基于用户签到行为的兴趣点推荐

任星怡 宋美娜 宋俊德

(北京邮电大学计算机学院信息网络工程研究中心教育部重点实验室 北京 100876)

摘 要 随着大数据技术的快速发展,推荐系统成为大数据领域里的一个重要的研究方向.随着基于位置社交网络(Location-Based Social Networks,LBSN)的快速发展,兴趣点(Point-Of-Interest,POI)推荐成为一个重要的研究热点,帮助人们发现有趣的并吸引人的位置,特别是当用户在异地旅行的时候.由于用户的签到行为具有高稀疏性,为兴趣点推荐带来很大的挑战.为处理用户签到数据的稀疏性问题,越来越多的研究结合地理影响、时间效应、社会相关性、内容信息和流行度影响这些方面的因素为提高兴趣点推荐的性能.然而,目前的研究缺乏一种综合分析上述所有因素共同作用的方法来处理兴趣点的数据稀疏问题,特别是异地推荐场景被目前大多数研究工作所忽略.针对以上所述的挑战,文中提出一种联合概率生成模型,称为 GTSCP,模拟用户签到行为的决策过程,该模型有效地融合上述因素来处理数据稀疏性,特别是异地推荐场景.文章所提的兴趣点推荐方法包含离线模型和在线推荐两个部分.文中所提的 GTSCP 联合模型支持本地和异地两种推荐场景.文章在多个真实 LBSNs 的大规模签到数据集上进行实验,结果表明该算法相比其它先进的兴趣点推荐算法具有更好的推荐效果.

关键词 基于位置的社交网络;兴趣点推荐;概率生成模型;用户签到行为;联合模型

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2017.00028

Point-of-Interest Recommendation Based on the User Check-in Behavior

REN Xing-Yi SONG Mei-Na SONG Jun-De

(*Engineering Research Center of Information Networks of Ministry of Education, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing* 100876)

Abstract With the rapid development of big data technology, recommendation systems become an important research direction in the field of big data. With the rapid development of location-based networks, point-of-interest (POI) recommendation has become an important way to help people discover interesting and attractive locations, especially when users travel out of town. However, because users check-in interaction is highly sparse, which brings a big challenge for POI recommendation. To tackle this challenge, a growing line of researches have exploited the geographical influence, temporal effect, social correlation, content information and popularity impact. However, current research lacks an integrated analysis of the joint effect of the above factors to deal with the problem of data sparsity, especially in the out-of-town recommendation scenario which has been ignored by most existing work. In light of the above, we propose a joint probabilistic generative model called GTSCP to imitate user check-in activities in a process of setting decision, which integrates the above factors to overcome the data sparsity effectively, especially for out-of-town users. The POI recommendation method consists of offline modeling and online recommendation. The GTSCP supports two recommendation scenarios in a joint model, i.e., home-town recommendation and out-of-town recommendation. We implement experiments

收稿日期:2016-03-28;在线出版日期:2016-09-19.本课题得到中国国家科技重点支撑项目(2014BAH26F02)资助.**任星怡**,女,1983 年生,博士研究生,主要研究领域为推荐系统、数据挖掘、大数据.E-mail:xyren@bupt.edu.cn.**宋美娜**,女,1974 年生,博士,教授,主要研究领域为服务计算、云计算、超大规模信息服务系统.E-mail:mnsong@bupt.edu.cn.**宋俊德**,男,1938 年生,博士,教授,主要研究领域为服务科学与工程、云计算、大数据、物联网.

on real LBSNs check-in datasets, experimental results show that GTSCP achieves significantly superior recommendation quality compared to other state-of-the-art POI recommendation techniques.

Keywords　location-based social network; point-of-interest recommendation; probabilistic generative model; user check-in activities; joint model

1 引　言

随着 Web 2.0 的快速发展，无线通信与位置采集技术促生很多基于位置的社交网络，例如，Foursquare、Yelp、Facebook Places 等. 用户能够以“签到”的形式发布他们的地理标签信息和物理位置，并对已访问的兴趣点（例如，商场、餐厅、博物馆、娱乐场所、酒店等）与朋友分享他们的体验与经验. 关于基于位置的社交网络的个性化推荐，利用用户签到数据是至关重要的，帮助用户探索新的区域与发现新的兴趣点，促进广告商对目标用户推送移动广告，从而使基于位置的社交网络更具有吸引力.

不同于传统推荐任务，兴趣点推荐是一个基于上下文信息的位置感知的个性化推荐. 通过下面的场景进行详细描述. 例如，星怡居住在中国北京，早晨她通常会去她家附近的庆丰包子铺吃早餐，中午她通常会去她工作附近的东北风味餐馆吃午餐，晚上在回家之前她通常会约她的朋友去酒吧娱乐，周末她有时会与家人去朝阳公园散步或者去西单购物. 现在，如果星怡想去杭州度假，那么在旅行中什么样的兴趣点她会感兴趣呢？这样的兴趣点推荐一定是基于上下文信息的位置感知的个性化推荐.

相比传统推荐系统的发展，兴趣点推荐系统的发展更加复杂. 兴趣点推荐面临一些新的挑战. 第一，用户-POI 的签到矩阵是高稀疏的，因为在基于位置的社交网络中用户访问兴趣点占有非常小的比例，因此兴趣点推荐面临数据稀疏性问题. 第二，随着不同的时间与地理位置的变化，用户的兴趣是动态变化的. 第三，兴趣点推荐包含不同类型的上下文信息，例如兴趣点的地理坐标、签到的时间、用户的社会关系、兴趣点的分类以及兴趣点的流行度等，与传统推荐不同的是和兴趣点相关的文本信息是不完整且模糊的. 用户-POI 矩阵极端稀疏是兴趣点推荐需要解决的一个最重要的问题. 在基于位置的社交网络中有数百万的 POIs，但用户只能访问有限数量的 POIs. 基于位置的社交网站数据显示用户在异地生成的签到记录只占用户在本地生成的签到记录的 0.47%[1]，这个观察凸显了数据稀疏问题[2-3].

