带有时间标签的流行社交位置发现

刘长赟 杨宇迪 周丽华 赵丽红
(云南大学信息学院 昆明 650091)

摘 要 流行社交位置是指大多数人日常生活中经常访问的位置,其广泛应用于推荐系统、定向广告应用等领域。随着基于位置的社交网络(Location-Based Social Network,LBSN)的迅速发展,流行社交位置的挖掘成为时空数据挖掘中的一个研究热点。然而,现有的研究主要是从LBSN中挖掘流行社交位置,忽略了流行社交位置的时间因素,因此,文中提出了带有时间标签的流行社交位置发现算法。该算法首先量化LBSN数据集中的时间信息,得到个体用户带有时间标签的频繁社交位置集合;然后计算这些带时间标签的位置在群体用户中的流行度;最后识别出符合要求的带时间标签的流行社交位置。文中采用约10个月的Foursquare东京用户签到数据对该算法的效率和正确性进行验证,结果表明,该算法能够较为准确地发现带有时间标签的流行社交位置。

关键词 时空数据挖掘,基于位置的社交网络,流行社交位置,带有时间标签的流行社交位置

中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.029

Discovering Popular Social Location with Time Label

LIU Chang-yun YANG Yu-di ZHOU Li-hua ZHAO Li-hong
(School of Information Science and Engineering, Yunnan University, Kunming 650091, China)

Abstract The popular social location means the places that most people visit frequently in daily life, which is widely used in recommendation systems, targeted advertisement applications, and other fields. With the rapid development of location-based social networks (LBSN), the identification of popular social locations has become an important hot research point in spatio-temporal data mining. However, the existing research mainly focuses on mining popular social locations from LBSN, but ignores the time factor of popular social locations. Therefore, this paper proposed a new algorithm for mining popular locations with time label. The proposed algorithm first quantifies the time information in the LBSN dataset to obtain a set of frequent social locations with respect to individual users, then calculates the popularity of these locations with respect to group of users, and then identifies popular social locations that meet the requirements. This paper validated the efficiency and correctness of the algorithm by using the Foursquare Tokyo user check-in data for about 10 months. The results show that the proposed algorithm can find the popular social location with time label more accurately.

Keywords Spatio-temporal data mining, Location-based social network, Popular social location, Popular social location with time label

1 引言

近年来,随着无线通信网络、全球定位系统等设施的飞速发展,社交网络已经成为人们日常生活中不可或缺的一部分。人们喜欢通过社交网络分享他们的位置信息,并通过手机、电脑等设备追踪他人的位置信息。这些分享引领了基于位置的社交网络数据分析。

利用LBSN中的位置信息和时间信息,可以发现人们日常生活和习惯中一些有价值的潜在信息,例如:哪些地方是人们经常访问的;哪些地方是人们经常在某个时间访问的;人们喜欢在哪个时间访问哪个地方;人们从来没有去过的地方。因此,本文的主要目的是发现带有时间标签的流行社交位置,即用户频繁访问的社交位置以及被频繁访问的时间点。

发现带有时间标签的流行社交位置不仅可以揭示用户在时间和空间上的偏好,还可以广泛应用于推荐系统、决策系统、市场营销、城市规划等方面。但是,发现带有时间标签的

到稿日期:2018-08-01 返修日期:2018-10-10 本文受国家自然科学基金(61762090,61262069,61472346,61662086),云南省自然科学基金(2016FA026,2015FB114),云南省创新研究团队项目(2018HC019),云南省高等学校科技创新团队项目(IRTSTYN)资助。

刘长赟(1992-),男,硕士,主要研究方向为数据挖掘;**杨宇迪**(1995-),女,硕士生,主要研究方向为数据挖掘;**周丽华**(1968-),女,博士,教授,CCF会员,主要研究方向为数据挖掘、社会网络分析,E-mail:lhzhou@ynu.edu.cn(通信作者);**赵丽红**(1974-),女,硕士,主要研究方向为数据挖掘。

流行社交位置仍然面临很多挑战。1）数据的稀疏性。由于用户不断地进行位置签到以及信息的快速扩散，基于位置的社会化网络积累了大量的签到位置轨迹数据和时间数据，但是单独观察每个用户可以发现，其留下的位置数据和时间数据仍十分稀疏。2）时间维度分析难度高。面对海量、高噪声和非线性的时间信息数据，现有的研究将连续的时间数据转化为离散的时间区块数据[1-3]，虽然这样便于计算，但造成了信息的丢失。3）人类行为复杂。人类行为在时间维度上具有高度非均匀性，并服从幂率分布[4]，例如：在多个时间段内没有活动记录，但这些时间空挡中却穿插着阵发的密集活动[5]。因此需要更加新颖且有效的方法来发现带有时间标签的流行社交位置。

目前，许多研究人员研究了流行社交位置的挖掘问题，并提出了几种解决方法。但是，这些研究中有些专注于位置的发掘，而未考虑时间因素；有些仅将时间信息作为一种度量单位，没有考虑详细时间信息对流行社交位置的作用；有些虽然考虑了时间，但仅将时间信息用于个人用户的位置发现，没有考虑群体用户中时间信息对位置挖掘的影响。

因此，本文首先量化 LBSN 数据集中的时间信息，得到个体用户带有时间标签的频繁社交位置集合，然后计算这些带时间标签的位置在群体用户中的流行度，并识别符合要求的带时间标签的流行社交位置。本文主要贡献包括：

（1）提出了位置访问率（Location Visit Rate，LVR）来量化 LBSN 数据集中的时间信息，将复杂的人类行动转化为大量可计算的数据。

（2）扩展用户群位置的流行度到用户群〈位置-时间〉对的流行度，考虑时间信息对用户群位置挖掘的影响。

（3）提出了带有时间标签的流行社交位置发现（Popular Social Location Mining with Time Label，SPLMTL）算法。

（4）使用真实数据集进行实验，对比了各个参数对算法性能的影响，并对算法的准确性进行了验证。

本文第 2 节介绍相关工作；第 3 节给出带有时间标签的流行社交位置发现问题的定义，并介绍 Dokuz 等[6] 提出的 SS-ILM（SocioSpatially Important Locations Mining）算法；第 4 节介绍相关概念及 SPLMTL 算法；第 5 节阐述实验结果；最后总结全文。

