适用于家乡和外地用户的位置感知推荐系统

**摘要**

空间项目推荐已成为帮助人们发现有趣位置的重要手段，尤其是当人们访问不熟悉的地区时。 目前的一些研究主要集中在根据用户的登记记录对空间项目推荐的个人和集体地理偏好进行建模，但他们未能探索用户兴趣跨越地理区域的现象，用户在不同区域旅行时会表现出不同的兴趣。此外，他们忽视了公众评论对于随后的用户登录行为的重要性。因此，有必要在正确的位置向正确的用户推荐正确的商品。在本文中，我们提出了一个潜在的概率生成模型LSARS来模拟用户在家乡和城外情况下的签入活动的决策过程。 针对用户的兴趣需求和人群情绪，从空间项目和用户评论的内容中学习位置感知和情感感知的个人兴趣。最后，我们将LSARS部署到两个实际的应用场景：空间项目推荐和目标用户发现。在两个大规模的基于位置的社交网络(LBSNs)数据集上的广泛实验表明，与现有的最先进的方法相比，LSARS获得了更好的性能。

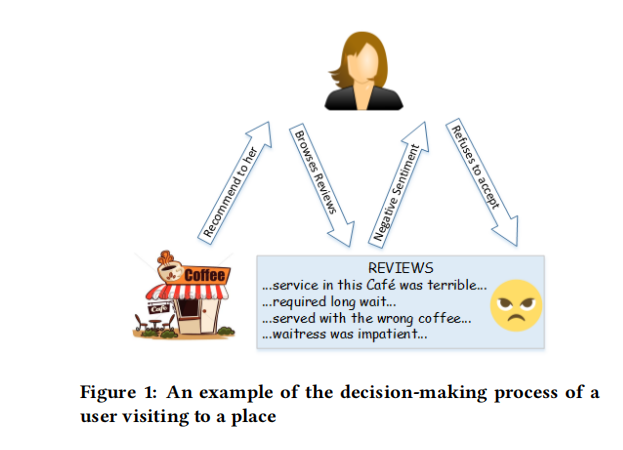
**1 介绍**

随着无线通信技术的快速发展，基于位置的社交网络（LBSNs）智能手机应用程序不断涌现，如Foursquare和Yelp，它们可以帮助用户找到附近有趣的地方。此外，用户还可以在这些LBSNs平台上发布他们的登入、主观评论和图片。大量的用户在物理世界中的行为被记录下来。如何深入分析用户的个性化推荐行为，是一项非常有价值的工作。

现有的一些方法在LBSNs中将矩阵因式分解应用于空间项目推荐。事实上，有数以百万计的空间项目，但是每个用户只访问其中的一小部分，造成了数据的稀疏性。为了缓解这一问题，一些研究将地理和社会两方面的研究融合到矩阵因式分解[4，12]中。然而，以往的方法只适合于本地用户推荐空间项目，即家乡推荐。此外，对于外地用户(如旅行者)来说，他们的大部分脚印都是在家乡地区留下的，这就加剧了数据稀疏的问题。[17]报道说，对于普通用户来说，她的家乡和外地登记记录的比率是1：0.0047。

利用空间项目的内容信息是解决数据稀疏和冷启动问题的有效方法。最近的一些研究[9，12-14，23，26，30]利用兴趣点(POIS)的分类和标签来推断外地用户的兴趣，并提出相应的POI建议。然而，这些方法没有考虑到不同地理区域的用户兴趣漂移现象[29]，即用户在离开家乡时往往有不同的兴趣。例如，在家乡很少吃海鲜的用户在海滨城市旅行时，经常会去海鲜餐厅。这种现象背后的原因是，外地的使用者更喜欢体验当地的一种可能不符合他们原有兴趣的种族。

此外，用户通常会在评论中表达他们对一些访问过的地方的主观情绪。一些文献[7,22]利用历史评论中的情感作为学习用户兴趣的明确特征，但很少有研究考虑到项目的人群情绪（即所有评论的整体情绪倾向）。事实上，在用户决定访问某个地方之前，他们通常会参考该地方的相关评论。我们认为，即使空间项目本身符合用户的兴趣，向用户推荐具有负面情绪的空间项目也是不合适的。图1示出了访问一个地方的用户的决策过程的示例。每个项目的人群情绪可以通过从所有评论中收集到的情绪来获得。



