



论文

基于时序特征的移动模式挖掘

陈劭, 刘洋*, 王月, 禹晓辉

山东大学计算机学院, 济南 250001

* 通信作者. E-mail: yliu@sdu.edu.cn

收稿日期: 2015-11-10; 接受日期: 2015-12-28; 网络出版日期: 2016-09-09

国家自然科学基金 (批准号: 61272092, 61572289)、山东省自然科学基金 (批准号: ZR2012FZ004, ZR2015FM002)、山东省科技发展计划基金 (批准号: 2014GGE27178)、国家重点基础研究发展计划 (973 计划) (批准号: 2015CB352500) 和泰山学者计划基金资助项目

摘要 定位设备 (如 GPS) 的广泛使用产生了大量时空轨迹, 合理地利用这些轨迹数据可以帮助挖掘用户移动模式. 本文基于轨迹的时序特征提出一种新的模型来挖掘用户隐含移动模式. 考虑到轨迹的特点: (1) 位置顺序对于理解用户移动模式很重要; (2) 用户的移动模式在不同时间段存在变化, 本文提出的方法首先对位置序列进行建模, 然后将时间信息加入到模型中. 为了验证模型的有效性, 本文在 Gowalla 签到数据集上进行了详细的实验, 实验结果表明提出的模型优于传统的 LDA 模型及 T-BiLDA 模型.

关键词 时空轨迹 移动模式挖掘 时序特征 签到数据 生成模型

1 引言

随着全球定位技术的快速发展, 获取大量的位置数据已经变得越来越简便. 比如, 人们常常利用智能手机在感兴趣的位置进行“签到”, 然后在社交网络 (如 Foursquare、街旁) 上与好友实时分享. 签到记录至少会包含对象、位置和时间 3 种属性, 我们把同一对象在连续时间内经过的所有空间位置收集起来, 就可以得到一条时空轨迹.

如何从时空轨迹数据中挖掘用户移动模式是本文关注的重点. 用户移动模式被认为是隐含在用户时空轨迹之中的, 例如, 现有两条轨迹 A 和 B, 分别为“安康小区→绿茶餐厅→家乐福超市”, “安康小区→肯德基→沃尔玛超市”. 虽然这两条轨迹通过的位置点不同, 但是隐含着相同的移动模式, 即“家→就餐→购物”. 此外, 用户移动模式和时间密切相关, 人们在不同的时间段有着不同的移动模式. 比如, 工作日早上 8 点, “家→工作”这个移动模式出现的概率很大, 而“家→娱乐”这个移动模式则更可能出现在晚上. 因此, 挖掘用户移动模式是一项极具挑战性的工作.

本文基于 LDA^[1] 提出了一种新的轨迹时序特征模型 (temporal and sequential model, TSM) 来发现用户隐含移动模式. LDA 是由 Blei 等提出的, 利用文本中词项的共现特征来发现文本的主题结构. 在 LDA 中, 每个文档中的主题服从多项分布; 每个主题下的单词也服从多项分布. 但是, 该方法并不

引用格式: 陈劭, 刘洋, 王月, 等. 基于时序特征的移动模式挖掘. 中国科学: 信息科学, 2016, 46: 1288–1297, doi: 10.1360/N112015-00237

能直接用于从时空轨迹中发现用户移动模式. 首先, 一条轨迹通常只包含几个位置点, 直接应用 LDA 会存在严重的数据稀疏问题^[2~4]. 因此, 将同一用户的所有轨迹合并起来, 作为一条训练数据. 其次, 在主题模型中, 文档中的单词通常被认为是无序的, 但是在轨迹中位置转移对于理解用户移动模式非常重要. 因此, 模型对位置序列进行建模, 因为不同的位置序列通常表示不同的移动模式. 例如, 两条轨迹“安康小区→家乐福超市”和“家乐福超市→安康小区”, 虽然包含相同的位置点, 但是其隐含的用户移动模式显然是不同的. 最后, 用户移动模式往往是周期性的, 并且在不同的时间段展现出不同的特性. 例如, 用户在工作日早上的移动模式往往和“上班”有关, 而晚上的移动模式则更多元化, 和“娱乐、吃饭、购物”等有关. 所以将轨迹中的时间信息加入到模型中.

本文的主要贡献如下:

- (1) 考虑到轨迹的特点, 基于轨迹时序特征提出了一种新的模型来挖掘用户隐含移动模式.
- (2) 基于提出的模型, 轨迹可以用隐含移动模式的分布来表示, 这为轨迹聚类、位置预测等应用奠定了基础.
- (3) 在 Gowalla 签到数据集上进行了深入的实验研究. 实验结果表明本文提出的模型能够清楚地发现人们隐含的移动模式, 并且表现得比 LDA 更好.

