引用格式：王培晓,王海波,傅梦颖,等.室内用户语义位置预测研究[J].地球信息科学学报,2018,20(12):1689-1698. [Wang P X, Wang H B, Fu M Y, et al. Research on semantic location prediction of indoor users[J]. Journal of Geo-information Science, 2018,20(12):1689-1698.] **DOI:** 10.12082/dqxxkx.2018.180411

室内用户语义位置预测研究

王培晓[1,2],王海波[3],傅梦颖[1,2],吴　升[1,2*]

1. 福州大学 福建省空间信息工程研究中心,福州 350002;2. 海西政务大数据应用协同创新中心,福州 350002;3. 湖北工业大学经济与管理学院,武汉 430068

Research on Semantic Location Prediction of Indoor Users

WANG Peixiao[1,2], WANG Haibo[3], FU Mengying[1,2], WU Sheng[1,2*]

1. Spatial Information Research Center of Fujian Province, Fuzhou University, Fuzhou 350002, China; 2. Fujian Collaborative Innovation Center for Big Data Applications in Governments, Fuzhou 350002, China; 3. Economic and management school, Hubei University of Technology, Wuhan 430068, China

Abstract: The location prediction technology can predict the location of the user at the next moment in advance, and plays an extremely important role in the field of Location-based Service (LBS).Most of the existing location prediction techniques only use the geographical location information and time information of the user's historical trajectory. The geographic trajectory is composed of a series of geographically-pointed time-stamped latitude and longitude points, and the geographic trajectory only mines users. Mobile mode is limited by geographic features. In this paper, we propose a novel approach for predicting the next semantic location of a user's movement based on the geographic and semantic characteristics of the group user trajectory. The semantic location prediction based on group users generally consists of three steps: Firstly, the specific algorithm is used to identify the staying area in the user's trajectory; Next, the semantic matching algorithm is used to associate the user's staying area with the semantic information; Finally, Mining the semantic location pattern of group users, using this pattern to predict the semantic location of the user at the next moment. In the stage of staying area identification, in order to reduce the influence of indoor stay time unfixed on the recognition of stay area, this paper proposes a new type of spatial-temporal agglomerative nesting (ST-AGNES), which can automatically identify the number of staying areas in the user's trajectory using only the distance threshold. In the semantic matching stage, this paper proposes a semantic matching method based on attractance rules, which makes uses all trajectory points in the stay area to be associated with indoor high-density semantic information. In the final forecasting stage, this paper uses Long Short-Term Memory (LSTM) to mine the semantic location patterns of group users and predict the future semantic location of users. The experimental results have achieved a prediction accuracy rate of 61.3%.

收稿日期：2018-08-31；**修回日期**：2018-10-29.

基金项目：国家重点研发计划项目(2017YFB0503500)；数字福建建设项目(闽发改网数字函(2016)23号)；湖北省教育厅人文社会科学研究项目(17Q071)。[**Foundation items:** National Key Research and Development Program of China, No.2017YFB0503500; Digital Fujian Program, No.2016-23; Hubei Provincial Education Department Humanities and Social Sciences Research Project, No.17Q071.]

作者简介：王培晓(1994-),男,山东济南人,硕士生,研究方向为地理信息服务、时空数据挖掘等。E-mail: 260129327@qq.com

***通讯作者**：吴　升(1972-),男,福建松溪人,工学博士,教授,研究方向为时空数据分析与可视化、信息共享与智慧政务、应急信息系统等。E-mail: ws0110@163.com

Key words: LSTM; ST-AGNES; attraction rule; indoor trajectory; location prediction
***Corresponding author:** WU Sheng, E-mail: E-mail: ws0110@163.com

摘要:位置预测技术可以提前预知用户下一时刻的位置,在基于位置的服务(Location-based Service,LBS)领域中发挥着极其重要的作用。现有的位置预测技术大多仅使用用户的地理轨迹,仅使用地理轨迹挖掘出来的用户移动模式易受地理特性的限制缺乏深层次的语义信息。本文基于某商场群体用户的室内轨迹数据和语义信息预测用户下一个时刻语义位置。语义位置预测包括停留区域识别、停留区域语义匹配、语义位置建模。在停留区域识别阶段,为减少室内停留时间不固定对停留区域识别的影响,本研究提出了一种新型的时空凝聚层次聚类算法(Spatial-Temporal Agglomerative Nesting, ST-AGNES),该算法具有思想简单、超参数少、自动生成聚类个数等优点。在语义匹配阶段,引入了吸引度规则,充分利用停留区域所有轨迹点与室内高密度的商铺名称信息做匹配。最后,采用长短型记忆神经网络模型(Long Short-Term Memory,LSTM)挖掘群体用户的语义位置模式并预测用户未来的语义位置,实验预测正确率达到61.3%。
关键词:LSTM模型;ST-AGNES算法;吸引度规则;室内轨迹;位置预测

1 引言

近年来,随着移动便携设备的普及和各种室内外定位技术的快速发展,获取用户实时位置信息成为可能。基于位置的服务(Location-based Service,LBS)也因此逐渐成为研究热点。位置预测研究是LBS研究的重要组成部分,受国内外研究学者的关注,该技术可根据用户的历史轨迹数据推断用户下一时刻的位置,从而为用户提供更加灵活的服务,如推荐服务[1]、提醒服务、智能化交通服务[2]等。