考虑到用户兴趣漂移和人群情绪的影响，我们提出了一种名为LSARS的潜在概率生成模型，用于模拟用户在家乡和城外区域的登记行为的决策过程。LSARS能够适应用户在不同区域的动态兴趣和人群对每个空间项目的情感。

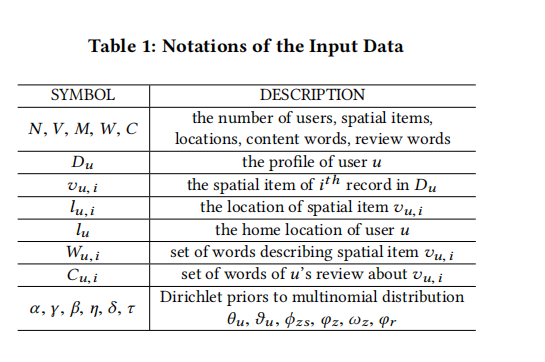
具体而言，我们首先将地理空间划分为一些子区域，在这些子区域中可以通过一组主题推断出个人的兴趣。由于区域和主题是相互依存的，因此LSARS将地理聚类和主题建模(即潜在Dirichlet分配(LDA)[2])结合为一个统一的过程。为了进一步缓解用户行为的数据稀疏性，LSARS整合了人群的偏好。为了解决用户兴趣分配问题，LSARS将地方特色与用户兴趣结合起来。具体而言，当用户u位于r区(即家乡或外地)时，LSARS会向用户推荐既受欢迎又能满足用户兴趣的空间项目。此外，人群对空间项目的感知性强，对其他用户的行为也有很强的依赖性，因此LSARS采用了一种基于情感的LDA模型。此外，LSARS利用托布勒的第一定律[15]来模拟用户的利益，即如果两个区域在地理上接近，那么用户在这两个区域的利益应该是类似。据我们所知，这是第一项通过调整用户兴趣漂移和空间项目的人群情绪来模拟用户在家乡和城外地区的登记活动的工作。在LSARS中构建了两个相应的组件：流行度感知用户移动性（PUM）和感知用户兴趣（SUI）。为了证明LSARS的潜在经济性，我们将其应用于两种应用场景：空间项目推荐和目标用户发现。在每种情景中，我们都会考虑家乡和城外案例。 主要贡献可归纳如下：

* **情绪影响**：我们将公众评论中的情感建模为用户兴趣。在用户决定登录到某个地方之前，他们不仅要检查该站点的内容信息，而且还要支付更多的相应的用户评论。
* **算法和模型**：我们提出了一种名为LSARS的潜类概率生成模型，它可以通过统一考虑用户偏好，地理影响，内容效应和情感影响来准确捕捉用户的登记行为。
* **经验评价**：我们通过使用来自LBSN平台的两个大规模数据集，进行了大量实验，以评估所提出的LSARS模型在空间项目推荐和目标用户发现的两个应用中的性能，结果表明我们的方法优于最先进的技术基线

本文其余部分按以下方式组织。第二节介绍了本文的基本原理和问题公式。第三节描述了所提出的lsars模型，并给出了相应的推理算法。第四节分别阐述了LSARS在空间项目推荐和目标用户发现中的两个应用。实验结果见第五节。第六节对相关工作进行了回顾，第七部分总结了本文的研究内容。

**2 相关工作与问题表述**

在这一部分中，我们使用了相关的数据结构和符号，并对问题进行了描述。为了便于表示，表1列出了输入数据的符号。



定义1。**(空间项目)**空间项目是与地理位置相关的项目(例如餐馆或电影院)。在我们的模型中，空间项具有三个属性：标识符，位置和内容。我们使用v来表示空间项目标识符，并使用lv来表示其对应的地理位置的经度和纬度坐标。此外，空间项目还包括文本语义信息，如类别和标记。我们使用符号Wv来表示描述空间项v的一组单词。

定义2.**（登记活动）**用户的登记活动由五元组（u，v，lv，Wv，Cv）表示，表示用户u访问具有内容描述Wv和用户的空间项v 点评位置lv的Cv。

定义3.**（用户归属位置）**给定用户u，我们将lu表示为用户所在的用户归属位置。然而，很难直接获得用户的家乡位置。我们，我们使用[17]开发的方法，这样我们就把用户经常登入的空间项目作为她的家所在地。