2 相关工作

近年来, 随着定位技术的发展, 出现了各种类型的时空轨迹数据, 如手机传感器搜集的数据^[5]和出租车轨迹数据^[6~8]. 微软亚洲研究院、浙江大学和上海交通大学等也在位置预测、城市规划等领域作了有益的探索, 并取得了很多成果^[6, 9, 10]. 但是传统的方法多采用 Markov 模型^[11~13]、频繁模式^[14]等来学习移动模式. 这些方法是将轨迹中的位置点映射成一系列离散的状态, 然后利用概率统计方法描述人们在不同状态之间的移动模式. 这类工作的重点往往是刻画观察到的时空轨迹数据本身, 而非数据背后的隐含移动模式.

与之对应, 近年来也出现了一些利用主题模型来挖掘用户隐含移动模式的方法, 因为从轨迹中发现移动模式自然地对应于从文本中发现主题. 其中, LDA^[1]是一种使用最广泛的主题模型. 在 LDA 中, 一篇文档被认为由多个主题组成, 一个主题是关于若干单词的多项分布. Zheng 等^[15]直接利用 LDA 试图发现时空轨迹中的异常行为, 他们将轨迹数据对应成文档. Long 等^[16]使用 LDA 模型从签到记录中发现当地地理主题. 具体而言, 他们将签到记录中的每一个地点当作一个单词, 将包含一个用户所有签到地点的轨迹当作一个文档. Cranshaw 等^[17]利用 LDA 模型, 通过分析地点类别中的共现模式来发现典型的邻域. 但是这些方法都是单纯地使用 LDA, 没有考虑到轨迹的时序特征. 同时, 也有一些方法^[18, 19]利用概率主题模型来发现时间活动序列中的隐含模式. Huynh 等^[18]利用 LDA 去建模序列中每一个活动出现的次数来发现隐含的模式, 可是他们没有考虑到序列活动的先后顺序. Shen 等^[19]改进了这个方法, 所提出的 T-BiLDA 方法除了考虑序列中活动的数目, 还考虑了活动之间的转移概率以及每个活动发生的时间来帮助发现更有意义的隐含模式.

此外, 在短文本中直接采用传统的主题模型 (例如 LDA) 可能面临严重的数据稀疏问题^[2, 3], 因此现有的一些方法注重解决这个问题. 一个简单而常用的方法是在训练模型之前将短文本聚集成冗长的文档. Hong 和 Davison^[4]在训练 LDA 模型之前, 将由同一用户生成的所有短消息聚集成一个用于训练的长文档, 这比直接在短文本上应用 LDA 模型要好得多. 所以本文也采用这种方法, 即在训练模型之前, 将属于同一对象的所有轨迹聚合起来构成轨迹集合.

表 1 用户签到记录例子
Table 1 Examples of user check-in records

User id	Check-in time	Latitude	Longitude	Location id
18	2010-07-23 22:07:33	33.133929563	-117.192530062	1490317
63	2010-09-12 23:51:15	37.776422099	-122.3943257332	16947
73	2010-03-14 15:00:39	37.4166508333	-122.1302035167	482948

3 基于时序特征的轨迹建模

本节首先介绍搜集的 Gowalla 上的签到数据的属性, 然后对这些数据进行预处理, 并给出用户轨迹的定义, 最后详细介绍了本文提出的模型.

3.1 数据

Gowalla 是一个基于位置的社交网站, 用户可以在网站上通过签到和朋友们分享自己的位置. 本文搜集了美国 107092 名用户在 12 个月 (2009 年 10 月到 2010 年 9 月) 的 6442892 条签到记录. 每条签到记录包含 5 种属性: 用户编号、签到时间、签到位置的纬度、签到位置的经度以及签到位置编号. 表 1 展示了 3 名用户的 3 条签到记录.

3.2 用户轨迹定义

为了进行时空轨迹挖掘, 首先将同一用户的签到记录收集起来, 然后依照定义 1 构造成轨迹.

定义1 (轨迹) 对于某位用户, 定义轨迹 T 为一系列按照时间排序的位置点, $\langle (l_1, t_1), (l_2, t_2), \dots, (l_n, t_n) \rangle$, 其中 l_i 表示用户签到位置编号, t_i 表示该位置点对应的签到时间.

这里, 两个相邻位置 l_i 和 l_{i+1} 之间的时间差应该小于一个时间阈值, 这个值是由用户根据自己的数据特点决定的, 本文设定时间阈值为 1 小时.