位置预测技术可以分为基于个人的位置预测和基于群体用户的位置预测[3-4]。基于个人的位置预测需要收集独立的用户信息,为每个用户产生独特的轨迹模式,多用于预测某用户独特的运动规律;基于群体的位置预测识别不同用户间的相似轨迹路径为相似用户创建通用的轨迹模式,多用于预测群体用户之间的相似行为。现有的位置预测研究大部分仅使用用户历史轨迹的地理位置信息和时间信息进行位置预测,地理轨迹是由一系列带有时间戳且由经纬度标记的地理位置点组成,仅由地理轨迹挖掘用户移动模式受地理特性的限制[3]。因此位置预测研究需要一种表达能力更强、更符合用户习惯的概念,即语义位置[5]。语义位置是一种以人为中心的位置表达方式,其隐含了与用户相关的深层次的知识(如目的意图、生活习惯、社会关系等)。语义位置预测包括:① 寻找用户轨迹中的用户停留区域;② 将用户停留区域标注上语义信息得到用户语义位置;③ 挖掘用户语义位置中存在的模式,利用该模式预测用户下一时刻的语义位置。目前,众多国内外学者建立了多种算法模型对用户进行位置预测:Jeung等[6]通过改进的Apriori算法预测用户的未来位置;Ye等[7]提出了个人生活模式用于描述单用户的周期性行为;Morzy等[8]使用改进的PrefixSpan算法挖掘用户频繁模式并预测用户的位置;郑宇等[9]建立了HITS的模型挖掘用户感兴趣的位置模式并预测用户位置。上述研究仅根据用户地理位置进行预测,并没有融合位置的语义信息,也有一些学者针对语义位置做了相关研究:Alvares等[10]提出了SMoT模型研究用户轨迹与语义信息的关联关系;窦丽莎等[11]在Alvares基础上使用SMoT模型推断用户的出行目的;齐凌艳[12]针对SMoT模型做了改进提高了语义匹配的准确度;Li等[13]从语义位置相似度的角度出发,提出了HGSM模型预测用户语义位置;宋路杰等[14]、彭曲等[15]、林树宽等[16]认为用户语义位置存在上下文相关性,采用马尔科夫及其变种模型预测用户语义位置;张心悦等[17]通过LDA主题模型对群体用户进行情感分类,后采用PrefixSpan挖掘用户语义位置的关联规则。但上述语义位置的研究多侧重于室外,室内的语义位置研究相对较少。由于室内位置密度高和位置的停留时间不固定等原因,室内空间中的语义位置预测仍是一个具有挑战性的问题。

本文旨在根据商场室内群体用户的轨迹数据挖掘相似用户之间的语义位置模式。首先,为避免停留时间对停留区域识别的影响,提出了一种新型的时空凝聚层次聚类算法(Spatial‐temporal agglomerative nesting,ST-AGNES),该算法仅需距离阈值即可识别轨迹中的用户停留区域;然后,针对室内空间位置高密度的特点,引入了一种基于吸引度规则的语义匹配方法,该方法利用停留区域内部的所有轨迹信息将语义信息与停留区域相关联;最后,采用长短型记忆(Long Short-Term Memory,LSTM)神

经网络模型对群体用户的语义位置建模并预测语义位置，从而有助于商场挖掘用户潜在的购物倾向，提高商场精准营销能力。

2　语义位置预测流程

语义位置预测的流程如图1所示，主要分为以下4步：① 数据清洗，去除原始轨迹中异常、冗余、错误等数据；② 采用ST-AGNES算法识别用户停留区域序列；③ 采用吸引度规则将所有用户的停留区域序列与商铺名称信息相关联，得到所有用户的语义轨迹；④ 先使用LSTM模型对语义轨迹建模，再根据用户的已知轨迹预测下一时刻的位置。

定义1：轨迹点 $pt=(macId, t, loc)$，pt 是移动设备采集到的位置点，$macId$ 是用户的唯一标识ID，t 代表该位置信息采集到的时间，$loc=(x, y, f)$ 代表该用户在 t 时刻的位置（x 表示经度，y 表示纬度，f 表示该用户所处的楼层ID）。

定义2：轨迹序列 $traj=\{pt_i\}$，单用户原始轨迹点清洗后按时间顺序排列的轨迹点称为用户的轨迹序列 $traj$。

定义3：用户停留区域 $stayArea=(startIndex, endIndex, deltaT \mid deltaT>timeThreh)$，用户在某区域内停留时间超过一定阈值的区域称为停留区域。$startIndex$ 表示停留区域中的起始轨迹点，$endIndex$ 表示停留区域中终止轨迹点，$deltaT$ 表示用户在该区域的停留时间，$timeThreh$ 表示时间阈值。

定义4：用户停留区域序列 $ST\_Seq=\{stayArea_i\}$，在用户轨迹中，将用户停留区域按时间顺序连接得到用户停留区域序列。

定义5：语义位置[5] $sema\_loc=(store, address)$，$store$ 表示某位置的语义信息，$address$ 表示语义信息的使用范围，如（Nike，北京市朝阳区万达二楼），在具体应用中 $address$ 往往被隐含的约定，在不产生歧义的情况下可省略。

定义6：用户语义轨迹[3] $sema\_traj=\{sema\_loc_i\}=\{(store_i, address_i)\}$，由用户语义位置按时间顺序连接得到语义轨迹，用户语义位置 $sema\_loc_i$ 由用户停留区域 $stayArea_i$ 语义匹配得到，当 $address$ 被省略时，语义轨迹可表示为 $sema\_traj=\{store_i\}$。