协同过滤（Collaborative Filtering，CF）是推荐系统最流行的方法[4]. 目前大量的研究工作[1-2,5-7]存储用户的历史签到数据到用户-POI 矩阵当中. 基于用户-POI 矩阵，利用基于 CF 的方法推断用户对每个未访问的 POI 的偏好. CF 的核心思想是根据相似的用户所提供的线索进行推荐. 由于用户移动位置的原因，大多数这些相似的用户更有可能与目标用户居住在同一地区. 考虑到这种推荐方法是通过相似用户已访问的 POIs 推荐给目标用户，大多数这些推荐给目标用户的 POIs 的物理位置位于目标用户的本地. 因此，对于异地推荐场景，这些基于 CF 的方法不能直接应用到兴趣点推荐中.

此外，兴趣点推荐不同于传统的 GPS 轨迹[8-9]，用户的时间顺序签到记录是低采样率的，用户移动信息的细节会被丢失. 在兴趣点推荐中任何两个连续签到的 POI 点之间的空间距离通常在 1 公里的范围内，然而在 GPS 轨迹中任何两个连续记录的 GPS 点之间的空间距离通常是 5 m～10 m. 而且，在兴趣点推荐中的两个连续签到的时间间隔远远大于在 GPS 轨迹中的两个连续记录点的时间间隔. 因此，现有的序列模式挖掘方法如马尔科夫链模型不能直接应用于用户签到数据稀疏的兴趣点推荐中.

随着基于位置的服务应用日益流行，关于兴趣点推荐，在地理、时间、社会、内容与流行度等方面的因素中，基于位置的社交网络为研究人类的移动行为提供一个前所未有的机会. 虽然有一些研究利用上述一个或几个因素来改善兴趣点推荐的效果，但目前已有的研究工作缺乏一种综合分析上述所有因素共同作用的方法来处理兴趣点的数据稀疏问题，特别是异地推荐场景被目前大多数研究工作所忽略. 针对以上所述的挑战，我们提出一种联合概率生成模型模拟用户签到行为的决策过程，称为 GTSCP，该模型有效地融合地理影响、时间效应、社会关系、内容信息以及流行度影响，并构建用户签到

行为的联合效应模型.

我们利用地理与流行度影响建模用户的移动模式.首先,我们通过用户历史已访问的 POIs 的位置分布或者用户的当前位置推断用户的活动范围.特别是,我们将地理空间划分成几个区域,并计算个人用户可能访问的每个区域的概率.然后,我们通过 POI 与区域中心的距离以及此区域的公众签到活动(即考虑地理影响和 POI 的区域级别流行度),推算每个区域的 POI 生成概率.其次,通过整合区域级别流行度影响(即群众的智慧),从而我们所提的 GTSCP 模型可以解决用户兴趣漂移问题.跨地理区域的数据表明用户在一个区域的兴趣推测不能总是被应用于另一个区域的推测.例如,一个用户生活在中国北京,他从来没有赌博行为,但是当他去澳门或者拉斯维加斯旅行时,他是最有可能去赌场的.为了进一步处理数据稀疏问题,GTSCP 融合了空间索引结构,即树型结构的空间金字塔,这是第一个划分空间项目位置到不同层次不同大小的空间网络中的模型.

我们利用兴趣点的内容信息、时间效应和社会关系建模用户的兴趣模式.我们通过用户已访问的 POIs 的内容信息推断用户的个人兴趣.因此,我们所提的 GTSCP 模型缓解了数据稀疏,特别是用户到异地旅行的场景.不幸的是,基于位置的社交网络当前的数据显示[10],大约所有 POIs 的 30% 缺乏任何有意义的文本描述.我们利用用户活动内容的时间效应来处理此问题,因为签到时间为 POIs 的内容信息提供了重要的线索.用户的社会关系也会影响用户签到行为.我们同时利用用户的社会关系处理数据稀疏问题.

我们所提的兴趣点推荐方法包含离线模型和在线推荐两个部分.离线模型此部分的目的是利用用户的签到行为并捕捉 POIs 的同现模式,学习每个用户的兴趣和每个地区的当地偏好.在线推荐部分需要查询用户、查询时间以及查询区域作为输入,并自动结合已学习的查询用户的兴趣和区域的当地偏好生成 top-k 推荐列表.我们利用一个可伸缩的查询处理技术通过扩展的阈值算法(Threshold Algorithm,TA)加速在线推荐过程.如图 1 所示.

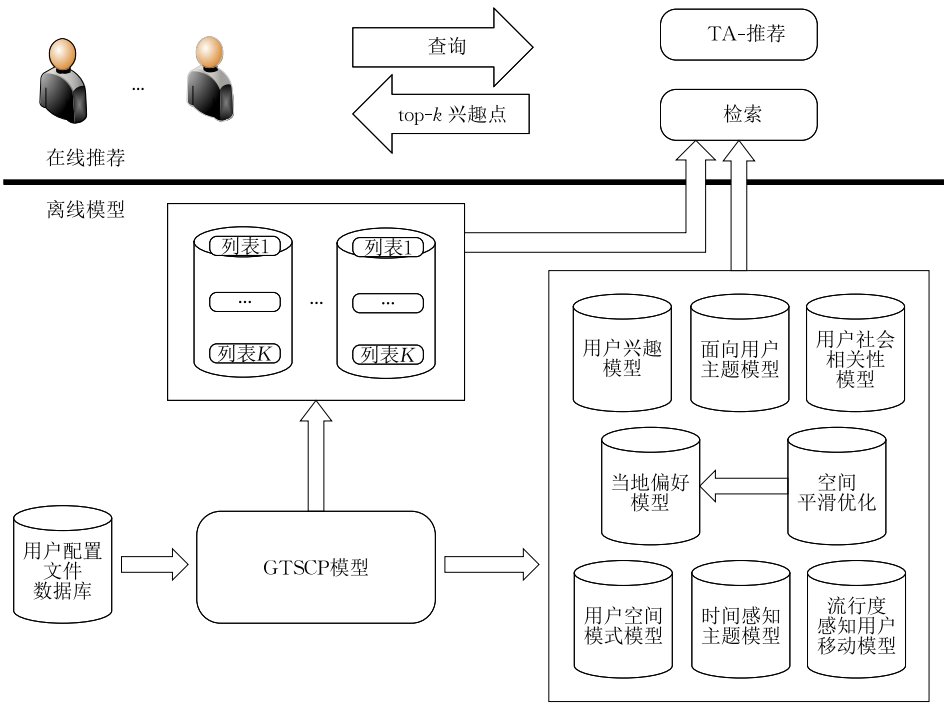


图 1 本文兴趣点推荐体系结构框架图

为了显示 GTSCP 的适用性,我们研究它是如何以一种联合的方式支持两种推荐场景:(1)本地推荐的目的是当目标用户在本地时需满足用户日常生活的信息需求;(2)异地推荐的目的是当目标用户在异地旅行时需满足用户外出旅行的信息需求,尤其是当目标用户在不熟悉的区域的时候.值得一提的是,这两个推荐场景应该是定位的、时间感知的、个性化的兴趣点推荐,目标用户在不同的位置与时间时,会推荐给同一目标用户不同的 POIs 排名列表.