定义4.**（用户配置文件）**对于每个用户u，我们创建一个用户配置文件Du，它是一组与u关联的用户登记活动。 实际上，数据集D是用户简档的集合，D = {Du：u∈U}。

给定数据集D，第一个目标是为家乡用户和外地用户提供空间项目建议。此外，空间项目的所有者也希望发现潜在的用户。我们的第二个目标是向每个空间项目的所有者推荐家乡和外地的用户。因此，我们用一个统一的方法，考虑了两个不同的情景，提出了我们的问题如下。

问题1.**（空间项目推荐）**给定用户的登记数据集D和用户u及其当前位置l，我们的目标是推荐您可能感兴趣的空间项目列表（即查询 是q =（u，l））。给定距离阈值d，如果用户的当前位置和家庭位置之间的距离（即，| l-lu |）大于d，则处于城外推荐。 否则，这是一个家乡的建议。

问题2.**（目标用户发现）**给定lv处的空间项v，我们的目标是发现可能偏爱空间项v的目标用户列表。给定距离阈值d，如果空间项目的位置与目标用户的归属位置之间的距离（即，| lv-lu |）大于d，则变为城外用户发现。 否则，它是一个家乡用户发现。

与[6,16]一致，我们在本文中设置了d = 100km，这需要至少一个小时的时间来驱动，因为大约100km的距离是人类日常体育活动的典型范围，

**3 LSARS模型**

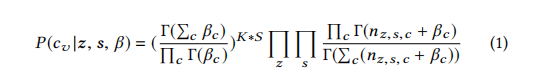
**3.1 模型的描述**

为了模拟在家乡和城外地区访问空间项目的用户的决策过程，我们提出了一个名为LSARS的联合概率生成模型，它假定用户的登记行为受到影响受以下因素影响：地理因素，项目内容效应和用户评论效果。表1显示了LSARS的图形表示，表2列出了模型的相关符号。在LSARS中，用户的登记记录被建模为观察到的随机变量，而主题，区域和情绪被认为是潜在的随机变量，分别表示为z，r和s。具体来说，LSARS包括两个组成部分：情感感知用户兴趣和流行度感知用户移动性。

**情感感知用户兴趣模型**. 直观地，用户最终基于以下假设选择空间项：(1)符合个人兴趣的空间项目内容；(2)获得最积极评价的项目。

受早期用户兴趣建模工作的启发[10,28]，LSARS采用潜在语义主题来表征用户的兴趣。具体而言，我们根据签入空间项的内容(例如标签和类别)推断用户对一组主题的兴趣分布，表示为θu。

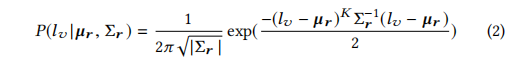
我们还对用户评论进行建模，以推断他们的偏好。从技术上讲，我们模型中的每个主题z不仅与项目内容的词ψz上的多项分布相关联，而且还与用户重新视图的词φzs上的多项分布相关。由于模型参数ϕ和φ是由所有用户共享的，为了获取项目的人群情绪，LSARS从该项目的所有评论中积累了情感。根据[11]中的工作，我们将用户评论与潜在情感集成到主题发现过程中，并在情感词ωz上采用多项分布。评论词cv的特征是P（cv | z，s，β）如下：



其中Γ(.)伽马函数。为了避免过度拟合，我们在多项式分布φzs上放置一个由β参数化的Dirichlet先验。类似地，θu，θu，ωz，φr，ψz的先验强加参数α，γ，δ，τ，η

请注意，在典型的主题模型（如[2]）中，文档包含多个主题，每个单词都有一个潜在的主题标签，适用于处理长文本。 但是，对于短文本Wv或Cv，文档更可能只包含一个主题。因此，在LSARS中，Wv或Cv中的所有单词都被赋予一个主题z。

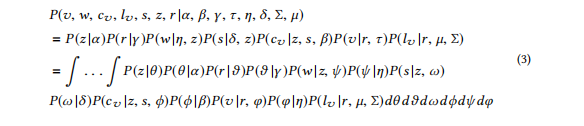
**位置感知用户移动建模**. 用户更有可能访问一些地理区域内的一些空间项目。我们将地理空间划分为R区域，并利用区域上的多项式分布ϑu对u的空间参数进行建模。根据[30]和[13]，我们对每个区域r采用高斯分布，空间项目v的位置用LV∼N(r，Σr)描述如下：