定义7：单点吸引度序列 $local\_attract=\{(store_i, p_i)\} \sum p_i=1\}$，单个轨迹点受不同商铺的吸引程度序列。其中，$(store_i, p_i)$ 代表轨迹点有 p_i 的概率被 $store_i$ 吸引。

定义8：区域吸引度序列 $reg\_attract=\{(store, attract_i)\}$，停留区域受不同商铺的吸引程度序列。其中，$attract_i$ 为停留区域内所有轨迹点受 $store_i$ 吸引的概率累加，$attract_i$ 最大的商铺即为该停留区域的语义位置。

3　研究方法

3.1　基于时空约束的凝聚层次聚类算法

轨迹序列 traj 中的各轨迹点具有不同的重要程

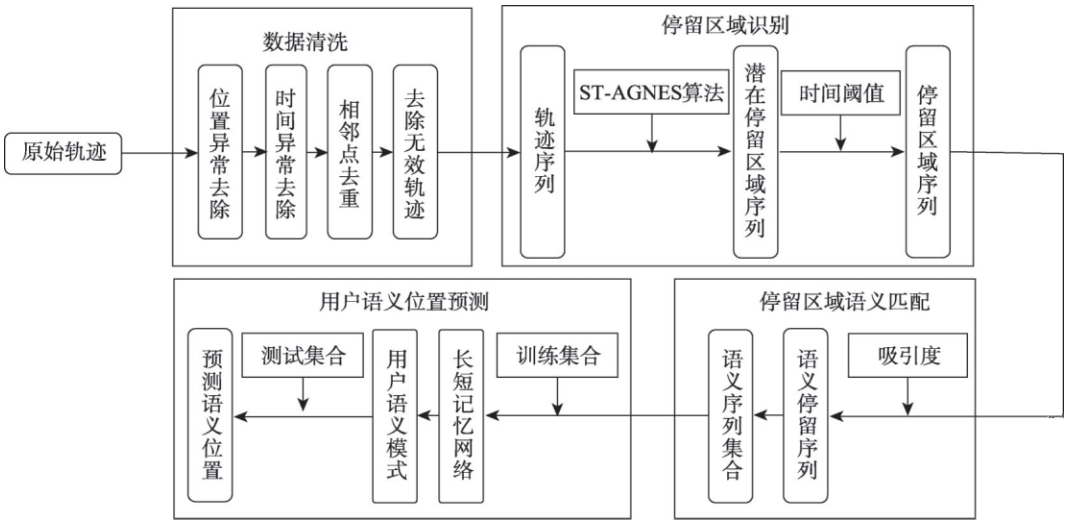


图1　室内用户位置预测总体流程

Fig. 1　Location prediction process

度，如图2所示，用户处于停留区域stayArea内部时有更大的概率查看商铺服务信息，因此停留区域内部的轨迹点的重要程度比外部轨迹点更高。目前停留区域识别算法主要应用了聚类算法，如Ashbrook等[10,18]采用传统的K-means算法和DBSACN算法识别停留区域。传统聚类算法通常只考虑了轨迹点的空间属性，忽略了时间属性对停留区域识别的影响。Zheng等[19]、Birant等[20]、Leiva等[21]提出了启发式算法、ST-DBSCAN算法和WKM算法聚类时空数据，但上述算法存在全局密度阈值、超参数过多、事先指定簇集个数等[22]缺点。针对上述缺点，本文提出了ST-AGNES算法。

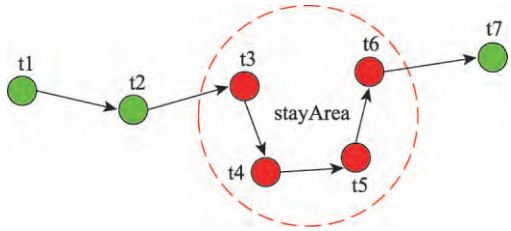


图2 用户停留区域

Fig. 2 Stay area

聚类是将 n 个 d 维向量 $X=\{x_1, x_2, \cdots, x_n\}$ 划分为 k 个不相交类 $\{C_1, \cdots, C_k\}$ 的一种方法[21]。传统的凝聚层次聚类算法（Agglomerative Nesting，AGNES）首先将每一个样本点 x_i 当做一个簇集 C_i，然后采用Linkage（Single Linkage、Complete Linkage、Average Linkage）方式计算任意2个簇集 C_i 和 C_j 之间的距离，通过迭代将最近的2个簇集合并成一个簇集，直到簇集个数等于 k 为止。ST-AGNES算法是AGNES算法的改进算法，该算法将时间顺序分布的数据集X划分为多个不相交顺序簇集 $\{C_1, C_2, C_3, \cdots\}$。如图3所示，$b_i$ 是簇 C_i 的左边界，即簇集 C_i 中第一个样本的索引，引入簇集边界索引b，顺序簇集 C_i 可以表示为 $\{x_{b_i}, \cdots, x_{b_{i+1}-1}\}$，由于ST-AGNES算法中存在时间约束，簇集 C_i 只能沿时间轴向前（簇集 C_{i-1}）或向后（簇集 C_{i+1}）合并，从而解决了AGNES仅考虑空间距离聚类时空数据的缺点，保证了簇集结果的时间连续性。其次，ST-AGNES算法采用距离阈值 dis_{thred}（相邻簇集距离均大于 dis_{thred}）作为算法的终止条件，避免了事先指定簇集个数 k 的局限性。本文将未加入时间阈值条件得到的聚类结果称为用户潜在停留区域序列，在聚类结果的基础上使用时间阈值 T_{threh} 过滤，得到最终的用户停留区域序列。ST-AGNES算法的具体流程如下：