2　相关工作

2.1　位置挖掘相关研究

基于 LBSN 的推荐包括基于内容的推荐、基于链路分析的推荐和基于协同过滤的推荐。基于内容的推荐，通过分析数据来判断用户喜好，从而进行推荐[7-8]。Maroulis 等[9] 利用与签到数据相关的基于分类的内容，提出了一种使用张量因子分解的内容感知兴趣点推荐系统。基于链路分析的推荐，通过利用社会网络的拓扑结构以及用户的历史信息，来识别有经验的用户和令人感兴趣的位置[10]。Yao 等[11] 提出了一种基于非负张量因子分解的协同过滤方法，获得了一个适合于场景感知兴趣点（Point-of-Interest，POI）推荐的紧凑的数据模型。陈功等[12] 提出了一种改进后的基于邻居节点的计算方法 INBIC 来计算影响力上界，并进行位置推荐。基于协同过滤（CF）的推荐，文献[13-15]从用户的历史数据出发，结合其他相似度较高的用户的经验来推断出用户的喜好。Dao 等[16] 提出了一种内容感知的基于协同过滤的遗传算法（CACF-GA），该算法通过定义离散的场景，将场景的相似性概念应用到传统的协同过滤技术上，以创建一个内容感知的推荐系统模型，然后使用遗传算法对该系统行优化，最终构建了一个基于位置的广告推荐系统。

Chen 等[17] 从签到数据中挖掘移动演化模式，提出了一个距离函数，并利用位置敏感哈希法来优化该函数并使其能更加快速地发现移动演化模式。Celikte 等[18] 通过分析 LBSN 网络中的签到数据，提取 LBSN 用户和各个区域的特征，从而得到一个将城市各个区域标签化的概率模型，用以发现不同城市相同标签的区域。Zhao 等[19] 通过对市民活动进行建模，提出了人口流动的预测因子，并利用深度学习预测器预测城市人口的流动性。Dokuz 等[20] 提出了社会时空重要位置，他们使用一些阈值来筛选数据，最终根据筛选出的数据分析社会时空重要位置。Zhang 等[21] 使用社会与地理融合模型（SGFM）构建了一种 POI 推荐方法。Ye 等[22] 将 LSBN 中的社会影响力和地理影响相结合，通过幂律分布对其进行建模，建立了基于地理影响的协同推荐算法。Levandoski 等[23] 使用非空间项目的空间评级、空间项目的非空间评级和空间项目的空间评级 3 种基于位置的评级分类方法组建了一个基于位置感知的推荐系统（LARS）。Bao 等[24] 提出了一个基于位置和偏好的推荐系统，该系统由加权类别层次结构离线建模和在线推荐两部分组成。Cui 等[25] 使用用户社交网络和超图模型共同构建了高阶关联信息，并将其与用户偏好一起组成矩阵，最后使用矩阵分解法来集成不同的偏好与关系信息，得到用户的偏好排名以用于 POI 推荐。Cho 等[26] 开发了一种人类活动模型，将周期性的短程运动与基于社交网络结构的旅行相结合，从而预测用户未来可能到达的位置。

上述文献对社交位置的挖掘做了很好的阐述，其中包括基于位置的个人推荐、城市模式发现及重要社交位置的挖掘，但这些方法仅仅将时间信息作为一种度量单位，并没有深入发挥详细的时间信息在位置挖掘中的作用和影响。

2.2　时间相关研究

位置挖掘中，时间是一个不可忽略的因素，因此许多研究考虑了时间因素。Yuan 等[1] 通过计算空间和时间的相似性，提出了一种基于协同过滤方法计算用户之间的空间相似性和时间相似性的推荐系统。Yuan 等[2] 于 2014 年提出了一种地理-时间感知图（GTAG），并提出了宽度优先传播（BPP）的偏好传播算法，该算法遵循一种放松呼吸优先搜索策略，在最多 6 个传播步骤内返回推荐结果并提供给用户一个给定时间访问的 POIs 列表。Gao 等[3] 将时间划分成各个时段，并在每个时段下学习用户签到的习惯，聚合时间签到得到位置推荐。Zhang 等[27] 建立了一种受连续时间影响的概率模型，并讨论了连续时间影响下用户和位置的相关性。Baltrunas 等[28] 将

用户数据划分为很多微数据，根据生活经验将这些微数据按照时间段分割，然后根据用户在各个时段的操作、行为等隐式反馈来发现兴趣，最后使用协同过滤的方法得到针对该用户的推荐。Si 等[29]将时间块特征与基于用户的协同过滤和空间影响相结合，提出了一种 POI 推荐方法(UPT)。Gao 等[30]对 LBSN 的时间效应进行了研究，并提出了一个通用的框架来开发和建模时间循环模式及其空间和社会数据的关系。Zhang 等[31]提出了一种位置推荐重力模型，通过个性化引力参数衡量每个已有位置对新位置的影响，提高了推荐精度。Ye 等[32]在 POI 推荐中使用语义标签，并将时间作为一个观测特征，在各语义标签下观察每个循环时间段的签到分布差异。Li 等[33]提出了一个四阶张量的排序方法，考虑用户的时变行为趋势，同时捕捉用户的长期偏好和短期偏好，从而可以产生符合用户兴趣的推荐。他们还建议对地点进行分类，以减轻数据稀疏和冷启动问题。Ying 等[34]使用用于用户偏好建模的上下文感知张量分解和基于 POI 评级的加权点击(超文本诱导的主题搜索)两个组件组成了一个 POI 推荐系统。

上述文献将时间因素和位置相因素结合应用于推荐系统、POI 发现等，提升了推荐系统的性能和精度。但是这些研究仅将时间信息用于个人用户的推荐及位置发现，没有考虑在群体用户中时间信息对位置挖掘的影响。

3 预备知识

本节首先给出带有时间标签的流行社交位置发现问题的定义，然后介绍由 Dokuz 等[6]提出的 SS-ILM(Socio Spatially Important Locations Mining)算法的相关定义。为了便于阐述，表 1 给出了本文的常用符号及含义。