给定一个目标查询用户 u_q、目标查询用户的当前位置 l_q 与时间 t_q,关于每个兴趣点,在线推荐部

分将计算得分排名，通过离线模型 GTSCP 来学习，自动结合用户兴趣模型、当地偏好模型、面向用户主题模型、用户社会相关性模型、用户空间模式模型、时间感知主题模型和流行度感知用户移动模型. 为了加快在线推荐的过程，关于 top-k 推荐，我们提出一个扩展查询过程技术，从在线评分计算中分离出离线评分计算从而减少查询时间. 特别是，在一个给定的区域划分兴趣点. 如图 1 所示，关于每个地理位置，我们存储此处地理位置的 K 个兴趣点列表. 在查询时间 t_q，我们检索目标查询用户 u_q 的当前位置 l_q 的所有兴趣点，然后利用扩展阈值算法(TA)通过结合 K 个候选项的分数来得出 top-k 个兴趣点.

本文中我们的研究工作的主要贡献总结如下：

(1) 我们提出一个联合概率生成模型，称为 GTSCP，是第 1 个同时融合地理影响、时间效应、社会相关性、内容信息和流行度影响的研究，并构建用户签到行为的联合效应模型.

(2) 我们所提的 GTSCP 是以一种联合模型的方式支持两种推荐场景，即本地推荐和异地推荐.

(3) 我们利用地理相关性，通过一个设计良好的空间索引结构即空间金字塔，对当地偏好进行平滑优化，进一步缓解数据稀疏问题.

(4) 我们所提的兴趣点推荐方法包括离线模型和在线推荐两个部分，然后利用一个可扩展的查询过程技术阈值算法加速在线推荐过程.

(5) 我们在一个真实 LBSN 的大规模签到数据集上进行大量的实验，评估我们所提的 GTSCP 的性能，实验结果清晰地表明 GTSCP 在本地和异地两种推荐场景下的推荐任务优于其它先进的兴趣点推荐.

本文第 2 节介绍 LBSN 中兴趣点推荐技术的相关工作；第 3 节提出 GTSCP 模型；第 4 节部署 GTSCP 到兴趣点推荐中；第 5 节介绍实验的方案设计与性能对比，验证该算法的有效性；第 6 节总结本文并探讨将来的研究工作.

2　相关工作

兴趣点推荐被认为是推荐系统领域中的一个重要任务，也称为位置推荐[11]. 最近，关于在 LBSNs 上的大规模用户签到行为记录，当前许多研究试图利用并融合地理影响、社会影响、时间效应、内容信息与流行度影响来提高兴趣点推荐的效果.

2.1　兴趣点推荐

随着基于位置社交网络的快速发展，兴趣点推荐为人们提供了更好的基于位置的服务，受到学术界和工业界的广泛关注. 基于记忆的协同过滤技术，如基于用户的协同过滤和基于项目的协同过滤被应用到兴趣点推荐中. 由于兴趣点推荐的签到数据具有高稀疏性，基于记忆的协同过滤方法很容易遭受数据稀疏问题. 因此，应用基于记忆的协同过滤技术不能有效地进行兴趣点推荐. 基于模型的协同过滤技术同样被应用到兴趣点推荐中，但是在这些研究的工作中，几乎都只考虑了显示反馈. 最近，一些研究工作开始考虑签到数据的隐式反馈. Lian 等人[6]提出一种结合地理影响的加权矩阵分解方法. 另外，Cao 等人[12]通过随机游走方法计算元路径特征值，以度量实例路径中的首尾节点间关联度的方式，利用监督学习方法获得各个特征的权值，计算特定用户在兴趣点的签到概率. Jiang 等人[13]提出一种基于模型的协同过滤的作者主题模型，关于社会用户该模型全面促进了兴趣点推荐的性能. 最新的兴趣点推荐开始应用多媒体数据. Jiang 等人[14]利用旅游指南和社区提供的照片以及与这些照片相关的异构元数据(如标签、地理位置和日期)，提出一种个性化旅行序列兴趣点推荐. 兴趣点推荐同样需要解决推荐系统的冷启动问题，一些研究工作[13-17]很大程度地解决了兴趣点推荐的冷启动问题.

2.2　基于地理社会影响的兴趣点推荐

地理社会信息表明，人们倾向于探索用户或者用户朋友已访问的 POI 邻近的 POIs[12]. Ference 等人[2]设计了一个协同推荐框架，通过用户朋友和相似用户生成的活动记录，构建一种混合的方式. Lian 等人[6]结合由地理影响所导致的空间聚类现象到一个加权矩阵分解框架中处理矩阵稀疏的问题. 最近兴趣点推荐的一些研究工作开始对社会媒体用户的信息进行挖掘. Qian 等人[15]利用 3 种社会因素即个人兴趣、人际关系之间的兴趣相似性和人际关系之间的影响，并基于概率矩阵分解将这 3 种社会因素融合到一个统一的个性化推荐模型中. Zhao 等人[16]充分利用移动用户位置的敏感特点进行评分预测，挖掘用户评分与用户-项目地理位置距离之间的相关性以及用户评分之间的差异与用户-用户地理位置距离之间的相关性. Zhao 等人[17]利用社会用户评分行为提出一种用户评分预测方法，并将四种因素即用户个人兴趣、人际交往之间的兴趣相似性、人际交往之间的评分行为和人际交往之间的评

分行为相似性融合扩散到一个统一的矩阵分解框架中.最近许多研究[18-20]表明用户签到行为与地理距离以及社会关系之间存在很强的相关性,所以大多数当前的兴趣点推荐工作主要集中利用地理社会影响提高推荐的精确度.Ye 等人[18]研究位置之间的地理影响并提出一个结合用户偏好、社会影响和地理影响的系统.关于位置推荐,Cheng 等人[19]通过结合一个多中心高斯分布模型、矩阵分解和社会影响来研究地理影响.