定义2 (r 阶序列) 对于一条轨迹 $\langle (l_1, t_1), (l_2, t_2), \dots, (l_n, t_n) \rangle$, 定义 r 阶序列 s 为 $(r+1)$ 个相邻的位置点, $s_i = \langle l_i, \dots, l_{i+r} \rangle$, 其中 $1 \leq i \leq n-r$.

由于两个相邻位置的时间差小于 1 小时, 利用 r 阶序列中所有位置点对应时间的平均值所在的时间段作为 r 阶序列的时间. 本文设置 1 天为 24 个时间段. 进而, 可以将轨迹 $\langle (l_1, t_1), (l_2, t_2), \dots, (l_n, t_n) \rangle$ 转变成 $\langle (s_1, \bar{t}_1), (s_2, \bar{t}_2), \dots, (s_{n-r}, \bar{t}_{n-r}) \rangle$. 在后文中, 除非作出特殊说明, 否则时间 $t, \bar{t}$ 都表示所对应的时间段.

定义3 (轨迹单元) 对于一条轨迹 $\langle (s_1, \bar{t}_1), (s_2, \bar{t}_2), \dots, (s_{n-r}, \bar{t}_{n-r}) \rangle$, 定义 $(s_i, \bar{t}_i)$ 为一个基本的轨迹单元.

模型中用到的相关符号以及对应解释如表 2 所示.

3.3 模型介绍

近年来, LDA^[1] 被广泛应用到文本挖掘中, 并取得了显著的成果. 轨迹数据和文本数据有很高的相似度, 他们都是由一系列离散的点 (位置或单词) 组成. 基于主题模型的文本挖掘采用词袋模型, 忽略单词间的依赖关系; 而轨迹中的位置序列和时间对于理解用户隐含移动模式非常重要. 具体来说, 轨迹中的位置序列能够反映用户不同的隐含移动模式, 例如从住宅区到超市显示用户的隐含移动模式为“购物”, 从超市到住宅区显示移动模式为“回家”. 此外, 不同的时间段通常存在不同的移动模式, 例如

表 2 符号和描述

Table 2 Notations and descriptions

Notations	Descriptions
l, s, t	Location, sequence, time
T, z	Trajectory, latent mobility pattern
T_s	Number of sequences of trajectory T
S, B	Number of sequences, number of time bins
M, K	Number of trajectories, number of latent mobility patterns
θ	Trajectory specific latent mobility pattern distribution
ϕ	Latent mobility pattern specific r th-order sequence distribution
φ	Latent mobility pattern specific time distribution
α, β, γ	Dirichlet priors

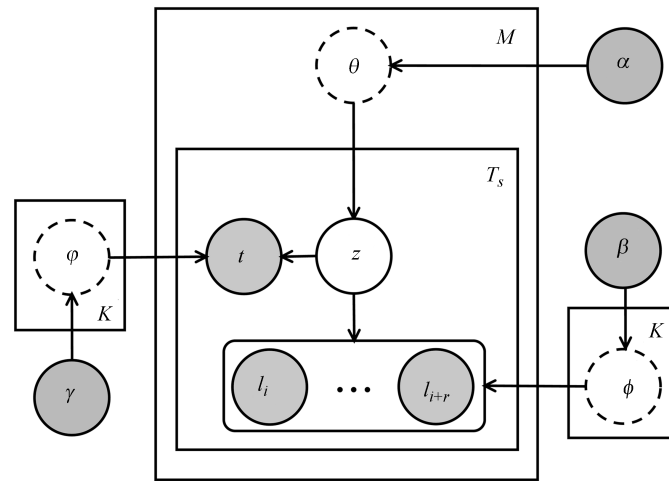


图 1 模型网络结构

Figure 1 The network structure of the proposed model

周一早上用户的移动模式更可能为“上班”, 周末下午用户的移动模式更可能为“购物”或者“娱乐”. 因此, 首先对位置序列进行建模, 然后将时间信息加入到模型中, 来发现用户隐含移动模式.

本文基于轨迹的时序特征提出一种新的模型对轨迹中的 r 阶位置序列和时间进行建模, 其结构如图 1 所示.