表 1 本文中的关键符号及含义

Table 1 Overview of key symbols

符号	含义
D	LBSN 数据集
L	位置集合
T	时间集合
t	l 被访问的时间
UP	用户流行度
LD	位置密度
VLT	访问生命周期
$SILU$	用户的社交重要位置
$\langle l-t\rangle$	〈用户-时间〉对
$\langle L-T\rangle$	〈位置-时间〉对集合
$min\_density$	位置密度阈值
min_visit	访问生命周期阈值
min_UP	用户流行度阈值
min_LVR	位置访问率阈值
$FSLUTL$	带时间标签的用户频繁社交位置
LVR	位置访问率

3.1 问题定义

本文将带有时间标签的社交位置表示为〈位置-时间〉对，简化为〈l-t〉对。带时间标签的流行社交位置发现问题定义为：在 LBSN 数据集 D 中，给定用户重要社交位置 $SILU$、位置访问率阈值 min_LVR、用户流行度阈值 min_UP，能够有效地发现带有时间标签的流行社交位置，即满足 $SILU$ 和用户流行度阈值的〈l-t〉对。

3.2 SS-ILM 算法[6]

SS-ILM 算法使用数据集 D，将用户的多维签到数据转换为位置密度和访问生命周期，并使用位置密度阈值和访问生命周期阈值从用户的位置集中筛选出符合要求的个人重要社交位置，然后将所有用户的个人重要社交位置整合到一个位置集中，轮流考查重要社交位置对所有用户的流行度，最后通过用户流行度阈值筛选出整个用户群的流行社交位置。SS-ILM 算法的相关定义如下。

定义 1(位置密度[6]) 对于用户 u，位置 l 的位置密度被定义为用户 u 访问位置 l 的次数在 u 所有访问次数中的占比，即

$$LD_l^u=\frac{C_{u\text{ in }l}}{C_{u\text{ in }L}} \tag{1}$$

其中，$C_{u\text{ in }l}$表示用户 u 访问位置 l 的次数，$C_{u\text{ in }L}$表示用户 u 访问 L 中位置的总次数。

定义 2(频繁位置[6]) 给定位置 l 关于用户 u 的位置密度值 LD_l^u和位置密度阈值 $min\_density$，若 $LD_l^u\geqslant min\_density$，则 l 被称为用户 u 的频繁位置。

定义 3(访问生命周期[6]) 用户 u 对位置 l 的访问生命周期 VLT_l^u定义为用户 u 对位置 l 的第一次访问和最后一次访问的时间差在用户 u 出现在社交媒体数据中的总时长中的占比。

$$VLT_l^u=\frac{Last\_Visit_{u\text{ in }l}-First\_Visit_{u\text{ in }l}}{Last\_Time_u-First\_Time_u} \tag{2}$$

其中，$Last_Visit_{u\text{ in }l}$表示用户 u 最后一次访问位置 l 的时间，$First_Visit_{u\text{ in }l}$表示用户 u 第一次访问位置 l 的时间；$Last\_Time_u$ 表示用户 u 最后一次出现在社交媒体中的时间，$First\_Time_u$ 表示用户 u 第一次出现在社交媒体中的时间。

定义 4(时间高比率位置[6]) 给定用户 u 对位置 l 的访问生命周期和访问生命周期阈值 min_visit，若 $VLT_l^u\geqslant min\_visit$，则称 l 为时间高比率位置。

定义 5(重要社交位置[6]) 给定社交媒体用户 u 访问的位置 l，若位置 l 既是频繁位置又是时间高比率位置，则位置 l 被称为 u 的重要社交位置(Social Important Location for User，SILU)。

4 带有时间标签的流行社交位置的发现

虽然 SS-ILM 算法能够在 LBSN 网络中准确地挖掘出流行社交位置，但是该算法没有考虑流行社交位置的时间效应。因此本文在此基础上提出带有时间标签的流行社交位置的发现。

本节首先介绍带有时间标签的流行社交位置发现的基本概念，然后介绍所提出的 SPLMTL 算法。

4.1 基本概念

定义 6(位置访问率 LVR(Location Visit Rate)) 给定用户 u、$SILU$ 中的位置 l 和观测时刻 t，则用户 u 在 t 时刻对位置 l 的访问率为位置 l 在 t 时刻被用户 u 访问的次数与观测时长的比值，记为 $LVR_u^{l\text{ at }t}$。

$$LVR_u^{l\text{ at }t}=\frac{C_{u\text{ in }l}(t)}{\Delta T} \tag{3}$$

其中，$C_{u\text{ in }l}(t)$表示位置 l 在 t 时刻被用户 u 访问的次数；ΔT 表示观测时长，即观测开始时间与观测结束的时间差。

$LVR_u^{l\text{ at }t}$说明了用户在 t 时刻对位置 l 的访问率。例如，用户在每天上午 9 点访问 a 地点，在一周之内(7 天)，用户在上午 9 点对 a 地访问了 5 次，则 $LVR_u^{l\text{ at }t}=5/7$。$LVR_u^{l\text{ at }t}$意味着用户在 t 时出现在位置 l 的概率。访问率 $LVR_u^{l\text{ at }t}$越高，则用户 u 在 t 时刻访问位置 l 的频繁度越高，反之频繁度越低。

定义 7(频繁时间)　给定一个位置访问率阈值 min_LVR，如果用户 u 在 t 时刻对位置 l 的访问率 LVR 大于或等于 min_LVR 的阈值，即 $LVR_u^{l\text{ at }t}\geqslant min\_LVR$，则该 t 时刻被称为用户 u 访问位置 l 的频繁时间。

定义 8(带时间标签的用户频繁社交位置(Frequent Socially Location for User with Time Label，FSLUTL))　给定用户 u 和〈位置-时间〉对集合〈L-T〉，若〈l-t〉∈〈L-T〉且位置 l 满足定义 5，时间 t 满足定义 7，则称该〈l-t〉为带时间标签的用户 u 的频繁社交位置。

定义 9(用户流行度(User popularity，UP))　给定用户 u、带时间标签的位置〈l-t〉对，那么满足定义 8 的 FSLUTL 的用户数量与用户总数的比值即为带时间标签的位置〈l-t〉在所有用户中的用户流行度，记为 $UP_{\langle l\text{-}t\rangle}$。

$$UP_{\langle l\text{-}t\rangle}=\frac{N_{\langle l\text{-}t\rangle\text{ in FSLUTL}}(u)}{N(u)} \tag{4}$$