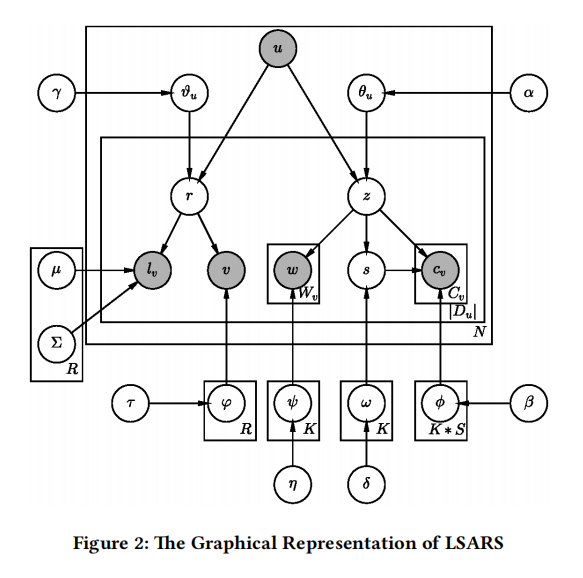
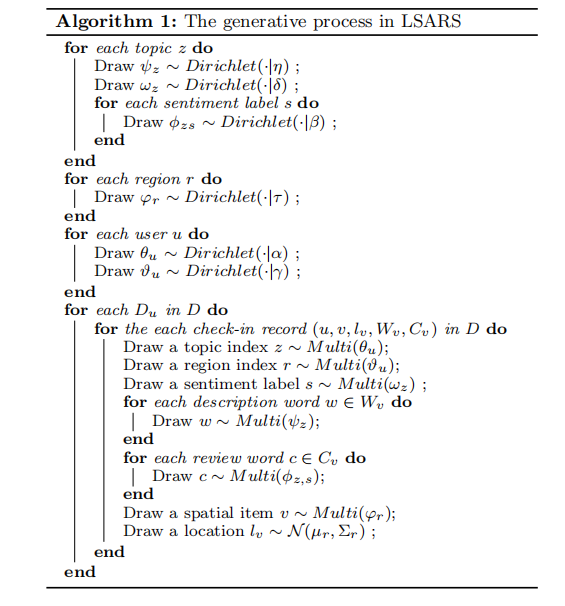
其中，Σr分别表示均值向量和协方差矩阵。

受欢迎程度对用户的入住活动也有很大的影响，特别是当用户在城外地区时。具体而言，用户的签入决策受到空间项的流行程度的强烈影响。我们使用多项分布φr来模拟区域级空间项的流行度。这是一个关键的模型设计，既支持家乡和外地的建议.

最后，我们得到了方程(3)中观测变量和隐变量的联合分布。



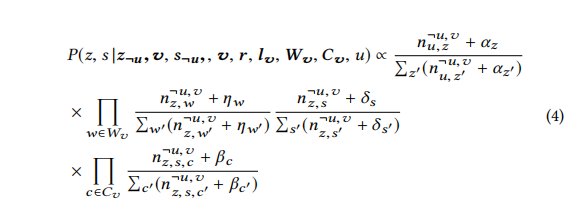
算法1列出了lsars模型的概率生成过程，并在图2中给出了lsars的图形表示。



**3.2模型推理**

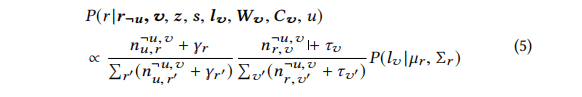
给出了超参数α，β，γ，η，δ，τ和观测值v、LV、WV和Cv，我们的目的是推断潜在变量θ，ϑ，ϕ，ψ，φ和ω，以及参数和Σ。对于θ，θ，φ，ψ，φ和ω，我们对每个用户使用潜在主题z，潜在情感s和潜在区域r的后验分布，而不是明确地将它们表示为要估计的参数，因为它是难以处理的 我们采用马尔可夫链蒙特卡罗（MCMC）方法绘制样本，如[18,19]所述。