在模型中, 假设每条轨迹存在若干个隐含移动模式 z . 对于一条轨迹 $T = m$, 隐含移动模式是通过多项分布 θ_m 产生的. 对于每种移动模式, r 阶位置序列是通过多项分布 ϕ_z 产生的, 每个序列的时间是通过多项分布 φ_z 产生的. 具体来说, 一条轨迹的生成过程如下所述:

- (1) 采样 K 个隐含模式 - 位置序列多项分布 ϕ_z , 其中先验分布为 Dirichlet 分布, $\phi_z \sim \text{Dir}(\beta)$.
- (2) 采样 K 个隐含模式 - 时间段多项分布 φ_z , 其中先验分布为 Dirichlet 分布, $\varphi_z \sim \text{Dir}(\gamma)$.
- (3) 对于轨迹 $T = m$, 采样一个轨迹 - 隐含模式多项分布 θ_m , 其中先验分布为 Dirichlet 分布, $\theta_m \sim \text{Dir}(\alpha)$; 然后对于轨迹中的每个轨迹单元:

- 从轨迹 - 隐含模式多项分布 θ_m 采样一种隐含模式 z_m ;

- 从隐含模式 – 位置序列多项分布 ϕ_{z_m} 采样一个 r 阶位置序列;
- 从隐含模式 – 时间段多项分布 φ_{z_m} 采样一个时间段.

给定先验参数 α, β, γ , 轨迹 – 隐含模式多项分布 θ , K 种隐含移动模式以及 T_s 个轨迹基本单元, 轨迹的边缘分布可以表示为

$$p(\{s_i, t_i\}_{i=1}^{T_s} | T; \alpha, \beta, \gamma) = \int p(\theta | \alpha) \prod_{i=1}^{T_s} \sum_{k=1}^K p(z_k | T; \theta) \cdot p(s_i | z_k; \beta) \cdot p(t_i | z_k; \gamma) d\theta. \quad (1)$$

4 模型求解和参数估计

第 3 节详细阐述了轨迹的生成过程, 本节将给出模型求解的过程以及参数估计. 求解模型的关键问题是计算在给定轨迹条件下隐含移动模式的后验分布, 即 $p(z|s, t)$. 但是, 准确的推断对于后验分布是不现实的. 这里采用主题模型中常用的近似推断方法 Gibbs 采样来解决这个问题.

首先, 明确模型中的变量和参数. 模型中的隐含参数是移动模式 z , 观测变量是位置序列 s 和时间段 t , α, β, γ 是 Dirichlet 分布的先验参数. 需要估计的参数是轨迹 – 隐含模式多项分布 θ , 隐含模式 – 位置序列多项分布 ϕ 和隐含模式 – 时间段多项分布 φ . 在 Gibbs 采样过程中, 首先需要计算后验分布模型 $p(z_i = k | z_{-i}, s, t)$:

$$\begin{aligned} p(z_i = k | z_{-i}, s, t) &= \frac{p(s|z)}{p(s_{-i}|z_{-i})p(s_i)} \frac{p(t|z)}{p(t_{-i}|z_{-i})p(t_i)} \frac{p(z)}{p(z_{-i})} \\ &\propto \frac{\Delta(n_k^s + \beta)}{\Delta(n_{k,-i}^s + \beta)} \cdot \frac{\Delta(n_k^t + \gamma)}{\Delta(n_{k,-i}^t + \gamma)} \cdot \frac{\Delta(n_m + \alpha)}{\Delta(n_{m,-i} + \alpha)} \\ &= \frac{n_{k,-i}^s + \beta_s}{\sum_{s=1}^S n_{k,-i}^s + \beta_s} \cdot \frac{n_{k,-i}^t + \gamma_t}{\sum_{t=1}^B n_{k,-i}^t + \gamma_t} \cdot \frac{n_{m,-i}^k + \alpha_k}{\sum_{k=1}^K n_{m,-i}^k + \alpha_k}, \end{aligned} \quad (2)$$

其中, $z_i = k$ 表示轨迹中第 i 个轨迹单元的隐含移动模式为 k . z_{-i} 表示除了第 i 个轨迹单元的其余轨迹单元, s_{-i} 和 t_{-i} 也有相同的含义. $n_{m,-i}^k$ 表示在第 m 条轨迹中, 去除第 i 个轨迹单元后, 隐含移动模式 k 出现的次数, $n_{k,-i}^t$ 和 $n_{k,-i}^s$ 也有相同的含义. S, K, B 分别代表位置序列的总数, 隐含移动模式的总数和时间段的总数.

根据 Gibbs 采样, 得到参数估计为

$$\theta_{m,k} = \frac{n_m^k + \alpha_k}{\sum_{k=1}^K n_m^k + \alpha_k}, \quad \phi_{k,s} = \frac{n_k^s + \beta_s}{\sum_{s=1}^S n_k^s + \beta_s}, \quad \varphi_{k,t} = \frac{n_k^t + \gamma_t}{\sum_{t=1}^B n_k^t + \gamma_t}. \quad (3)$$

5 实验分析

本节利用 Gowalla 签到数据集对提出的模型进行评测. 首先介绍下实验中的数据集和实验设置, 然后通过用户隐含移动模式发现来评测模型的效果, 最后和 LDA 进行比较.