其中，$N_{\langle l\text{-}t\rangle\text{ in FSLUTL}}(u)$表示满足定义 8 的 FSLUTL 的用户数量，$N(u)$表示社交媒体用户的总数量。

定义 10(带时间标签的流行社交位置(Socially Popular Location with Time Label，SPLTL))　给定一个用户流行度阈值 min_UP、带时间标签的位置〈l-t〉的用户流行度 $UP_{\langle l\text{-}t\rangle}$，如果 $UP_{\langle l\text{-}t\rangle}\geqslant min\_UP$，则〈$l$-$t$〉被称为带时间标签的流行社交位置。

4.2　带有时间标签的流行社交位置挖掘算法

为了从包含用户集 U、用户访问位置集 L 和位置访问时间集 T 的 LBSN 数据集 D 中挖掘出满足条件的〈位置-时间〉对，本文提出了 SPLMTL 算法。该算法将重要社交位置 $SILU$ 与用户的位置访问时间信息进行整合，并使用位置访问率阈值筛选出个人用户带时间标签的流行社交位置，然后将所有个人用户的带时间标签的频繁社交位置整合为一个〈L-T〉集合，从〈L-T〉集合中找到不小于用户流行度阈值 min_UP 的〈l-t〉对。

算法 1　SPLMTL 算法

输入：社交网络数据集 D，包含用户信息 U(如用户 ID、用户名等)、位置信息 L(如经纬度、地名等)和时间信息 T；位置访问阈值 min_LVR，用户流行度阈值 min_UP

输出：带有时间标签的〈位置-时间〉对集合 SPLTL

```
1. 初始化所有〈位置-时间〉对集合〈L-T〉，FSLUTL=Ø，SPLTL=Ø
2. FOR each D 中的社交媒体用户 u
3.    使用 SS-ILM 算法计算得到每个 u 的重要社交位置集合 SILU
4.    FOR each SILU 中的位置 l
5.       转换位置 l 为〈l-t〉
6.       计算每个〈l-t〉的 LVR
7.       IF LVR≥min_LVR：
8.          将〈l-t〉对加入到 FSLUTL 中
9.       END IF
10.   END FOR
11.   将 FSLUTL 中元素加入〈L-T〉
12. END FOR
13. 计算〈L-T〉中所有元素的用户流行度 UP
14. 使用 min_UP 从〈L-T〉中选取符合条件的元素加入 SPLTL
15. 输出 SPLTL
```

SPLMTL 算法的目的是发现带有时间标签的流行社交位置。在 SPLMTL 算法中，第 1 步初始化〈位置-时间〉对集合〈L-T〉，〈L-T〉将用于保存所有用户的 FSLUTL，第 2－11 步描述如何挖掘 SPLTL。算法中第 5 步得到所有 SILU 中的位置后对其中每个位置的访问时间进行考量，得到一组用户带有时间信息的重要社交位置，即〈l-t〉对。第 6 步计算上述〈l-t〉的位置访问率 LVR，第 7－8 步将大于或等于 min_LVR 的〈l-t〉对加入 FSLUTL 中。第 11 步计算〈L-T〉集合内元素的用户流行度，第 14 步使用用户流行度阈值 min_UP 进行筛选，最终得到 SPLTL。

4.3　SPLMTL 算法的时间复杂度分析

SPLMTL 算法在计算用户的 SILU 时需要计算每个用户每个位置的位置密度 LD 和访问生命周期 VLT，该步骤的时间复杂度为 $O(n^2)$。计算每个用户所有〈位置-时间〉对的位置访问率的复杂度同样为 $O(n^2)$。而计算用户流行度时，仅对所有〈l-t〉对进行对比，则该步骤算法的复杂度 $O(n)$。由此可得，SPLMTL 算法的时间复杂度为 $O(n^2)$，其中 n 为 LBSN 数据集的用户总数。

5　实验设计

本节将对实验效果以及相关参数对算法性能的影响进行说明，实验使用 Inter Core I7 CPU 2.60 GHz 的 CPU，Kingston 8Gx2 DDR3 1 600 MHz 的存储器，SAMSUNG 850 EVO 120 G 的 SSD，SEAGEAT 2 TB 10 000r/m SATA Rev2.5 的 HDD，Windows 10 专业版的软件平台，JAVA jdk1.8 的开发平台。

5.1　数据集

本文实验所使用数据来自 Foursquare 用户的真实数据集。该数据集由 Yang 等[35]从 2012 年 4 月 3 日至 2013 年 1 月 16 日采集的约 10 个月的 Foursquare 用户签到数据组成，采集地点为美国纽约和日本东京。本文实验使用的是日本东京数据集，总量为 573 703 条，共包含 2 293 个用户，13 493 个位置。数据结构中包含经纬度信息，且全部指向日本东京及其周边。

由于不同的时间段可能具有不同的带有时间标签的流行社交位置，因此在保持每个覆盖位置的平均访问量相同的情况下，大致按照四季将数据集划分为以下 4 个子集：

数据集 1：2012 年 4 月 3 日－2012 年 5 月 13 日，用户数

为 1977，数据量为 136492 条，覆盖位置为 8300 个。

数据集 2：2012 年 5 月 13 日－2012 年 7 月 12 日，用户数为 2104，数据量为 135266 条，覆盖位置为 8016 个。

数据集 3：2012 年 7 月 12 日－2012 年 11 月 20 日，用户数为 2204，数据量 140487 条，覆盖位置为 8448 个。

数据集 4：2012 年 11 月 20 日－2013 年 1 月 16 日，用户数为 2109，数据量为 161458 条，覆盖位置为 8692 个。

5.2 数据预处理

5.2.1 数据清洗

在实验中，由于本数据集中含有大量噪声以及不需要的字段，例如显示为乱码的位置 ID 以及位置类型 ID 等，因此需要进行数据清洗。首先将混杂在一起的用户进行分割，通过识别每条数据用户 ID，将数据分割成由用户 ID 作为主键的用户数据集。同时在分割的过程中，删除不需要的字段，使数据集占用的存储空间尽可能小。然后对时间项进行整理，将该数据集中采用的世界协调时间 *UTC* 转换为本地时间。