（1）输入时间连续的用户轨迹 $traj=\{pt_1, pt_2, \cdots, pt_n\}$，将每一个轨迹点初始化为一个簇集，簇集的边界索引集合 $B=\{b_1, b_2, \cdots, b_n\}=\{1, 2, \cdots, n\}$。

（2）计算相邻簇集之间的距离，得到距离序列 $dist=\{d_{i,\ i+1}\}$，$d_{i,\ i+1}$ 是簇集 C_i 和簇集 C_{i+1} 之间的距离。

（3）寻找dist中的最小值 $d_{\min}$，如果 $d_{\min}$ 小于距离阈值 d_{threh}，将最近两个簇集合并，更新边界索引集合 B，重新计算相邻簇集之间的距离序列 $dist$，如果 $d_{\min}$ 大于距离阈值 dis_{threh}，得到最终的簇集边界索引集合 B，否则跳转到步骤（3）。

（4）根据集合 B 得到用户潜在停留序列，去除用户潜在停留序列中不满足停留时间阈值 T_{threh} 的区域，得到用户停留区域序列。算法实现伪代码如图4所示。

ST-AGNES算法与现有的启发式算法、ST-DBSCAN算法、WKM算法相比主要有如下优点：① 基于层次的聚类方法，不存在全局密度阈值；② 仅具有一个超参数 dis_{threh}（T_{threh} 不参与聚类结果的生成）；③ 通过超参数 dis_{threh} 自动生成簇集的个数不需要事先指定。

3.2 基于吸引度规则的语义匹配方法

停留区域的语义匹配是语义位置预测的前期准备工作，传统的语义匹配方式[13]首先计算停留区域轨迹点的算术平均值，然后与距离算术平均值最

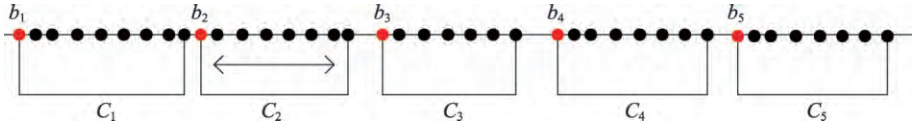


图3 按时间顺序分布的数据集X

Fig. 3 Time-series dataset X

算法 1 基于时间序列的凝聚层次聚类

输入: 用户轨迹$traj=\{p_i\}$，距离阈值$disThreh$

输出: 簇集的边界$boundaries=\{b_i\}$

```
function ST_AGNES(traj, distThreh)
    //将每一个轨迹点初始化为一个簇集
    //边界索引集合初始化为每一个点的索引
    boundaries = initialBoundaries(traj)
    //计算相邻簇集之间的距离(本文使用Average Linkage计算方法)
    disArr = distance(boundaries)
    minDis = disArr.min()
    while minDis < disThreh do
        //将最近的两个簇集合并（需要做四件事）
        //1.在boundaries对象中查找需要合并的两个簇集（新簇集）
        //2.计算新簇集与上一个簇集之间的距离，并更新到disArr
        //3.计算新簇集与下一个簇集之间的距离，并更新到disArr
        //4.在boundaries中删除被合并簇集的起始索引
        mergeCluster(boundaries, minDis, disArr)
        minDis = disArr.min()
    end while
    return boundaries
end function
```

图4　时空凝聚层次聚类算法

Fig. 4　Spatio-temporal agglomerative nesting

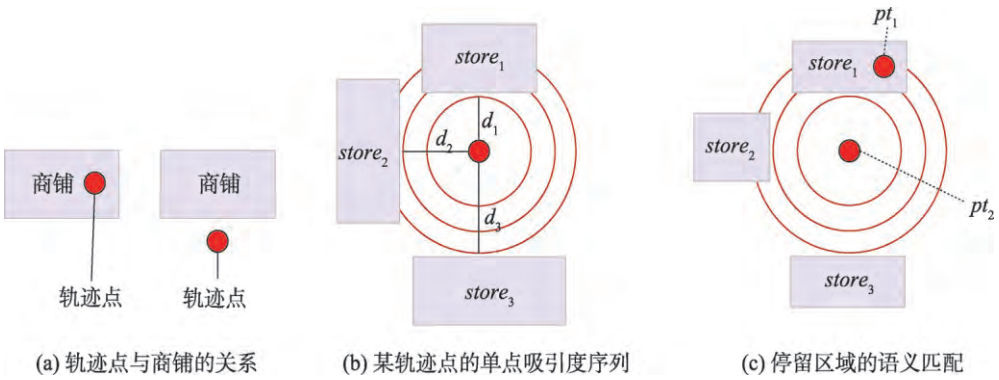
近的标志性建筑物信息做匹配。但算术平均值大概率位于所有点的中央，很容易落在实际停留范围之外[12]，在商铺密集的商场内，此种匹配方式将会导致较大的匹配误差。因此，本文针对室内商铺相距较近的特点，提出一种基于吸引度规则的语义匹配方法。