2.3 基于时间效应的兴趣点推荐

在 LBSNs 中用户签到行为的时间效应也吸引了研究人员的关注.在 LBSNs 中,兴趣点推荐具有时间序列模式和时间循环模式.Gao 等人[7]依据时间是非统一、连续的特性,研究用户签到的时间循环模式.Cheng 等人[21]通过嵌入时间序列模式,介绍了在 LBSNs 中的连续、个性化的兴趣点推荐任务.关于时间感知的兴趣点推荐,Yuan 等人[22]融合时间循环模式到一个基于用户的协同过滤框架中.

2.4 基于内容信息的兴趣点推荐

最近大多数研究人员探讨 POIs 的内容信息来缓解数据稀疏问题.关于基于位置的社交网络,Ye 等人[10]利用个体位置的显性模式和相似位置间的隐式相关性,提出一种用分类标签标注位置的语义注释工作.Jiang 等人[13]提出一种基于用户信息挖掘的协同过滤兴趣点推荐方法.Ferrari 等人[23]分析 Twitter 上的帖子,并在数据集上执行 LDA 算法提取城市模式,例如热点与人群行为.一些研究人员结合矩阵分解方法与主题模型来挖掘每个兴趣点相关的文本信息.Gao 等人[24]和 Zhao 等人[25]在兴趣点推荐中利用了 POIs 相关的内容信息与用户的情绪信息.

2.5 基于流行度影响的兴趣点推荐

POIs 的流行度反映了 POIs 所提供的服务和产品的质量.因此,关于兴趣点推荐利用 POIs 的流行度是有用的.当前大多数研究普遍认为 POIs 的流行度是用户对 POIs 的先验偏好.Ying 等人[26]对于未访问的 POIs,在完全二部图中,利用用户的先验偏好作为用户与 POIs 之间的加权边.其它一些研究工作[22,27-29]利用地理信息获取先验偏好来调整后验偏好.然而,在这些研究中,先验偏好不是个性化的用户偏好.

如上所述,这些研究工作缺乏一种综合分析上述所有因素共同作用的方法来处理兴趣点的数据稀疏问题,特别是异地推荐场景被目前大多数研究工作所忽略.我们提出一种联合概率生成模型,称为 GTSCP,融合地理影响、时间效应、社会关系、内容信息以及流行度影响,有效地解决了数据稀疏问题,特别是异地推荐场景.

3 用户签到行为联合模型

3.1 问题定义

本节定义数据结构、阐述研究问题与展示模型框图.从 LBSN 的丰富信息中提取数据结构即 POIs 上的用户历史签到数据,包括 POIs 的地理信息、POIs 上的用户签到时间,用户间的社会关系、POIs 的内容信息以及 POIs 的流行度.为了便于说明,表 1 列出用于本文的输入数据的符号.

表 1 输入数据的符号

符号	意义
u,v,l,s,t,r	用户,POI,位置,社会关系,时间段,区域
U,V	用户集合,POIs 集合
L^U,L^V	用户本地位置集合,POIs 位置集合
N,M	用户的数量,POIs 的数量
L,P	位置、POIs 位置的数量
K,W	主题的数量,唯一词的数量
R,T	区域的数量,时间段的数量
D_u	用户 u 的配置文件
W_v	描述 POI 的词集合
l_u	用户 u 的本地位置
l_v	POIv 的位置

定义 1. 兴趣点.一个兴趣点(POI)定义为一个具有唯一标识符的与地理位置有关的特定地点.

在我们的模型中,一个兴趣点有 3 个属性:标识符、位置和内容.我们使用 v 代表兴趣点的标识符,l_v 代表兴趣点的位置即经纬度坐标.请注意的是,许多不同的 POIs 会共享相同的经纬度坐标.另外,一个 POI 有很多相关的语义信息,例如分类和标签.我们使用符号 W_v 描述 POI 的词集合.然后我们使用空间金字塔划分并索引整个地理区域.粒度取决于应用的性质和从城市到街道的范围.

定义 2. 用户本地位置.根据文献[30],给定一个用户 u,我们定义用户的本地位置为用户所居住的地理位置,用符号 l_u 代表.

在我们的模型中,我们假设用户居住的地理位置是"固定"的.换言之,用户本地位置是静态的,而实时位置是动态的,即用户旅行时与其相关的地理位置是暂时的.由于隐私问题,用户居住的位置并不总是可以获取的.由于用户居住的位置不能很明确地给定,我们采用文献[31]的方法,在大多数用户签

到的范围内，我们划分 25 km×25 km 的区域作为空间金字塔的单元格.

定义 3.　签到行为. 用户签到行为由 5 个元组 (u,v,l_v,W_v,t) 组成，表明用户 u 访问 POIv，在时间 t，位置在 l_v，由 W_v 所描述.

定义 4.　用户配置文件. 关于每一个用户 u，我们创建一个用户配置文件 D_u，其是关于用户 u 的签到行为的集合. 在我们的模型中，数据集 D 由用户配置文件组成，即 $D=\{D_u:u\in U\}$.

我们所提的兴趣点推荐模型支持本地和异地推荐场景. 在我们的模型中，给定一个数据集 D 作为用户配置文件的集合，公式化我们所提的问题，并以一种统一的方式考虑了两种推荐场景.

定义 5.　主题. 给定一个词的集合 W，一个主题 z 被定义为关于 W 的多项分布，即 $\phi_z=\{\phi_{z,w}:w\in W\}$，每一个 $\phi_{z,w}$ 中，w 代表主题 z 生成词 w 的概率. 通常，一个主题是一个语义连贯的词的软集群.