5.2.2 用户筛选

由于社交用户的多样性，无视用户类型会导致计算结果出现偏差，例如：过多的旅游用户会将流行社交位置导向各个观光景点，然而对于在该城市长期生活的人来说，观光景点很少作为其社交场所。因此，用户筛选是非常必要的工作。

本文计划挖掘的流行社交位置主要是城市居民大量访问的位置。该位置具有更多本地化特征。例如长期在城市生活的居民，因为其本身对城市的熟悉程度，会在特定的时候到达某些本地人才知道的位置。于是从用户群中筛选出本地人，对发现流行社交位置具有重大的影响。如果用户有至少 50 条 Foursquare 签到记录在以东京为中心的半径 50km 以内的区域，则该用户被认定为东京本地人。

5.2.3 减少数据过度重复

在使用 Foursquare 的过程中，总是会出现一些问题，例如一个用户在极短时间内连续发送几次推特、朋友圈是一件很常见的事[36]。但是这样会对数据的有效性造成影响，本文把这类问题称为数据过度重复。为了避免数据的过度重复，需要对比每条数据中的时间信息，如果发现用户在短时间内发送过量信息，则抛弃部分重复信息，只收集有效、可用的信息。本文设定一个时间方位，若前后两条信息之间的时差不超过 20 min，则抛弃后一条，只记录前一条信息。若超过 20 min，即使记录在相同或者相近位置，也将其作为有效信息予以采用。

5.3 实验结果

本节首先对 SPLMTL 算法的正确性进行验证，然后介绍 SPLMTL 算法的性能。

5.3.1 SPLMTL 算法的正确性验证

本文从两方面讨论居住在东京的社交媒体用户带有时间标签的流行社交位置的发现的正确性。第一个方面是算法挖掘得到的位置是否是流行社交位置，其中流行社交位置的描述通过人工提取；第二个方面是算法所发现的时间标签是否符合日常生活规律。SPLMTL 算法中参数 *min_density*，*min_visit*，*min_LVR* 和 *min_UP* 分别设置为 0.06，0.05，0.02 和 0.01。

(1)流行社交位置的正确性。图 1 和表 2 显示了 Top 10 带时间标签的流行社交位置。表 2 中列出了 Top 10 流行社交位置，第一个位置秋叶原是全球最大的电器商业街，同时也是日本偶像文化、动漫文化中心，其属于东京居民的流行社交位置。位置 4，6，7 属于秋叶原周边地区，亦属于流行社交位置；位置 2 新宿是东京最繁华的商业街；位置 3 涩谷是各种时髦及流行的发源地；位置 5 池袋是东京市区中的主要商业及娱乐地区；位置 8 东京桥是百货商店聚集地；位置 9 新桥是交通枢纽；位置 10 品川包含了微软日本总部在内的商务办公中心。根据以上结果可知，本文提出的 SPLMTL 算法可以准确地发现流行社交位置。

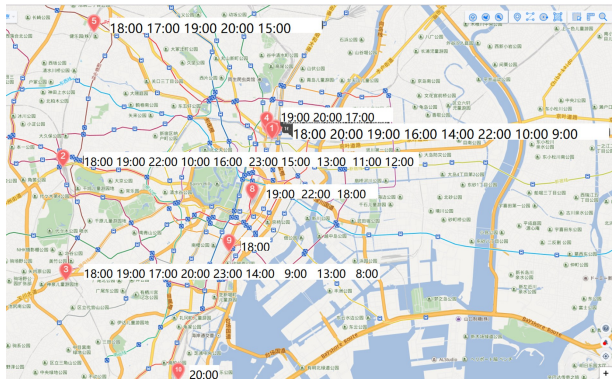

图 1 1977 个用户在东京的带有时间标签的流行社交位置

Fig. 1 Socially popular location with time label of 1977 users in Tokyo

表 2 SPLMTL 算法发现的 Top10 流行社交位置

Table 2 Socially popular location of Top-10 discovered by SPLMTL algorithm

序号	流行社交位置
1	秋叶原
2	新宿
3	涩谷
5	池袋
8	东京桥
9	新桥
10	品川
4，6，7	秋叶原周边

(2)流行社交位置时间标签的正确性。以位置 1，4，6，7，(秋叶原及其周边位置)为例，图 2 给出秋叶原周边位置的详细情况。以位置 1 为例，图 3 表示该位置的用户流行度与时间的关系，0 点到 6 点位置的流行度为 0；8 点开始，流行度开始上升；10 点到 12 点流行度缓慢下降；13 点出现小高峰；15 点略有下降；16 点至 18 点流行度再次上升，到达峰值；21 点到 23 点，流行度迅速下降至最低。通过调查东京居民的日常生活规律可知，其上班时间通常是 9 点至 17 点，12 点至 14 点为午休时间。图 3 位置 1 的流行度随时间的变化规律与居民日常生活规律基本吻合。同样观察其他位置，其时间标签与用户流行度的关系均符合当地居民的日常生活习惯。因此算法 SPLMTL 可以准确地挖掘出流行社交位置的时间标签。

图 2　东京秋叶原周边地段详情

Fig. 2　Details of surrounding area of Akihabara, Tokyo

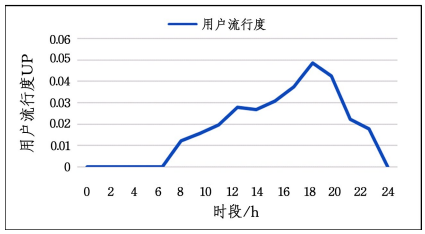

图 3　秋叶原周边位置流行度-时间图

Fig. 3　Akihabara location popularity-time chart

综上所述，SPLMTL 算法能够准确地挖掘出带有时间标签的流行社交位置。

5.3.2　算法挖掘性能

图 4—图 7 分别为 SPLMTL 算法在 4 个数据集上的挖掘结果。其中各项参数设定如下：SS-ILM 算法中计算 SILU 所用参数 *min_density* 为 0.06，*min_visit* 为 0.05；SPLMTL 算法中 *min_LVR* 为 0.02，*min_UP* 为 0.01。