5.1 数据集和实验设置

为了更好地理解人们的移动模式, 从 Gowalla 上收集了人们从 2009 年 10 月到 2010 年 9 月的签到记录, 数据集包含 107092 名用户在 1280969 个兴趣点的 6442892 条签到记录, 每条记录包含 5 种属

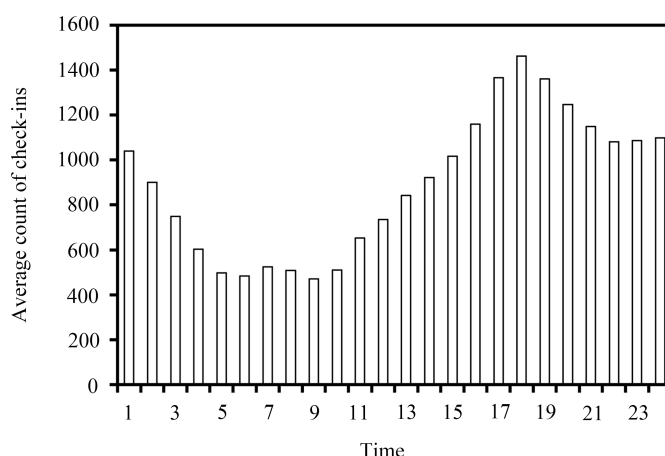


图 2 不同时间段的签到次数

Figure 2 The number of check-ins in different time periods

性: 用户、签到时间、经度、纬度以及签到位置. 这里将一天的时间平均划分成 24 个时间段, 并统计用户这一年时间平均每天的签到次数, 如图 2 所示.

从图 2 中可知, (1) 用户在 3 点到 11 点间签到次数比较少, 因为大部分人在睡觉或者上班途中, 无暇使用社交网络; (2) 从 12 点开始, 用户签到数量开始增加, 因为部分用户会在上班时使用社交网络; (3) 下午 18 点, 用户可能正在回家途中, 或者在饭店吃饭, 有比较多的空闲时间, 并且愿意将当前活动分享给朋友, 所以签到次数达到顶峰; (4) 18 点以后, 用户签到次数开始降低, 并且保持相对稳定的状态直到凌晨 2 点, 这个时间段用户多在娱乐, 也会乐意通过 Gowalla 和朋友分享自己的位置, 所以签到次数也比较多.

为了更好地训练模型, 过滤掉一天签到次数少于 3 条的记录, 然后基于定义构造轨迹, 最后, 为了解决数据稀疏性问题, 将同一用户的所有轨迹合并起来作为 1 条训练数据. 具体来说, 假设用户有 3 条轨迹 $\langle (s_1, \bar{t}_1), (s_2, \bar{t}_2), (s_3, \bar{t}_3) \rangle, \langle (s_4, \bar{t}_4), (s_2, \bar{t}_2) \rangle, \langle (s_5, \bar{t}_5), (s_3, \bar{t}_3) \rangle$, 那么合并后的轨迹为 3 条轨迹的简单连接, 即 $\langle (s_1, \bar{t}_1), (s_2, \bar{t}_2), (s_3, \bar{t}_3), (s_4, \bar{t}_4), (s_2, \bar{t}_2), (s_5, \bar{t}_5), (s_3, \bar{t}_3) \rangle$. 在所有的实验中, 设置 Dirichlet 超参数 $\alpha = 50/K$, $\beta = 0.01$, $\gamma = 0.01$. 在 Gibbs 采样过程中, 设置迭代次数为 1000, 并计算 10 次实验的平均结果来评测本文提出的模型.

给定训练轨迹, 能够评测用户隐含移动模式的一致性. 这里采用广泛使用的点间互信息 (point-wise mutual information) 来计算隐含移动模式的一致性^[20]. 点间互信息是用来计算每种移动模式中前 10 个位置序列中所有序列对的相关程度, 形式化定义为

$$\text{PMI}(s) = \frac{1}{45} \sum_{i < j} \text{pmi}(s_i, s_j), i, j \in \{1, \dots, 10\},$$

$$\text{pmi}(s_i, s_j) = \log \frac{p(s_i, s_j)}{p(s_i)p(s_j)},$$

其中, $p(s_i)$ 表示位置序列 s_i 在轨迹中出现的概率, $p(s_i)$ 和 $p(s_i, s_j)$ 可以利用所有的轨迹提前计算出来. 显然, PMI 越大表示隐含移动模式的一致性越好.