问题 1.　兴趣点推荐. 给定一个用户签到行为数据集 D、一个目标查询用户 u_q、目标查询用户的当前位置 l_q 与时间 t_q，即查询 $q=(u_q,t_q,l_q)$，我们的目标是推荐给用户 POIs 的列表即目标查询用户 u_q 感兴趣的 POIs. 给定一个距离阈值 d，如果用户当前位置与其本地位置的距离大于 d，即 $|l_q-l_u|>d$，则问题变成异地推荐. 否则，问题为本地推荐.

根据相关研究[2,20]，在我们的工作中，我们设置 $d=100$ km，因为 100 km 左右的距离是人们“到达”的典型半径，是需要开车 1～2 h 的距离.

3.2　模型描述

为模仿用户在决策过程中的签到行为，我们提出一个联合概率生成模型，并融合地理、时间、社会、内容和流行度信息. GTSCP 模型的框图如图 2 所示. N,L,K,R 分别代表用户、位置、主题和区域的数量. 表 2 介绍了我们的模型参数符号.

表 2　模型参数的符号

符号	意义
θ_u	用户 u 兴趣，由主题集合上的多项分布所表达
θ'_l	位置 l 的当地偏好，由主题集合上的多项分布所表达
ς_u	用户 u 的社会相关性，由社会关系集合上的多项分布所表达
ϑ_u	用户 u 的活动范围，由区域集合上的多项分布所表达
ϕ_z	词具体对应到主题 z 上的多项分布
ψ_z	时间具体对应到主题 z 上的贝塔分布
φ_r	POIs 具体对应到区域 r 的区域级别流行度分布
μ_r	区域 r 的位置平均值
Σ_r	区域 r 的位置协方差
λ_u	特定于用户 u 混合加权：关于采样二进制变量 a 的参数
b,b'	生成 λ_u 的贝塔先验
$\alpha,\alpha',\delta,\gamma,\beta,\tau$	$\theta_u,\theta'_l,\varsigma_u,\vartheta_u,\phi_z,\psi_z,\varphi_r$ 多项分布的狄利克雷先验

在我们的方法中，输入数据，即用户的签到记录，被建模为观察随机变量，在图 2 中显示为阴影圈. 签到记录的社会、区域和主题指数被考虑为潜在随机变量，表示为 s,r,z. 特别是，GTSCP 模型由 7 个部分组成：用户兴趣模型、当地偏好模型、面向用户主题模型、用户社会相关性模型、用户空间模式模型、时间感知主题模型和流行度感知用户移动模型，将在下文分别描述. 本文提出的 GTSCP 算法是一个灵活的模型，由上述的 7 个部分组成，由于兴趣点推荐会受多种因素的影响，本文所提的模型可以灵活地增减各个因素，所以这 7 个部分模型之间的因素是相互独立的.

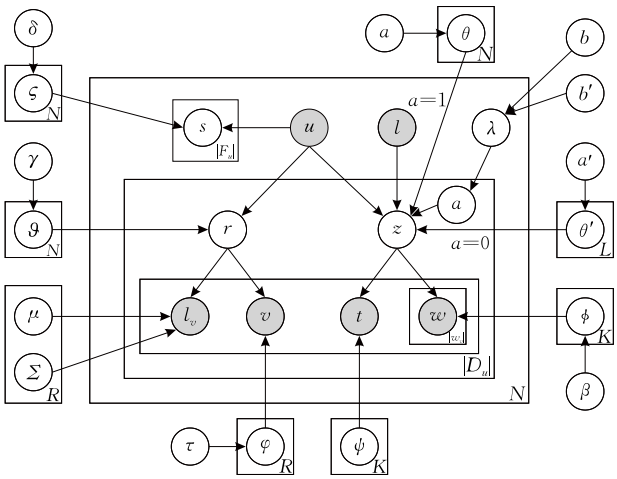


图 2　GTSCP 模型的框图

3.2.1　用户兴趣模型

关于用户兴趣模型，由以前的研究工作[3,32]所启发，我们利用潜在主题模型提取用户的兴趣解决数据稀疏问题. 我们通过主题提取用户配置文件，将每个用户签到的 POIs 的所有文本评论聚集到一个用户文档. 因此，我们得到一个大的文档集合，每个文档对应一个用户. 特别是，利用在主题集合上对应的用户签到的 POIs 的内容（即分类和标签），我们推导个体用户的兴趣分布，表示为 θ_u. 因此，用户兴趣与主题分布相关，从签到行为的主题中提取用户兴趣.

3.2.2　当地偏好模型

当用户访问一个城市，尤其是一个新的城市，他们更有可能参加那个城市受欢迎的活动以及参观当地的景点. 因此，给用户推荐兴趣点时，其他旅行者在这个城市旅行进行查询时所留下来的签到行为历史记录是有价值的资源，特别是用户前往一个陌生的城市，他们对邻近的区域几乎没有认识的经验. 位置 l 的当地偏好由 θ'_l 所表示，即主题集合上的多项

分布.这个模型可以捕捉当地的民俗风情和当地具有吸引力的地方,从而解决用户兴趣漂移问题.

3.2.3 面向用户主题模型

在大量的签到 POIs 的内容文档的采集过程中,我们利用主题模型算法是用来识别潜在主题信息.我们通过主题提取 POI 配置文件,将所有用户在相同 POI 上的文本评论聚集到一个 POI 文档.因此,我们得到一个大的文档集合,每个文档对应一个 POI.它是一个语料库的生成概率模型,基本的思想是文档被表示为潜在主题上的随机混合物,通过词分布提取主题特征.给定一组词集合 W,主题 z 被定义为 W 上的多项分布,表示为 ϕ_z.这个主题分布为面向用户主题模型.

3.2.4 用户社会相关性模型

朋友通常会分享相似的兴趣.这是一个常见的假设,源于"同质性"理论,阐明用户会连接相似的用户,通过"社会影响"理论,阐明连接的用户相比其他未连接的用户会更加具有相似性.这些同质性与社会影响的现象相互作用,他们的集体效应被称为"社会相关性".另外,用户朋友的兴趣对用户在 POIs 进行签到的决策时起到重要的作用[33].例如,北京故宫为一个兴趣点,其在北京属于旅游分类中的热门旅游景点.如果一个用户 u 评论了北京故宫,然后用户 u 就属于旅游分类的圈子中了.在旅游分类中,用户 u 的朋友用户 v 会受其影响,由于用户 u 和用户 v 之间存在着社会关系,彼此之间存在信任感并互相影响对方,则用户 v 会对北京故宫此兴趣点产生兴趣,很有可能对北京故宫进行访问并签到加以评论.F_u 代表用户 u 的朋友的集合.在分类中,用户 u 与用户 v 之间的直接或者加权的社会关系通过一个正值 $s_{u,v}\in[0,1]$来表示,s 代表社会关系.我们使用高斯分布 ς_u 来建模用户与其朋友之间的社会相关性,即 $s\sim Multi(\varsigma_u)=N(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{s},\varepsilon)$,$\boldsymbol{c}$ 是 n 维的权值向量,$\boldsymbol{s}$ 是 n 维的相似度向量,ε 是高斯模型中的方差.该建模方法可以检测与用户相似的兴趣以及与 POIs 相似的功能来处理冷启动问题.