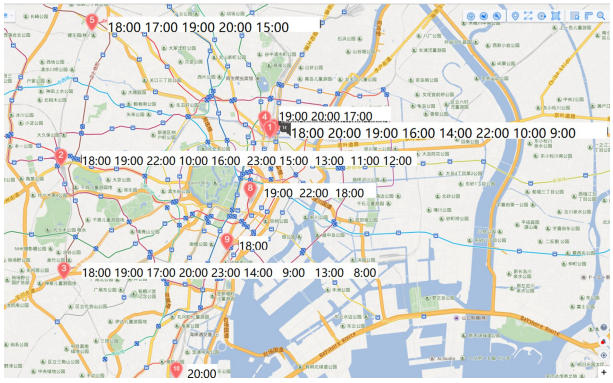

图 4　1977 个用户在东京的带有时间标签的流行社交位置

Fig. 4　Socially popular location with time label of 1977 users in Tokyo

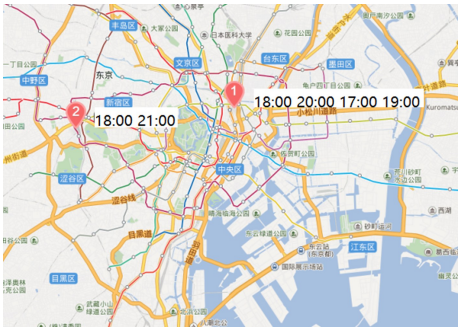

图 5　2104 个用户在东京的带有时间标签的流行社交位置

Fig. 5　Socially popular location with time label of 2104 users in Tokyo

观察图 4—图 7 可以得到，SPLMTL 在不同的数据集上显示出不同的挖掘结果，即不同时段 SPLMTL 算法的挖掘结果不同。其中，图 4 的挖掘结果最丰富，图 5 和图 7 均有挖掘结果且类似，但图 6 中没有任何结果。

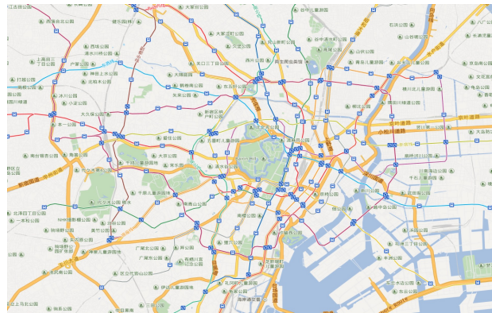

图 6　2204 个用户在东京的带有时间标签的流行社交位置

Fig. 6　Socially popular location with time label of 2204 users in Tokyo

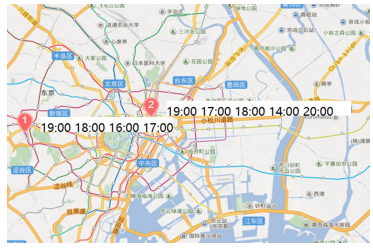

图 7　2109 个用户在东京的带有时间标签的流行社交位置

Fig. 7　Socially popular location with time label of 2109 users in Tokyo

因此，针对图 6 出现的情况，分析了每个数据集中用户日均记录条数和参数 *min_LVR* 对挖掘结果产生的影响。

(1)用户日均数据条数。数据集 1 的单个用户记录量为 1.53 条/日，数据集 2 的单个用户记录为 1.13 条/日，数据集 3 的单个用户记录量为 0.5 条/日，数据集 4 的单个用户记录量为 1.37 条/日。可以看到，在参数不变的情况下，当用户日均记录条数为 0.5 时，算法挖掘性能最低。

(2)参数 *min_LVR*。图 8、图 9 分别给出其他参数不变的情况下，参数 *min_LVR* 分别为 0.011，0.007 时，SPLMTL 算法在数据集 3 上的挖掘结果。可以看出，在用户日均记录不变的情况下，随着 *min_LVR* 的减小，图 9 相对于图 8 挖掘得到的带时间标签的流行社交位置更多。

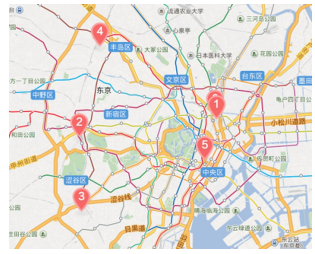

图 8　当 *min_density* 为 0.01，*min_visit* 为 0.05，*min_LVR* 为 0.011，*min_UP* 为 0.01 时，SPLMTL 算法在数据集 3 上的挖掘结果

Fig. 8　The result of SPLMTL algorithm mining on dataset 3, when *min_density* is 0.01, *min_visit* is 0.05, *min_LVR* is 0.02, and *min_UP* is 0.01

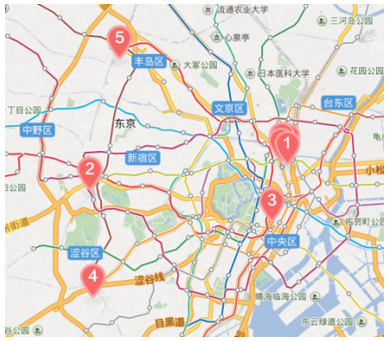

图 9 当 *min_density* 为 0.01，*min_visit* 为 0.05，*min_LVR* 为 0.007，*min_UP* 为 0.01 时，SPLMTL 算法在数据集 3 上的挖掘结果

Fig. 9 The result of SPLMTL algorithm mining on dataset 3, when *min_density* is 0.01, *min_visit* is 0.05, *min_LVR* is 0.02, and *min_UP* is 0.01

综上所述，参数 *min_LVR* 和用户日均记录条数都会影响算法挖掘结果。同时，在参数不变的情况下，SPLMTL 算法对数据集稀疏性非常敏感，用户日均记录量越高，算法的效果越好，反之则越差。

5.3.3 实验对比

本文主要研究 SPLTL 挖掘问题。正如相关工作中讨论的，Dokuz 等[6]提出的 SS-ILM 研究了如何挖掘出用户群中的重要社交位置。因此，在实验中，将 SS-ILM 算法结果与本文提出的算法挖掘结果进行了比较。