在MCMC中，构造一个马尔可夫链以收敛于目标分布，然后从该马尔可夫链中提取样本。链的每一种状态都是对被采样变量的值分配，状态之间的二重转换规则需要不同的采样器,采用美国吉布斯抽样法。首先，需要估计z和s：p(z，s|，v，，v，r，LV，WV，CV，u)的条件概率。在方程(3)中给出了潜在变量和观测变量的联合概率分布，并利用贝叶斯链规则得到了条件概率如下：



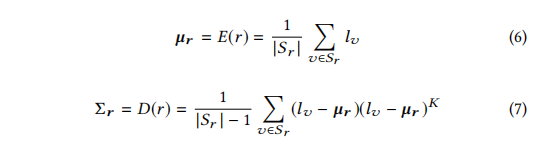
其中nu，z是从用户u中抽取潜在主题z的次数；NZ，w是单词w从主题z生成的次数；NZ，s是情感s从主题z生成的次数；NZ，s，c是由带有情感标签s的主题z生成单词cv的次数。e数n\_u，v带有上标u，v表示不包括当前实例的数量。带有上标¬uv的数字n¬uv表示不包括当前实例的数量。

我们根据下列后验概率对区域r进行采样：



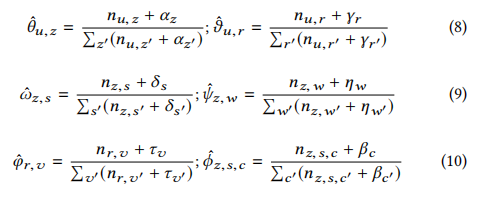
其中nu，r是区域r从用户u中采样的次数，nr，v是空间项v由区域r生成的次数。

每迭代一次，参数r和Σr按方程(6)和(7)更新。



其中Sr表示分配给潜在区域r的空间项集。

**推理框架**. 为了简单起见，我们在实现中将超参数x为α=50/K，γ=50/R，β=η=τ=δ=0.01。利用k-均值聚类算法初始化潜在地理区域r，然后随机初始化主题z和情感的赋值以进行签入记录，重复1600次。最后，后验样本可以通过检查检入记录的z和r赋值来估计参数。然后估计ϑ，θ，φ，ϕ，ψ和ω参数如下：



**3.3 时间复杂度**

分析了推理框架的时间复杂度。假设整个过程运行i迭代。在每次迭代中，所有用户的签入记录都会被扫描。对于每一次签入记录，都需要O(Ks)运算来计算潜在主题和潜在情绪的后验分布，并需要O(R)运算来计算后验区域，取样潜伏区的分布。整个时间复杂度为

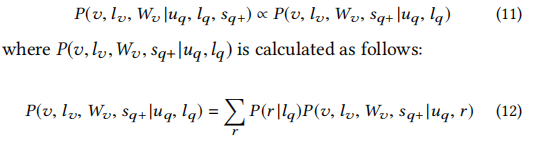
O(I(KS + R))。

**4 使用LSARS的应用程序**

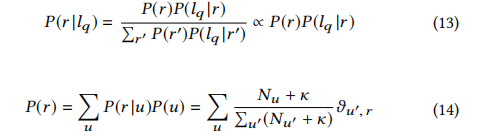
在本节中，我们将LSARS部署到两个应用程序：空间项推荐和目标用户发现。

**4.1 空间项目推荐**

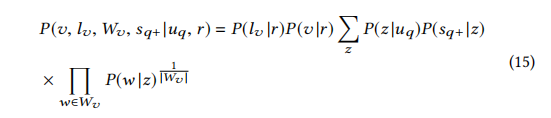
给定查询用户uq和位置lq，即q =（uq，lq），任务是计算用户uq检查每个空间项v的概率，然后向用户uq返回概率更高的top-k项。具体而言，用户uq访问空间项目v的概率由方程(11)计算，其中我们将正情绪表示为sq+。



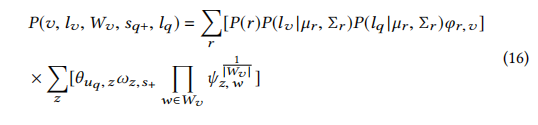
其中P(r| lq)表示给定用户u访问区域r的概率，并根据Bayes准则用方程(13)计算。可以通过等式（14）估计潜在区域r的先验概率。