3.2.5 用户空间模式模型

空间聚类现象表明用户最有可能访问的一些 POIs 的地理位置,并且这些 POIs 通常仅限于某些地理区域[18].在本文中,地理空间划分为 R 个区域.我们采用一个多项分布 ϑ_u 在区域上建模用户 u 的空间模型.依据文献[28,34],用户在所有 $\mathbb{R}$ 区域上遵循一个多项分布 $r\sim P(r|\eta_u)$,η_u 是用户 u 在潜在区域上的用户从属分布.每一个 POI 的位置 l_v 是从一个区域从属多元正态分布 $l_v\sim N(\boldsymbol{\mu}_r,\boldsymbol{\Sigma}_r)$中提取的.建模 POI 空间密度的一个简单方法是使用一个高斯密度函数,如下:

$$P(l_v|\boldsymbol{\mu}_r,\boldsymbol{\Sigma}_r)=\frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}_r|}}\mathrm{e}^{-\frac{1}{2}(l_v-\boldsymbol{\mu}_r)^{\mathrm{T}}\boldsymbol{\Sigma}_r^{-1}(l_v-\boldsymbol{\mu}_r)} \quad (1)$$

其中 $\boldsymbol{\mu}_r$ 和 $\boldsymbol{\Sigma}_r$ 代表平均值向量和协方差矩阵.

3.2.6 时间感知主题模型

最近,在基于位置社交网络的用户签到数据中显示,大约所有 POIs 的 30%是缺乏任何有意义的文本信息.为处理这个问题,我们利用已签到的 POIs 的内容信息与签到时间之间的相关性,挖掘用户活动内容与用户活动时间的同现模式.直观上,那些被大多数用户在相同或者相似的时间签到的 POIs 具有相同或者相似的内容和功能.因此,使用签到时间推断 POIs 的主题是有用的.技术上,在我们所提的 GTSCP 模型中,每个主题 z 不仅与词 ϕ_z 的多项分布相关而且与时间 ψ_z 的连续分布相关.这个方法使得 ϕ_z 与 ψ_z 可以在主题的发现过程中相互影响并彼此增强相关性.另外,这个方法的优点是推断个人兴趣时不仅是语义感知的而且是时间感知的.

为了在主题发现过程中融合签到时间的信息,时间分割的设计是至关重要的.现有的离散时间分解方法是把时间分割成基于小时的时间片段[7,22].然而我们的方法与其不同的是,我们处理的时间为连续的真实值,并正则化一天的时间范围为 0～1,考虑以下两点:(1)时间的本质是连续的;(2)时间离散化通常回避了选择恰当的时间间隔问题,这些时间的间隔总是太大或者太小.依据文献[35],我们采用贝塔分布来描述主题 z 的时间分布.

3.2.7 流行度感知用户移动模型

流行度在很大程度上也会影响用户签到行为.当用户到异地旅行时,尤其是一个新的区域,用户的决策很大程度受当地的口碑意见所影响,即 POIs 的流行度.我们采用多项分布 φ_r 建模区域级别的 POIs 的正则化流行度,即 $\varphi_{r,v}$ 代表区域 r 的 POIv 的正则化流行度.当地流行的 POIs 代表当地具有吸引力的地方,有可能为用户提供更好的体验.这就是为什么在同一地区内的具有相同语义的两个 POIs 的评价会不同.

3.2.8 GTSCP 模型的概率生成过程

我们将上述 7 个部分通过概率生成过程集成到 GTSCP 模型中.为了避免过度拟合,我们设置了多项分布 θ_u 的狄利克雷先验,由参数 α 所表示:

$$P(\theta_u|\alpha)=\frac{\Gamma(\Sigma_z\alpha)}{\Pi_z\Gamma(\alpha)}\prod_z\theta_{u,z}^{\alpha-1} \tag{2}$$

其中 $\Gamma(\cdot)$ 是伽马函数. 相似的，多项分布 $\theta'_l,\phi_z,\varsigma_u,\vartheta_u,\varphi_r$ 的狄利克雷先验，分别由参数 $\alpha',\beta,\delta,\gamma,\tau$ 所表示如下：

$$P(\theta'_l|\alpha')=\frac{\Gamma(\Sigma_z\alpha')}{\Pi_z\Gamma(\alpha')}\prod_z\theta'^{\alpha'-1}_{l,z} \tag{3}$$

$$P(\phi_z|\beta)=\frac{\Gamma(\Sigma_w\beta)}{\Pi_w\Gamma(\beta)}\prod_w\phi_{z,w}^{\beta-1} \tag{4}$$

$$P(\varsigma_u|\delta)=\frac{\Gamma(\Sigma_s\delta)}{\Pi_s\Gamma(\delta)}\prod_s\varsigma_{u,s}^{\delta-1} \tag{5}$$

$$P(\vartheta_u|\gamma)=\frac{\Gamma(\Sigma_r\gamma)}{\Pi_r\Gamma(\gamma)}\prod_r\vartheta_{u,r}^{\gamma-1} \tag{6}$$

$$P(\varphi_r|\tau)=\frac{\Gamma(\Sigma_v\tau)}{\Pi_v\Gamma(\tau)}\prod_v\varphi_{r,v}^{\tau-1} \tag{7}$$

我们在算法 1 中正式地描述了 GTSCP 模型算法的概率生成过程.

算法 1.　GTSCP 模型的概率生成过程.