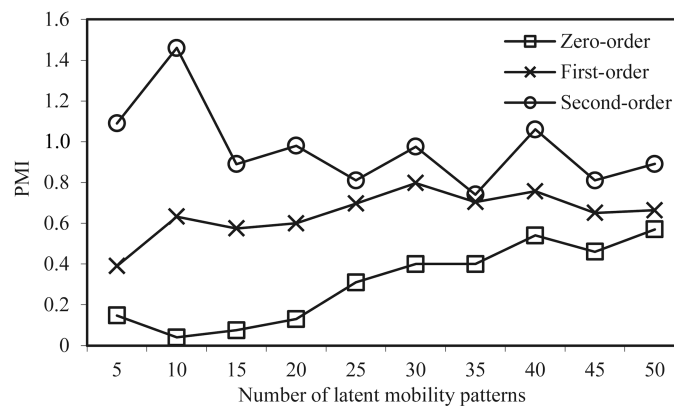


图 3 隐含移动模式数目的影响

Figure 3 The impact of the number of latent mobility patterns

表 3 位置类别分类

Table 3 Location category taxonomy

Id	Location category	Id	Location category	Id	Location category
1	Parking	6	Residence	11	Food
2	Sport	7	Station	12	Shopping
3	Hospital	8	Bank	13	Education
4	Hotel	9	Company	14	Entertainment
5	Night life	10	Road	15	Life service

5.2 参数设置和调整

分别设置位置序列为 0 阶, 1 阶和 2 阶, 并调整隐含移动模式的数目 (K) 来评测它的影响. 结果如图 3 所示.

从图 3 中可知 (1) 对于不同阶数的模型, PMI 达到最大值时对应的隐含移动模式数目是不相同的, (2) 1 阶模型表现得比 0 阶模型好, 2 阶模型表现得最好, 这也证明了位置序列对于挖掘用户隐含移动模式是非常重要的.

为了进一步验证用户隐含移动模式的质量, 人工标注了一些位置序列, 来可视化用户移动模式. 某些用户签到位置存在对应的语义标签 (商场、电影院等), 可以为少量位置点爬取语义标签, 然后利用爬取的语义标签显式地标记每种隐含移动模式. 由于轨迹中的位置编号不具有任何语义信息, 利用 Foursquare 提供的 api (<https://developer.foursquare.com/docs/venues/search>), 基于位置的经纬度返回位置对应的类别, 并用来标记隐含的移动模式. 参考文献 [21] 中给出的位置类别分类并简化它以满足我们的需求. 新的位置类别分类如表 3 所示.

首先利用 2 阶模型训练数据, 然后随机选择 3 种隐含移动模式, 并在每种模式中选择前 5 个 2 阶位置序列. 表 4 展示了这 3 种移动模式下的前 5 个位置序列对应的语义标签. 从表 4 中可以得到如下结论: (1) 移动模式 1 主要是关于“购物”的; (2) 移动模式 2 主要是“娱乐”和“吃饭”, 因为吃饭和娱乐经常是相伴发生的; (3) 移动模式 3 主要是“上下班”, 因为在上下班途中人们通常有比较多的空闲时间去签到. 总体来说, 我们能够明确不同的用户隐含移动模式. 实际上, 很多位置点对应的类别是缺失的并且很难获得, 基于提出的模型, 仅需要少量的位置类别就能明确用户隐含的移动模式,

表 4 用户移动模式
Table 4 User mobility patterns

Mobility pattern 1	Mobility pattern 2	Mobility pattern 3
Residence→Shopping→Shopping	Food→Parking→Entertainment	Residence→Station→Company
Road→Station→Shopping	Residence→Station→Food	Station→Station→Road
Shopping→Entertainment→Road	Parking→Food→Shopping	Road→Station→Station
Shopping→Road→Residence	Entertainment→Road→Station	Station→Station→Company
Road→Station→Life service	Station→Entertainment→Road	Station→Station→Road

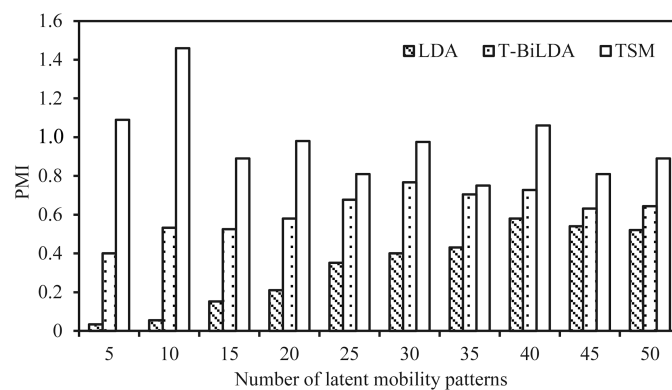


图 4 不同方法的结果比较

Figure 4 Performance comparison with baselines

这对于提高基于位置的推荐效果是有很大帮助的.