表 3 列出了 SS-ILM 算法挖掘出的 Top 10 重要社交位置，SPLMTL 算法的挖掘结果已在表 2 中列出。通过比较可以看出，4 个位置的顺序是相同的，即 1-秋叶原，2-新宿，3-涩谷，6-秋叶原周边。其他位置在两个算法结果中的位置顺序发生改变。特别地，算法 SPLMTL 不仅得到了 SS-ILM 算法中的重要社交位置，还发现了新的位置，如 9-新桥。由于 SPLMTL 算法不仅考虑用户频繁访问的位置，还考虑每个位置在不同时间点的流行度，而 SS-ILM 仅通过用户的普遍性来定位位置，因此使得 SS-ILM 算法和 SPLMTL 算法的位置顺序不同，这也是出现新的流行社交位置的原因。因此，相比于 SS-ILM 算法，SPLMTL 算法能更全面地发现流行社交位置。

表 3 SS-ILM 算法发现的 Top 10 流行社交位置

Table 3 Socially important locations Top-10 discovered by SS-ILM algorithm

序号	流行社交位置
1	秋叶原
2	新宿
3	涩谷
4	池袋
5	东京桥
8	品川
10	品川附近
6,7,9	秋叶原周边

5.3.4 算法中参数对算法的运行效率的影响

本节使用数据集 1 测试用户数量和 *min_LVR* 参数对算法的影响。在测试用户数量对算法影响的过程中将以 SS-ILM 算法的运行时间作为参考。

(1)用户数量的影响

在本实验中，参数设定如下：SS-ILM 算法中计算 SILU 所用的参数 *min_density* 设置为 0.06，*min_visit* 设置为0.05，SPLMTL 算法中 *min_LVR* 设置为 0.02，*min_UP* 设置为 0.01。用户人数分别设置为 400，800，1 200，1 600。图 10 显示了算法运行时间随用户数量变化的情况。随着用户数量的上升，算法运行时间也不断上升，且 SPLMTL 算法运行效率的增长率明显高于 SS-ILM 算法。其原因是 SPLMTL 算法需要对 SS-ILM 算法计算所得的 SILU 进行整理，收集各个位置的时间信息，这极大地影响了算法的运行效率。

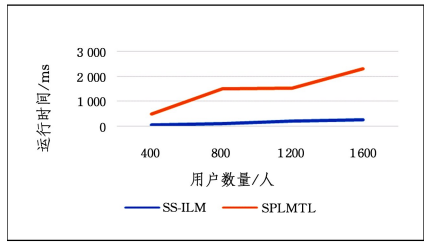

图 10 用户数量对算法 SPLMTL 运行效率的影响

Fig. 10 Effect of number of users on function efficiency of SPLMTL algorithm

(2)*min_LVR* 的影响

本实验参数设定如下：SS-ILM 算法中计算 SILU 所用的参数 *min_density* 设置为 0.06，*min_visit* 设置为 0.05，用户人数为 1 600，SPLMTL 算法中 *min_UP* 为 0.01，*min_LVR* 分别设置为 0.005，0.010，0.015，0.020，0.025，0.030。图 11 显示了算法运行时间随 *min_LVR* 变化的情况。

由图 11 可观察到，算法的运行效率随着 *min_LVR* 的上升而提高。但同时，由于数据的稀疏性，提升 *min_LVR* 也会导致数据被过多筛选，导致运行结果不理想。

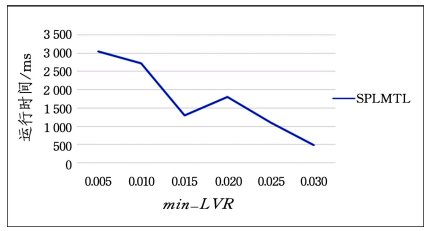

图 11 *min_LVR* 对算法 SPLMTL 运行效率的影响

Fig. 11 Effect of *min_LVR* threshold on function efficiency of SPLMTL algorithm

结束语 本文通过将时间因素引入位置挖掘，提出了一种带有时间标签的流行社交位置挖掘算法。该算法扩展 Dokuz 等提出的 SS-ILM 算法，使用位置访问率 *LVR* 对数据进行筛选，得到多个带有时间标签的〈位置-时间〉对，并使用用户流行度 *UP* 衡量每个〈位置-时间〉对在用户群中的流行程度，最终得到带有时间标签的流行社交位置。实验表明，本文提出的算法能够解决带时间标签的流行社交位置的发现问题。

将来的工作主要从以下两个方面进行。首先，将考虑通过计算用户之间的相似性来对相似用户的数据进行整理，得到期望的数据，以降低数据的稀疏性对算法的影响。然后，将考虑如何减少算法对原数据的读取次数，使算法性能得到提升。

参 考 文 献

[1] YUAN Q, CONG G, MA Z, et al. Time-aware point-of-interest recommendation[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2013.

[2] YUAN Q, CONG G, AIXIN S. Graph-based Point-of-interest Recommendation with Geographical and Temporal Influences [C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai: ACM Press, 2014: 659-668.

[3] GAO H, TANG J, HU X, et al. Exploring temporal effects for location recommendation on location-based social networks[C]// Proceeding of the 7th ACM Conference on Recommender Systems. Hong Kong: ACM Press, 2013: 93-100.

[4] BARABÁSI A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039): 207-211.

[5] FAN C, GUO J L, HAN X P, et al. A Review of Research on Human Dynamics[J]. Complex Systems and Complexity Science, 2011, 8(2): 1-17. (in Chinese)
樊超，郭进利，韩筱璞，等. 人类行为动力学研究综述[J]. 复杂系统与复杂性科学，2011，8(2)：1-17.

[6] DOKUZ A S, CELIK M. Discovering socially important locations of social media users[J]. Expert Systems With Applications, 2017(86): 113-124.

[7] JIMÉNEZ M, SORIANO J, CANTERA J M, et al. New Trends in Semantic-Based Location and Context-Aware Adaptation for Mobile Web Applications Development[J]. Current Topics in Medicinal Chemistry, 2012, 3(5): 513.