其中Nu表示用户u生成的签入数。为了避免过度拟合，我们引入Dirichlet先验参数κ作为伪计数. 我们采用几何平均作为主题z生成词集WV的概率。

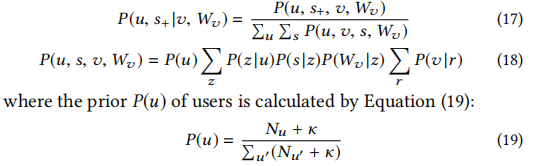


基于方程(11)-(15)，我们提出了公式(16)来推断每个空间项目的得分，并为用户提供顶k项。



**4.2 发现目标用户**

我们应用我们的模型来发现空间项目V的潜在用户。给定一个空间项目v及其描述wv，任务是对项目v可能有积极情绪(表示为s)的用户和顶级用户。具体而言，用户u访问空间项目v的概率是通过考虑项目的地理属性和情感属性来计算的：

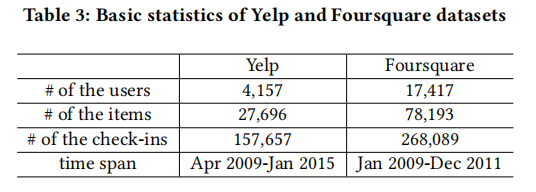


**5 实验**

在这一部分中，我们首先描述了实验的结果，然后给出了实验结果。

**5.1 实验集**

我们的实验是在两个真实世界的数据集上进行的：Yelp和Foursquare.e这两个数据集的详细情况见表3。



**Yelp.** Yelp的挑战数据集1包含4157个用户，这些用户至少有20个评论，有来自四个城市的27696个空间项目。每个登记记录都存储为用户ID，项目ID，项目位置，项目内容，项目审查和登记日期。请注意，此数据集不包含确切的签入时间，只提供粗略的签入日期(例如“2010-01-01”)。

**Foursquare.** Foursquare数据集2包含17417个用户的登记记录，这些用户分别在两个城市（即纽约和洛杉矶）发布至少6条评论。每个登记记录都存储为用户ID，项目ID，项目位置，项目内容，项目审查和登记时间。

**5.2 比较方法**

我们将LSARS与以下具有良好调谐参数的最先进方法进行了比较.

**CKNN:** CKNN[1]将用户的活动投影到类别空间，并使用加权类别层次结构对用户首选项进行建模。当收到查询时，CKNN检索查询区域中的所有用户和项，并制定用户项矩阵。它应用一种基于用户的协作性绑定(CF)方法来预测查询用户对未访问项的评分。两个用户之间的相似性是根据他们在类别层次结构中的权重来计算的。

**UPS-CF：**Ups-CF[6]是一种针对外地用户的协作推荐框架，它考虑了用户偏好、社会化程度和地理位置邻近性。具体而言，UPS-CF根据其朋友和类似用户的签入记录向目标用户推荐空间项目。

**CAPRF：**CAPRF [7]是一个统一的POI推荐框架，它集成了情感指示，用户兴趣和POI属性，其中提出了一种新的情感增强加权方案来结合个人情感信息。

**GCF：**GCF[24]是一种结合地理信息的协同决策模型。

**LCA-LDA：**LCA-LDA[28]是一种支持外城用户空间项目推荐的位置感知推荐模型。LCA-LDA通过利用共同访问模式和空间项目的内容来考虑每个城市的个人兴趣和当地偏好。 与我们的LSARS相比，LCA-LDA无法模拟地理影响和人群情绪。

**JIM：**JIM [30]是一种联合概率生成模型，通过整合时间效应，内容效应，地理影响和口碑效应等因素来模拟用户的登记行为，特别是对于外地用户。与我们的lsars相比，Jim也没有考虑到人群的情绪，而是融入了时间等。

**5.3评价方法和计量**

在这一部分中，我们介绍了lsars模型的两种应用的评价方法和指标。

*5.3.1空间项目推荐的评估。*由于LSARS是为本地和外地推荐而设计的，我们分别在这两种情况下评估了空间项目推荐的推荐效果。为了确定用户的活动是否发生在家乡地区，我们测量了用户的家庭位置lu与访问项目位置LV之间的位置距离。如果距离（即，| lu-lv |）大于d，则认为用户处于城外区域。 根据以前的研究[20,25]，我们选择d为100km。