5.3 方法比较

本文的模型是基于主题模型提出的, 为了进一步验证模型的有效性, 将它和 LDA^[1] 以及 LDA 在时序任务挖掘应用中的变种 T-BiLDA^[18] 进行比较.

LDA^[1]. LDA 是一种文档生成模型, 其中一篇文档被认为由多个主题组成, 一个主题是关于若干单词的多项分布. 在移动模式发现应用中, 把位置点当作单词, 移动模式当作主题, 轨迹集合当作文档.

T-BiLDA^[19]. 在时间活动序列分析任务中, T-BiLDA 同时考虑了序列的 3 种特征: (1) 不同活动的数目, (2) 活动之间的 Markov 转移以及 (3) 每个活动发生的时间, 这些特征的组合能够帮助发现更有意义的隐含主题分布.

在所有方法中, 设置 Dirichlet 超参数 $\alpha = 50/K$, $\beta = 0.01$, 同样, 也将同一用户的所有轨迹合并起来作为 1 条训练数据.

图 4 展示了提出的模型 TSM 和 LDA, T-BiLDA 方法的比较结果. 能够明显地看到, 对于任何数目的隐含移动模式, TSM 比 LDA 和 T-BiLDA 表现得更好. LDA 在建模过程中只考虑位置因素, 没有考虑位置序列和时间因素, 所以比其他方法表现得差. T-BiLDA 考虑了位置转移, 所以表现得比 LDA 更好. TSM 考虑了轨迹中的位置序列和时间因素, 并且表现得最好. 原因是两方面的: (1) 位置序列在轨迹隐含移动模式挖掘中起到重要的作用; (2) TSM 中加入了时间因素, 使得相似的用户移动行为处于同一种模式中.

6 结论

基于对大量轨迹数据的观察和学习, 本文提出了一种新颖的时序特征分析模型用以挖掘用户的隐含移动模式. 以该方法为基础, 可以开展广泛的时空数据挖掘研究 (如基于位置的推荐, 广告的定向投放和时空轨迹聚类). 为了验证方法的有效性, 本文利用真实的 Gowalla 签到数据集进行了深入的实验, 实验结果证明提出的模型是有效的, 并且比 LDA 和 T-BILDA 表现得更好.

在未来的工作中, 我们计划开展基于该模型的“下一位置预测”问题研究. 现实中的轨迹数据往往规模巨大, 同时又涉及海量的用户. 为了更好地理解用户的移动模式, 可以考虑采用轨迹聚类的方法对不同的用户轨迹加以区分. 而通过本文所提出的方法, 用户轨迹可以通过使用定长的隐含移动模式向量来表示, 这为开展有效的轨迹聚类提供了帮助. 更进一步地, 我们将针对具有相似移动模式的特定用户组群进行模式挖掘, 从而提供更加精确的位置预测.