用户停留区域的每一个轨迹点与商场内的商铺存在2种空间关系，即包含和未包含，如图5(a)所示，当停留区域内部的轨迹点落在商铺内部时，可认为该轨迹点仅被该商铺所吸引，即$local_attract=\{(store,1)\}$。对停留区域中落在商铺外面的轨迹点使用同心圆相切法计算当前轨迹点的单点吸引度序列，即以当前轨迹点为圆心，以半径$r_i(i=1,2,3,\cdots)$画圆，当轨迹点到商铺$store_i$的距离与半径r_i相同时，该圆与商铺相切。以切到商铺的顺序对商铺吸引度排序，求得与轨迹点最近的前n间商铺$\{store_1, store_2, \cdots, store_n\}$，此时该轨迹点的单点吸引度序列由这$n$家商铺共同计算得到。如图5(b)所示，轨迹点由3间商铺共同吸引，该轨迹点的单点吸引度序列为$local_attract=\{(store_1,p_1),(store_2,p_2),(store_3,p_3)\}$，其中$p_i$的计算过程如式(1)所示。

$$p_i=\frac{\left(1/d_i\right)}{\sum_{j=1}^{n}\left(1/d_j\right)} \tag{1}$$

式中：p_i代表轨迹点被第i家商铺吸引的概率；d_i代表轨迹点到第i家商铺的距离。

用户的停留区域stayArea由若干个用户轨迹点组成，每一个轨迹点的单点吸引度序列共同组成该区域的区域吸引度序列，如图5(c)所示，停留区域stayArea中存在2个位置$\{pt_1, pt_2\}$，其中pt_1落在商铺$store_1$的内部，那么pt_1被商铺$store_1$唯一吸引，其单点吸引度序列为$\{(store_1,p_1)\}$，其中$p_1=1$；而pt_2落在商铺外部，此时采用同心圆相切法求得pt_2单点吸引度序列为$\{(store_1,p_2),(store_2,p_3),(store_3,p_4)\}$，那么各商铺对停留区域的区域吸引度序列可表示为$reg_attract=\{(store_1,attract_1),(store_2,attract_2),(store_3,attract_3)\}$，其中$attract_1=p_1+p_2$，$attract_2=p_3$，$attract_3=p_4$。最大的$attract$值对应的商铺$store$即为该用户停留区域的语义信息，用户的每一个停留区域都将唯一对应一个语义位置，将用户停留区域序列中的每一个停留区域语义



(a) 轨迹点与商铺的关系　(b) 某轨迹点的单点吸引度序列　(c) 停留区域的语义匹配

图5　用户停留区域的语义匹配

Fig. 5　Semantic matching of user stay areas

匹配后得到用户的语义轨迹 sema_ traj 。

3.3 基于LSTM的语义位置预测模型

经语义匹配后的语义轨迹 sema_traj 在一定程度上反映了该用户的兴趣爱好和购物习惯，所有用户的语义轨迹组合在一起即可挖掘群体用户的行为模式，从而预测相似用户下一时刻的位置。传统的时序数据预测多采用马尔科夫模型或标准的循环神经网络模型（Recurrent Neural Network，RNN）[14−15,23]，但这些算法的记忆状态有限，难以预测长时序数据，为解决长时序数据预测的问题，本文采用LSTM模型预测用户的语义位置。在LSTM中，采用3种“门”（遗忘门、输入门和输出门）结构增强了模型的记忆能力，如图6所示，遗忘门 f_t 决定从细胞状态中丢弃的信息；输入门 i_t 决定被存放到细胞状态中的新信息；输出门 o_t 一个细胞状态输出的值，通过3种“门”结构成功解决了RNN难以预测长时序数据的问题。本文基于LSTM神经网络搭建的用户语义位置预测模型如图7所示，预测模型主要有输入层（inputs）、词向量层（embedding）、LSTM隐藏层和输出层（outputs）4部分组成。在基于群体用户的位置预测中，模型输入群体用户的语义轨迹，根据群体用户的语义轨迹训练相似用户的行为模式。

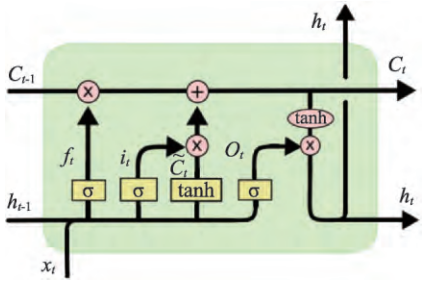


图6 LSTM单元结构

Fig. 6 LSTM cell structure

4 实验结果与分析

4.1 实验数据

实验数据来源于济南市某广场一周内的移动用户蓝牙定位数据以及该商场的商铺数据。室内蓝牙定位数据从2017年12月20日至2017年12月27日，覆盖广场5个楼层，平均采样为1−10 s不等，定位精度约为3 m，数据字段包括用户ID、记录时间、用户的位置（经纬度及所在楼层ID），如表1所示。一周用户总记录量逾300万，轨迹点总记录量为69 070 836个，经过预处理后共剩五万多条轨迹。商铺数据采用爬虫程序从百度地图爬取，共爬取了352间商铺数据，经坐标转换后与室内用户定位数据相匹配。每条商铺数据包括商铺唯一标识ID、商铺范围（坐标序列组成的面要素）、商铺名称、所在楼层ID，如表2所示，商铺经纬度范围和所在楼层ID共同确定该商铺在商场的具体位置。