[8] MOKBEL M F, LEVANDOSKI J J. Toward context and preference-aware location-based services[C]// Proceedings of the 8th ACM International Workshop on Data Engineering for Wireless and Mobile Access. Providence RI: ACM Press, 2009: 25-32.

[9] MAROULIS S, BOUTSIS I, KALOGERAKI V. Context-aware point of interest recommendation using tensor factorization[C]// Proceedings of the 4th IEEE International Conference on Big Data. Washington DC: IEEE Press, 2016: 963-968.

[10] PARK H, RHO S, PARK J. A Link-Based Ranking Algorithm for Semantic Web Resources: A Class-Oriented Approach Independent of Link Direction[J]. Journal of Database Management, 2017, 22(1): 1-25.

[11] YAO L, SHENG Q Z, QIN Y, et al. Context-aware Point-of-Interest Recommendation Using Tensor Factorization with Social Regularization[C]// Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago: ACM Press, 2015: 1007-1010.

[12] CHEN G, LU J, ZHANG Z N. Research on topic-based key node discovery in mobile social network[J]. Application Research of Computers, 2017, 34(7): 2010-2015. (in Chinese)
陈功，卢菁，张仲楠. 基于主题划分的移动社交网络关键位置发现研究[J]. 计算机应用研究，2017，34(7)：2010-2015.

[13] LINDEN G, SMITH B, YORK J. Amazon. com Recommendations: Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2003, 7(1): 76-80.

[14] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the International Conference on World Wide Web. Hong Kong: ACM Press, 2001: 285-295.

[15] ZHANG C, WANG K. POI recommendation through cross-region collaborative filtering[J]. Knowledge and Information Systems, 2016, 46(2): 369-387.

[16] DAO T H, JEONG S R, AHN H. A novel recommendation model of location-based advertising: Context-Aware Collaborative Filtering using GA approach[J]. Expert Systems with Applications, 2012, 39(3): 3731-3739.

[17] CHEN C C, CHIANG M F, PENG W C. Mining and clustering mobility evolution patterns from social media for urban informatics[J]. Knowledge and Information Systems, 2016, 47(2): 381-403.

[18] CELIKTEN E, FALHER G L, MATHIOUDAKIS M. Modeling Urban Behavior by Mining Geotagged Social Data[J]. IEEE Transactions on Big Data, 2017, 3(2): 220-233.

[19] ZHAO K, TARKOMA S, LIU S Y, et al. Urban Human Mobility Data Mining: An Overview[C]// Proceedings of the 4th IEEE International Conference on Big Data. Washington DC: IEEE Press, 2016: 1911-1920.

[20] CELIK M, DOKUZ A S. Discovering Socio-spatio-temporal Important Location of Social Media Users[J]. Journal of Computation Science, 2017(22): 85-98.

[21] ZHANG D C, LI M, WANG C D. Point of interest recommendation with social and geographical influence[C]// Proceedings of the 4th IEEE International Conference on Big Data. Washington DC: IEEE Press, 2016: 1070-1075.

[22] YE M, YIN P, LEE W C, et al. Exploiting geographical influence for collaborative point-of-interest recommendation[C]// Proceedings of the 34th ACM SIGIR international conference on Research and development in Information Retrieval. Beijing: ACM Press, 2011: 325-334.

[23] LEVANDOSKI J J, SARWAT M, ELDAWY A, et al. LARS: A Location-Aware Recommender System[C]// Proceedings of the IEEE 28th International Conference on Data Engineering. Arlington VA: IEEE Computer Society, 2012: 450-461.

[24] BAO J, ZHENG Y, MOKBEL M F. Location-based and preference-aware recommendation using sparse geo-social networking data[C]// Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Sys-

tems. Redondo Beach CA: ACM Press, 2012: 199-208.

[25] CUI C, SHEN J, NIE L, et al. Augmented Collaborative Filtering for Sparseness Reduction in Personalized POI Recommendation[J]. ACM Transactions on Intelligent Systems and Technology, 2017, 8(5): 1-23.

[26] 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. San Diego CA: ACM Press, 2011: 1082-1090.

[27] ZHANG J D, CHOW C Y. TICRec: A Probabilistic Framework to Utilize Temporal Influence Correlations for Time-Aware Location Recommendations[J]. IEEE Transactions on Services Computing, 2016, 9(4): 633-646.

[28] BALTRUNAS L, AMATRIAIN X. Towards time-dependant recommendation based on implicit feedback[R]. New York: Workshop on Context-aware Recommender Systems, 2009.

[29] SI L F, ZHANG F Z, LIU W Y. A Time-aware POI Recommendation Method Exploiting User-based Collaborative Filtering and Location Popularity[C]// Proceedings of the 2nd International Conference on Communications, Information Management and Network Security. Beijing: Destech Publicat. 2017: 17-25.

[30] GAO H, TANG J, HU X, et al. Modeling temporal effects of human mobile behavior on location-based social networks[C]// Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco CA: ACM Press, 2013: 1673-1678.

[31] ZHANG J D, CHOW C Y. Spatiotemporal Sequential Influence Modeling for Location Recommendations: A Gravity-based Approach[J]. ACM Transactions on Intelligent Systems & Technology, 2015, 7(1): 1-25.

[32] YE M, JANOWICZ K, LEE W C. What you are is when you are: the temporal dimension of feature types in location-based social networks[C]// Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Chicago IL: ACM Press, 2011: 102-111.

[33] LI X, JIANG M, HONG H, et al. A Time-Aware Personalized Point-of-Interest Recommendation via High-Order Tensor Factorization[J]. Acm Transactions on Information Systems, 2017, 35(4): 1-23.

[34] YING Y, CHEN L, CHEN G. A temporal-aware POI recommendation system using context-aware tensor decomposition and weighted HITS[J]. Neurocomput., 2017(242): 195-205.

[35] YANG D, ZHANG D, ZHENG V W, et al. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2014, 45(1): 129-142.

[36] LI M, WANG X C, ZHANG J, et al. Study on Check-in and Related Behaviors of Location-based Social Network Users[J]. Computer Science, 2013, 40(10): 72-76. (in Chinese)
李敏, 王晓聪, 张军, 等. 基于位置的社交网络用户签到及相关行为研究[J]. 计算机科学, 2013, 40(10): 72-76.