FOR 每个主题 z DO
　抽取 $\phi_z\sim Dirichlet(\cdot|\beta)$；
END
FOR 每个区域 r DO
　抽取 $\varphi_r\sim Dirichlet(\cdot|\tau)$；
END
FOR 每个社会相关性 s DO
　抽取 $\varsigma_u\sim Dirichlet(\cdot|\delta)$；
END
FOR 每个用户 u DO
　抽取 $\theta_u\sim Dirichlet(\cdot|\alpha)$；
　抽取 $\vartheta_u\sim Dirichlet(\cdot|\gamma)$；
END
FOR 每个位置 l DO
　抽取 $\theta'_l\sim Dirichlet(\cdot|\alpha')$；
END
FOR 在 D 集合中的每个文档 D_u DO
　FOR 每个签到$(u,v,l_v,W_v,t)\in D_u$ DO
　　抛一枚硬币 a 通过伯努利$(a)\sim beta(b,b')$
　　IF $a=1$ THEN
　　抽取一个主题指数 $z\sim Multi(\theta_u)$；
　　END
　　IF $a=0$ THEN
　　抽取一个主题指数 $z\sim Multi(\theta'_l)$；
　　END
　　抽取一个时间指数 $t\sim Beta(\psi_{z,1},\psi_{z,2})$
　　FOR 每个令牌 $w\in W_v$ DO
　　抽取 $w\sim Multi(\phi_z)$；
　　END
　　抽取一个社会相关性指数 $s\sim Multi(\varsigma_u)$；
　　抽取一个区域指数 $r\sim Multi(\vartheta_u)$；
　　抽取一个 POI 指数 $v\sim Multi(\varphi_r)$；
　　抽取一个位置 $l_v\sim N(\boldsymbol{\mu}_r,\boldsymbol{\Sigma}_r)$；
　END
END

最终，我们根据式(8)的描述，推导出观察潜在变量的联合分布.

$$\begin{aligned}&P(v,l_v,w_v,t,z,r,s|\alpha,\alpha',\beta,\gamma,\tau,\psi,\mu,\Sigma)=\\&\quad P(z|\alpha)P(z|\alpha')P(r|\gamma)P(s|\delta)\\&P(w_v|z,\beta)P(v|r,\tau)P(t|z,\psi)P(l_v|r,\mu,\Sigma)=\\&\quad\int P(z|\theta)P(\theta|\alpha)\mathrm{d}\theta\int P(z|\theta')P(\theta'|\alpha)\mathrm{d}\theta'\\&\quad\int P(r|\vartheta)P(\vartheta|\gamma)\mathrm{d}\vartheta\int P(s|\varsigma)P(\varsigma|\delta)\mathrm{d}\delta\\&\quad\int P(w_v|z,\phi)P(\phi|\beta)\mathrm{d}\phi\int P(v|r,\varphi)\mathrm{P}(\varphi|\tau)\mathrm{d}\varphi\\&\quad P(t|z,\psi)P(l_v|r,\mu,\Sigma)\end{aligned} \tag{8}$$

3.3　模型推理

3.3.1　模型参数学习

我们的目的是学习观察潜在随机变量 v,l_v,W_v 和 t 的最大边缘对数似然参数. 精确推断 GTSCP 模型是困难的，由于对后验分布常量进行正则化是棘手的，因此我们利用马尔可夫链蒙特卡洛方法(MCMC)最大化式(8)中的完全数据似然性. 请注意我们采用共轭先验狄利克雷多项分布，因此我们很容易集成出 $\theta,\theta',\phi,\varphi,\varsigma$ 和 ϑ. 为了方便采样，我们根本不需要采样 θ，θ'，ϕ,φ,ς 和 ϑ. 至于超参数 $\alpha,\alpha',\beta,\delta,\gamma,\tau,b$ 和 b' 为简单起见，我们采用固定值，即 $\alpha=\alpha'=50/K$，$\gamma=50/R,\delta=\beta=\tau=0.01,b=b'=0.5$.

在吉布斯采样过程中，关于每一个用户签到行为(u,v,l_v,W_v,t)，我们需要获取正在采样的潜在社会相关性 s、潜在主题 z、潜在区域 r 的后验概率. 我们从式(8)所显示的观察潜在变量的联合概率分布开始，采用贝叶斯链式法则，可以方便地获得条件概率. 首先，我们通过以下后验概率采样社会相关性 s：

$$P(s|s_{-u,v},z,r,v,l_v,W_v,t,u,\cdot)\propto\frac{n_{u,s}^{-u,v}+\delta}{\Sigma_{s'}(n_{u,s'}^{-u,v}+\delta)} \tag{9}$$

其中：$s_{-u,v}$ 代表除了当前所有用户关系的社会相关性；$n_{u,s}$ 代表从用户 u 的社会相关性分布中采样潜在社会相关性 s 的次数；$n^{-u,v}$ 的上标 $-u,v$ 代表除了当前所有签到记录的数量.

然后我们采用两步吉布斯采样过程从而保证主题的一致性，首先根据后验概率采样硬币 a 如下：

$$P(a=1|a_{-u,v},z,u,\cdot)\propto \frac{n_{u,z}^{-u,v}+\alpha}{\sum_{z'}(n_{u,z'}^{-u,v}+\alpha)}\cdot\frac{n_{u,a1}^{-u,v}+b}{n_{u,a1}^{-u,v}+n_{u,a0}^{-u,v}+b+b'} \tag{10}$$

$$P(a=0|a_{-u,v},z,u,\cdot)\propto \frac{n_{l,z}^{-u,v}+\alpha'}{\sum_{z'}(n_{l,z'}^{-u,v}+\alpha')}\cdot\frac{n_{u,a0}^{-u,v}+b'}{n_{u,a1}^{-u,v}+n_{u,a0}^{-u,v}+b+b'} \tag{11}$$

其中：$n_{u,a1}$是在用户配置文件 D_u 中当 $a=1$ 时被采样的次数；$n_{u,a0}$在用户配置文件 D_u 中当 $a=0$ 时被采样的次数；$n_{u,z}$代表从用户 u 的兴趣分布中采样潜在主题 z 的次数；$n_{l,z}$代表从位置 l 的当地偏好分布中采样潜在主题 z 的次数.