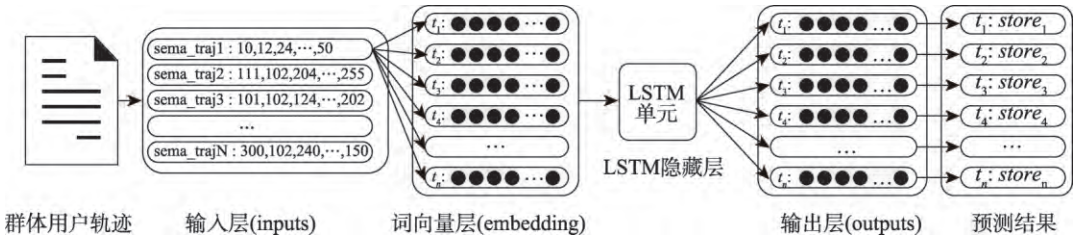


图7 室内用户语义位置预测模型

Fig. 7 Indoor user semantic location prediction model

表1 用户轨迹数据实例

Tab. 1 Samples of user's records

用户ID	时间	X/m	Y/m	所在楼层ID
0000CE***	2017−12-20 10:46:45	130219***	43904***	1
0000CE***	2017−12-20 10:46:57	130219***	43903***	1
0000CE***	2017−12-20 10:47:05	130219***	43904***	1
…	…	…	…	…
0000CE***	2017−12-20 19:20:33	130219***	43904***	4
0000CE***	2017−12-20 19:20:45	130219***	43904***	4

表2　商场商铺实例

Tab. 2　Samples of semantic stores

商铺ID	商铺形状	商铺名称	所在楼层ID
1	Shape(面)	***	2
2	Shape(面)	***	2
3	Shape(面)	***	4
…	…	…	…
351	Shape(面)	***	4
352	Shape(面)	***	3

4.2 停留区域序列识别结果与分析

停留区域识别结果依赖于算法中时间阈值 T_{threh} 和距离阈值 dis_{thred} 的选择。本文参考相邻商铺中心点之间的距离(约10 m)，将停留时间超10 min且相距大于10 m的轨迹簇视为用户的停留区域。实验结果共获得51 894条有效的用户停留区域序列(图8)。与此同时，本文采用Li等[13]和Zheng等[19]提出的启发式算法识别轨迹中的停留区域，采用同样的阈值参数(距离阈值 dis_{thred} =10，时间阈值 T_{threh} =10)，结果只获得8297条有效的用户停留区域序列，图9为分别采用启发式算法和ST-AGNES算法得到的某用户停留区域对比结果，可以看出采用ST-AGNES算法可以保留更多的用户停留信息。将本文方法得到的某用户停留区域轨迹点与商场平面图叠加显示(图8、9)，用户的停留区域基本出现在商铺内部，符合基本常识，同时验证了本文算法的可靠性，为下一步的语义位置匹配奠定了基础。

图8　某用户停留区域轨迹点与室内商铺

Fig. 8　User's stay area points and indoor store

4.3 室内语义位置匹配结果与分析

语义轨迹是语义位置预测的核心，本文将商铺名称作为用户语义位置，基于本文提出的吸引度规则将用户停留区域与商铺名称相匹配，共匹配出352间商铺信息，同时本文采用传统做法将停留区域的中心点与商铺相匹配，共匹配出308间商铺信息，可以看出传统语义匹配方法会漏掉部分商铺信息。对所有用户轨迹进行语义匹配处理后总共获得24 267条用户语义轨迹(当语义轨迹的长度小于2时，本文认为该语义轨迹价值不高，删除该语义轨迹)，表3是经编码后的用户语义轨迹，用户语义轨迹为下一步的语义位置预测提供了数据支持。

图9　启发式算法和ST-AGNES算法得到的某用户停留区域对比

Fig. 9　Comparison of a user's stay area obtained by heuristic algorithm and ST-AGNES algorith

表3　用户语义轨迹实例

Tab. 3　Samples of semantic stores

用户ID	用户语义轨迹
0000CE***	$S_{48}, S_{91}, S_{34}, S_{231}, S_{34}, S_{91}, S_{11}, S_{79}$
FA8170***	$S_{301}, S_{60}, S_{286}, S_{132}, S_{94}, S_{292}, S_{285}, S_{310}, S_{48}$
FAA378***	$S_{20}, S_{211}, S_{223}, S_{20}, S_{343}, S_{20}$
…	…
FE53FA***	$S_{107}, S_{132}, S_{107}, S_{296}, S_{107}, S_{132}, S_{124}, S_{132}$
0AEE45***	$S_{234}, S_{43}, S_{297}, S_{60}, S_{48}, S_{32}, S_{322}, S_{271}, S_{94}, S_{95}$

4.4 用户语义位置预测结果与分析

为预测用户语义位置，本文选取20 000条语义轨迹作为LSTM的训练集，剩下的所有语义轨迹作测试集，表4为经多次试验最终确定的LSTM神经网络参数。

本文以复杂度和准确率作为预测结果的评价指标：

(1)复杂度：用来评价模型预测语义位置是否很好的标准，模型的复杂度越低，代表模型的预测能力越好。复杂度可理解为平均分支系数，即模型通过已知语义位置预测下一个语义位置时的平均