参考文献

- 1 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *J Mach Learn Res*, 2003, 3: 993–1022
- 2 Tang J, Meng Z, Nguyen X, et al. Understanding the limiting factors of topic modeling via posterior contraction analysis. In: *Proceedings of the 31st International Conference on Machine Learning*, Beijing, 2014. 190–198
- 3 Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts. In: *Proceedings of the 22nd International Conference on World Wide Web*. New York: ACM, 2013. 1445–1456
- 4 Hong L, Davison B D. Empirical study of topic modeling in twitter. In: *Proceedings of the 1st Workshop on Social Media Analytics*, Washington, 2010. 80–88
- 5 Farrahi K, Gatica-Perez D. Discovering routines from large-scale human locations using probabilistic topic models. *Trans Intell Syst Tech*, 2011, 2: 3
- 6 Yuan J, Zheng Y, Zhang L, et al. Where to find my next passenger. In: *Proceedings of the 13th International Conference on Ubiquitous Computing*, Beijing, 2011. 109–118
- 7 Pan G, Qi G, Wu Z, et al. Land-use classification using taxi GPS traces. *IEEE Trans Intell Transport Syst*, 2013, 14: 113–123
- 8 Castro P S, Zhang D, Chen C, et al. From taxi GPS traces to social and community dynamics: a survey. *Comput Surv*, 2013, 46: 17
- 9 Li X, Pan G, Wu Z, et al. Prediction of urban human mobility using large-scale taxi traces and its applications. *Front Comput Sci*, 2012, 6: 111–121
- 10 Zheng Y, Liu Y, Yuan J, et al. Urban computing with taxicabs. In: *Proceedings of the 13th International Conference on Ubiquitous Computing*, Beijing, 2011. 89–98
- 11 Chen M, Liu Y, Yu X. NLPMM: a next location predictor with Markov modeling. In: *Advances in Knowledge Discovery and Data Mining*. Berlin: Springer, 2014. 186–197
- 12 Chen M, Liu Y, Yu X. Predicting next locations with object clustering and trajectory clustering. In: *Advances in Knowledge Discovery and Data Mining*. Berlin: Springer, 2015. 344–356
- 13 Chen M, Yu X, Liu Y. Mining moving patterns for predicting next location. *Inf Syst*, 2015, 54: 156–168
- 14 Monreale A, Pinelli F, Trasarti R, et al. Wherenext: a location predictor on trajectory pattern mining. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, 2009. 637–646
- 15 Zheng B, Fan X, Li M, et al. Trajectory segment-based abnormal behavior detection method using LDA model. *J Comput Appl*, 2015, 35: 515–518
- 16 Long X, Jin L, Joshi J. Exploring trajectory-driven local geographic topics in foursquare. In: *Proceedings of the ACM Conference on Ubiquitous Computing*, Pittsburgh, 2012. 927–934
- 17 Cranshaw J, Yano T. Seeing a home away from the home: distilling proto-neighborhoods from incidental data with latent topic modeling. In: *Proceedings of the 1st Workshop on Computational Social Science and the Wisdom of Crowds at NIPS*, Vancouver, 2010. 10
- 18 Huynh T, Fritz M, Schiele B. Discovery of activity patterns using topic models. In: *Proceedings of the 10th Interna-*

- tional Conference on Ubiquitous Computing. New York: ACM, 2008. 10–19
- 19 Shen Z, Luo P, Xiong Y, et al. Topic modeling for sequences of temporal activities. In: Proceedings of the 9th IEEE International Conference on Data Mining, Miami, 2009. 980–985
 - 20 Newman D, Bonilla E V, Buntine W. Improving topic coherence with regularized topic models. In: Proceedings of Advances in Neural Information Processing Systems, Granada, 2011. 496–504
 - 21 Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, 2012. 186–194

Mining mobility patterns based on temporal and sequential features

Meng CHEN, Yang LIU*, Yue WANG & Xiaohui YU

School of Computer Science and Technology, Shandong University, Jinan 250001, China

*E-mail: yliu@sdu.edu.cn

Abstract The wide-spread use of positioning devices (e.g., GPS) has given rise to a mass of spatio-temporal trajectories, which enable us to mine user-mobility patterns. In this paper, we proposed a model based on sequential and temporal trajectory features to mine people's latent movement patterns. Considering the following trajectory characteristics—(1) location sequences play a pivotal role in understanding user-mobility patterns, and (2) user-mobility patterns change over time—we first modeled the location sequences and then incorporated the temporal information into the model. To verify the effectiveness of our model, we performed thorough empirical studies on a check-in dataset of the Gowalla social network. The experimental results confirmed that the proposed method performed better than Latent Dirichlet Allocation (LDA) and T-BiLDA.

Keywords spatio-temporal trajectory, mobility patterns mining, temporal and sequential features, check-in data, generative model



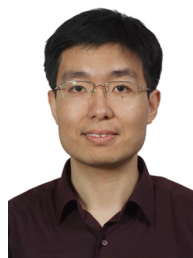
Meng CHEN was born in 1990. He received his B.S. degree from Shandong University, China. He is currently a Ph.D. student in the School of Computer Science and Technology, Shandong University. His research interests are in the areas of trajectory mining and urban computing.



Yang LIU was born in 1977. She received her Ph.D. degree from York University, Canada. She is currently an associate professor in the School of Computer Science and Technology, Shandong University, China. Her main research areas are data mining and information retrieval.



Yue WANG was born in 1993. She received her B.S. degree from Shandong University, China. She is currently a postgraduate in the School of Computer Science and Technology, Shandong University. Her research interests include trajectory mining and recommendation systems.



Xiaohui YU was born in 1977. He received his Ph.D. degree from the University of Toronto, Canada. His research interests are in the areas of database systems and data mining. He is a professor in the School of Computer Science and Technology, Shandong University, China, and an associate professor in the School of Information Technology, York University, Canada (on leave). He is a member of the IEEE.