得本文模型的训练更为充分。而 SUP 算法性能不佳的可能原因在于稀疏签到数据集不易获取活动及用户的相似性。此外,对比结果也表明在个体活动语义表达上,MFAR 提取的文本及空间特征优于 LLDA 挖掘的空间信息与潜在活动主体分布。

综合上述实验结果,本文提出的 MFAR 算法在基于 Foursquare 数据集的个体用户活动识别任务中具有更好的表现。

5 结论

本文提出了一种结合空间偏好和 POI 语义信息的个体活动语义识别算法。重点研究了空间访问偏好和 POI 语义对个体活动语义识别的影响,并且通过实验对比验证了这两个特征在活动语义识别中的作用,与其他算法的对比也证明了本文算法的性能优势。本文在特征挖掘中还存在一些不足,对空间信息的挖掘中主要通过区域访问频率来反映个体对空间区域的偏好,缺少对群体空间访问偏好与个体空间访问偏好之间的联系的挖掘,因此,未来工作中将就社交网络关系等群体与个体活动的潜在关联进一步融合到特征表示中进行研究,提高模型的活动识别能力。

参考文献(References)

[1] CHO E, MYERS S A, LESKOVEC J. Friendship and mobility: user movement in location-based social networks [C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011: 1082 – 1090.

[2] ZHU Z Q, CAO J X, WENG C H. Location-time-sociality aware personalized tourist attraction recommendation in LBSN[C]// Proceedings of IEEE 22nd International Conference on Computer Supported Cooperative Work in Design, 2018: 636 – 641.

[3] QIAO Y Q, LUO X Y, LI C L, et al. Heterogeneous graph-based joint representation learning for users and POIs in location-based social network[J]. Information Processing & Management, 2020, 57(2): 102151.

[4] CAO K Y, GUO J J, MENG G J, et al. Points-of-interest recommendation algorithm based on LBSN in edge computing environment[J]. IEEE Access, 2020, 8: 47973 – 47983.

[5] GONZÁLEZ M C, HIDALGO C A, BARABÁSI A L. Understanding individual human mobility patterns [J]. Nature, 2008, 453: 779 – 782.

[6] VAN WEERDENBURG D, SCHEIDER S, ADAMS B, et al. Where to go and what to do: extracting leisure activity potentials from Web data on urban space[J]. Computers, Environment and Urban Systems, 2019, 73: 143 – 156.

[7] CAI L, XU J, LIU J, et al. Sensing multiple semantics of urban space from crowdsourcing positioning data[J]. Cities, 2019, 93: 31 – 42.

[8] HUANG W, XU S S, YAN Y W, et al. An exploration of the interaction between urban human activities and daily traffic conditions: a case study of Toronto, Canada[J]. Cities, 2019, 84: 8 – 22.

[9] YANG C, BAI L X, ZHANG C, et al. Bridging collaborative filtering and semi-supervised learning: a neural approach for POI recommendation [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 1245 – 1254.

[10] RIZWAN M, WAN W G, GWIAZDZINSKI L. Visualization, spatiotemporal patterns, and directional analysis of urban activities using geolocation data extracted from LBSN[J]. ISPRS International Journal of Geo-Information, 2020, 9(2): 137.

[11] ARAÚJO F, ARAÚJO F, MACHADO K, et al. Ensemble mobility predictor based on random forest and Markovian property using LBSN data[J]. Journal of Internet Services and Applications, 2020, 11(1): 1 – 11.

[12] MARTÍ P, SERRANO-ESTRADA L, NOLASCO-CIRUGEDA A. Social media data: challenges, opportunities and limitations in urban studies[J]. Computers, Environment and Urban Systems, 2019, 74: 161 – 174.

[13] KIM J S, JIN H, KAVAK H, et al. Location-based social network data generation based on patterns of life [C]// Proceedings of 21st IEEE International Conference on Mobile Data Management, 2020: 158 – 167.

[14] KAVAK H, KIM J S, CROOKS A, et al. Location-based social simulation[C]//Proceedings of the 16th International Symposium on Spatial and Temporal Databases, 2019: 218 – 221.

[15] ZHONG H D, LYU H B, ZHANG S Z, et al. Measuring user similarity using check-ins from LBSN: a mobile recommendation approach for e-commerce and security services[J]. Enterprise Information Systems, 2020, 14(3): 368 – 387.

[16] WANG Y, QIN Z X, PANG J, et al. Semantic annotation for places in LBSN through graph embedding[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017: 2343 – 2346.

[17] ZHANG Y F, LIU A, LIU G F, et al. Deep representation learning of activity trajectory similarity computation [C]// Proceedings of IEEE International Conference on Web Services, 2019: 312 – 319.

[18] SHI H Z, LI Y, CAO H C, et al. Semantics-aware hidden Markov model for human mobility[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(3): 1183 – 1194.

[19] PAN X, HU R M, LI D S. Social-IFD: personalized influential friends discovery based on semantics in LBSN[C]// Proceedings of IEEE International Conference on Communications, 2020: 1 – 6.

[20] PAPANGELIS K, CHAMBERLAIN A, LYKOURENTZOU I, et al. Performing the digital self: understanding location-based social networking, territory, space, and identity in the city[J]. ACM Transactions on Computer-Human Interaction, 2020, 27(1): 1 – 26.

[21] LI Y H, ZHAO X G, ZHANG Z, et al. Annotating semantic tags of locations in location-based social networks [J]. GeoInformatica, 2020, 24(1): 133 – 152.