表4 LSTM模型参数

Tab. 4 LSTM model parameters

BATCH_SIZE	NUM_LAYERS	HIDDEN_SIZE	EMBEDDING_SIZE	LEARNINT_ RATE
64	2	256	128	0.01

可选择数量,其中 $p(w_i]w_1, \cdots, w_{i-1})$ 代表通过前 $i-1$ 个语义位置预测正确第 i 语义位置的概率:

$$perplexity(S) = p(w_1, w_2, w_3, \cdots, w_m)^{-1/m} = \sqrt[m]{\frac{1}{p(w_1, w_2, w_3, \cdots, w_m)}} = \sqrt[m]{\prod_{i=1}^{m}\frac{1}{p(w_i]w_1, \cdots, w_{i-1})}} \quad (2)$$

(2)准确率:准确率也叫做正确率,即预测结果中正确的数量占预测结果集的比例。在本文中即用户下一地点预测正确的次数 N_{step} 与该用户语义轨迹长度 N_{traj} 的比值:

$$Accuracy = \frac{N_{step}}{N_{traj}} \quad (3)$$

模型复杂度和预测正确率的变化情况如图10所示。由图10(a)可知,模型的复杂度随着迭代次数和训练数据的增加先急剧下降后处于平稳波动中,最终稳定在7左右;由图10(b)可知,模型的预测正确率也随着迭代次数和训练数据的增加逐渐上升后稳定在61.3%左右。由此可见,LSTM神经网络随着迭代次数和训练数据的增加逐渐在群体用户的语义轨迹中发现了语义位置模式。模型对某用户预测的商铺如表5所示,将商铺信息按照访问概率的大小进行排序,括号内为某用户下一时刻访问该商铺的概率,为节约篇幅,只显示排名前三的商铺信息。从表5可看出,假设该用户去过名为YAGERRIS的商铺,那么接下来他要去商铺reemoor的概率为0.09,去KISS KITTY的概率为0.09,去AESOMINO的概率为0.08,而他实际去的商铺为FAmecoco,这是由于模型的输入轨迹过短,模型获得的已知知识过少,导致预测难度较大,但随着语义轨迹长度的增加,模型的预测正确率越来

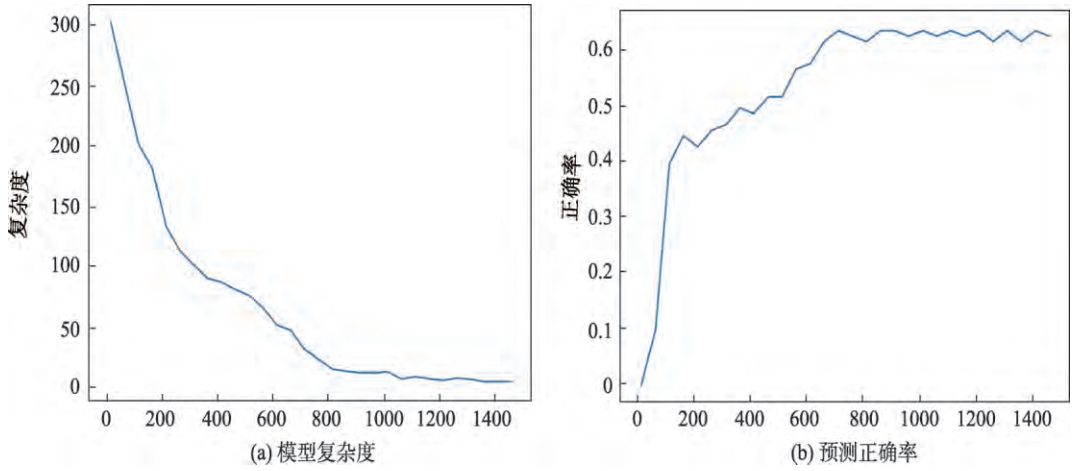


图10 模型复杂度与预测正确率

Fig. 10 Model perplexity and prediction accuracy

表5 某用户语义位置预测结果

Tab. 5 User semantic location prediction results

已知语义轨迹	预测语义位置	实际位置
YAGERRIS	reemoor(0.09),KISS KITTY(0.09),AESOMINO (0.08)	FAmecoco
YAGERRIS,FAmecoco	marfeel (0.13),reemoor (0.11),FIOCCO(0.11)	reemoor
YAGERRIS,FAmecoco,reemoor	KISS KITTY (0.26),FIOCCO (0.19),marfeel (0.08)	KISS KITTY
YAGERRIS,FAmecoco,reemoor,KISS KITTY	FIOCCO(0.35) ,marfeel(0.17),ILAPAOE(0.11)	FIOCCO

越高。同时，从表中多条语义轨迹可以看出，该用户光顾的商铺为时尚女装女鞋系列，模型推荐的也多为该系列，由此可看出模型推荐的商铺类别基本与用户的爱好一致。

5 结论

本文基于济南市某广场群体用户的蓝牙定位轨迹数据预测用户的语义位置。首先提出了ST-AGNES算法，该算法仅需要距离阈值即可自动生成簇集的个数，克服了其他时空聚类算法提前指定簇集个数、超参数过多和全局密度阈值等缺点。在语义匹配阶段，引入了吸引度规则将用户停留区域与商场的商铺信息相关联。与时空聚类算法和语义匹配算法进行比较可知，ST-AGNES算法、吸引度规则可以识别和匹配更全的停留区域和语义信息。最后，采用LSTM模型对商场群体用户语义位置的建模并预测用户语义位置。预测结果表明商场室内群体用户的语义位置存在语义联系，可帮助商场提前预知用户的行为习惯，有助于提高商场精准营销能力。

参考文献（References）：

[1] Wu F, Fu K, Wang Y, et al. A spatial-temporal-semantic neural network algorithm for location prediction on moving objects[J]. Algorithms, 2017,10(2):37.

[2] 谭娟,王胜春.基于深度学习的交通拥堵预测模型研究[J].计算机应用研究,2015,32(10):2951-2954. [Tan J, Wang S C. Research on prediction model for traffic congestion based on deep learning[J]. Application Research of Computers, 2015,32(10):2951-2954.]

[3] Ying J C, Lee W C, Weng T C, et al. Semantic trajectory mining for location prediction[C]. ACM Sigspatial International Conference on Advances in Geographic Information Systems, 2011:34-43.

[4] Sabarish B A, Karthi R, Gireeshkumar T. A survey of location prediction using trajectory mining[M]. Springer India, 2015:119-127.

[5] Liu J, Wolfson O, Yin H. Extracting semantic location from outdoor positioning systems[C]. MDM2006 workshop MCISME, 2006.

[6] Jeung H, Liu Q, Shen H T, et al. A hybrid prediction model for moving objects[C]. IEEE International Conference on Data Engineering, 2008:70-79.

[7] Ye Y, Zheng Y, Chen Y, et al. Mining individual life pattern based on location history[C]. Tenth International Conference on Mobile Data Management: Systems, Services and MIDDLEWARE, 2009:1-10.

[8] Morzy M. Mining frequent trajectories of moving objects for location prediction[C]. International Conference on Machine Learning and Data Mining in Pattern Recognition, 2007:667-680.

[9] Zheng Y, Zhang L, Ma Z, et al. Recommending friends and locations based on individual location history[J]. Acm Transactions on the Web, 2011,5(1):5.

[10] Alvares L O, Bogorny V, Kuijpers B, et al. A model for enriching trajectories with semantic geographical information[C]. ACM International Symposium on Advances in Geographic Information Systems, 2007:22.

[11] 窦丽莎,曹凯.出行者子停留语义推断模型框架[J].山东理工大学学报(自然科学版),2012,26(6):17-22. [Dou L S, Cao K. A model framework for inferring sub-stays semantics of traveler[J]. Journal of Shandong University of Technology (Natural Science Edition), 2012,26(6):17-22.]

[12] 齐凌艳,陈荣国,温馨.基于语义轨迹停留点的位置服务匹配与应用研究[J].地球信息科学学报,2014,16(5):720-726. [Qi L Y, Chen R G, Wen X. Research on the LBS matching based on stay point of the semantic trajectory [J]. Journal of Geo-information Science, 2014,16(5):720-726.]

[13] Li Q, Zheng Y, Xie X, et al. Mining user similarity based on location history[C]. ACM Sigspatial International Conference on Advances in Geographic Information Systems, 2008:34.

[14] 宋路杰,孟凡荣,袁冠.基于Markov模型与轨迹相似度的移动对象位置预测算法[J].计算机应用,2016,36(1):39-43. [Song L J, Meng F R, Yuan G. Moving object location prediction algorithm based on markov model and trajectory similarity[J]. Journal of Computer Applications, 2016,36(1):39-43.]

[15] 彭曲,丁治明,郭黎敏.基于马尔可夫链的轨迹预测[J].计算机科学,2010,37(8):189-193. [Peng Q, Ding Z M, Guo L M. Prediction of trajectory based on markov chains[J]. Computer Science, 2010,37(8):189-193.]

[16] 林树宽,李昇智,乔建忠,等.基于用户移动行为相似性聚类的Markov位置预测[J].东北大学学报(自然科学版),2016,37(3):323-326. [Lin S K, Li S Z, Qiao J Z, et al. Markov location prediction based on user mobile behavior similarity clustering[J]. Journal of Northeastern University (Natural Science), 2016,37(3):323-326.]

[17] 张心悦,王光霞,吴月,等.室内用户语义位置模式挖掘研究——以商场为例[J].测绘与空间地理信息,2016,39(2):12-16. [Zhang X Y, Wang G X, Wu Y, et al. Research on semantic location pattern mining of indoor users: Take shopping malls as an example[J]. Geomatics & Spatial

Information Technology, 2016,39(2):12-16.]

[18] Ester M, Kriegel H P, Xu X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]. International Conference on Knowledge Discovery and Data Mining, 1996:226-231.

[19] Zheng Y, Zhang L, Xie X, et al. Mining interesting locations and travel sequences from GPS trajectories[C]. International Conference on World Wide Web, 2009:791-800.

[20] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data[J]. Data & Knowledge Engineering, 2007,60(1):208-221.

[21] Leiva L A. Vidal E. Warped K-Means: An algorithm to cluster sequentially-distributed data[M]. Elsevier Science Inc., 2013:196-210.

[22] 唐建波,邓敏,刘启亮.时空事件聚类分析方法研究[J].地理信息世界,2013(1):38-45. [Tang J B, Deng M, Liu Q L. On spatio-temporal events clustering methods[J]. Geomatics World, 2013(1):38-45.]

[23] 马春来,单洪,李志,等.移动用户下一地点预测新方法[J].浙江大学学报(工学版),2016,50(12):2371-2379. [Ma C L, Shan H, Li Z, et al. New next place prediction method for mobile users[J]. Journal of ZheJiang University (Engineering Science), 2016,50(12):2371-2379.]