然后，我们使采样主题 z 通过下面的后验概率，当 $a=1$ 时：

$$P(z|a=1,z_{-u,v},s,r,v,l_v,W_v,t,u,\cdot)\propto \frac{n_{u,z}^{-u,v}+\alpha}{\sum_{z'}(n_{u,z'}^{-u,v}+\alpha)}\cdot\frac{(1-t)^{\psi_{z,1}^{-1}}t^{\psi_{z,2}^{-1}}}{B(\psi_{Z,1},\psi_{Z,2})}\cdot\prod_{w\in W_v}\frac{n_{z,w}^{-u,v}+\beta}{\sum_{w'}(n_{z,w'}^{-u,v}+\beta)} \tag{12}$$

当 $a=0$ 时：

$$P(z|a=0,z_{-u,v},s,r,v,l_v,W_v,t,u,\cdot)\propto \frac{n_{l,z}^{-u,v}+\alpha'}{\sum_{z'}(n_{l,z'}^{-u,v}+\alpha')}\cdot\frac{(1-t)^{\psi_{z,1}^{-1}}t^{\psi_{z,2}^{-1}}}{B(\psi_{Z,1},\psi_{Z,2})}\cdot\prod_{w\in W_v}\frac{n_{z,w}^{-u,v}+\beta}{\sum_{w'}(n_{z,w'}^{-u,v}+\beta)} \tag{13}$$

其中：$z_{-u,v}$代表除了当前所有签到记录的主题配置；$n_{z,w}$代表从主题 z 生成词 w 的次数.

最终，我们通过以下后验概率采样区域 r：

$$P(r|r_{-u,v},s,z,v,l_v,W_v,t,u,\cdot)\propto \frac{n_{u,r}^{-u,v}+\gamma}{\sum_{r'}(n_{u,r'}^{-u,v}+\gamma)}\cdot\frac{n_{r,v}^{-u,v}+\tau}{\sum_{v'}(n_{r,v'}^{-u,v}+\tau)}\cdot P(l_v|\boldsymbol{\mu}_r,\boldsymbol{\Sigma}_r) \tag{14}$$

其中：$n_{u,r}$代表从用户 u 的空间活动分布中采样区域 r 的次数；$n_{r,v}$代表区域 r 生成的 POIv 的次数. 为了简化并加速，我们根据配置的潜在变量 r 和 z 在每次迭代后采用矩量法更新贝塔分布与高斯分布的参数 μ，Σ 和 ψ. 特别是，参数 $\boldsymbol{\mu}_r$ 和 $\boldsymbol{\Sigma}_r$ 通过式(15)与(16)所更新.

$$\boldsymbol{\mu}_r=E(r)=\frac{1}{|s_r|}\sum_{v\in s_r}l_v \tag{15}$$

$$\boldsymbol{\Sigma}_r=D(r)=\frac{1}{|s_r|-1}\sum_{v\in s_r}(l_v-\boldsymbol{\mu}_r)(l_v-\boldsymbol{\mu}_r)^{\mathrm{T}} \tag{16}$$

其中，s_r 代表配置在潜在区域 r 的 POIs 的集合. 更新贝塔分布参数 ψ 如下：

$$\psi_{z,1}=t_z\left(\frac{t_z(1-t_z)}{s_z^2}-1\right) \tag{17}$$

$$\psi_{z,2}=(1-t_z)\left(\frac{t_z(1-t_z)}{s_z^2}-1\right) \tag{18}$$

其中，t_z 和 s_z^2分别代表配置在主题 z 的时间戳的采样平均值和采样协方差.

3.3.2 模型推理框架

经过大量采样迭代后，通过检验 s，r 和 z 的数量，使用近似后验输出估计参数. 模型推理框架如算法 2 所示. 我们首先使用 K-Means 方法初始化地理位置的聚类(算法中 3～4 行). 然后基于签到记录随机初始化主题配置(5～9 行). 然后，关于每一次迭代，每个签到记录(u,v,l_v,W_v,t)，通过式(9)～(14)更新社会相关性、主题和区域配置(算法中 12～18 行). 在每一次迭代后，我们更新高斯分布和贝塔分布的参数(算法中 19～20 行). 重复迭代直至收敛(算法中 11～26 行)，另外，在初始几百次的迭代中引入一个老化过程来移除不可靠的采样结果，并且仅在周期性采样之后，我们又引入采样滞后(即老化过程后样本之间的间隔)，以避免样本之间的相关性(算法中 21～25 行).

$$\theta_{u,z}^{sum}\mathrel{+}=\frac{n_{u,z}+\alpha}{\sum_{z'}(n_{u,z'}+\alpha)} \tag{19}$$

$$\theta'^{sum}_{l,z}\mathrel{+}=\frac{n_{l,z}+\alpha'}{\sum_{z'}(n_{l,z'}+\alpha')} \tag{20}$$

$$\phi_{z,w}^{sum}\mathrel{+}=\frac{n_{z,w}+\beta}{\sum_{w'}(n_{z,w'}+\beta)} \tag{21}$$

$$\varsigma_{u,s}^{sum}\mathrel{+}=\frac{n_{u,s}+\delta}{\sum_{s'}(n_{u,s'}+\delta)} \tag{22}$$

$$\vartheta_{u,r}^{sum}\mathrel{+}=\frac{n_{u,r}+\gamma}{\sum_{r'}(n_{u,r'}+\gamma)} \tag{23}$$

$$\varphi_{r,v}^{sum}\mathrel{+}=\frac{n_{r,v}+\tau}{\sum_{v'}(n_{r,v'}+\tau)} \tag{24}$$

$$\psi_z^{sum}\mathrel{+}=\psi_z \tag{25}$$

$$\mu_r^{sum}\mathrel{+}=\mu_r \tag{26}$$

$$\Sigma_r^{sum}\mathrel{+}=\Sigma_r \tag{27}$$

$$\hat{\lambda}_u=\frac{n_{us1}+b}{n_{us1}+n_{us0}+b+b'} \tag{28}$$

根据已有的数据来递归估计似然函数. 参数估计法能在数据模型未知参数和确实数据之间建立一种关系. 如果已知缺失变量与其他变量间的关系，则可直接估计该分布模型的未知参数. 若已知分布模型参数，则可以获得缺失值的无偏估计值. 这样模型参数与缺失值之间的相互依赖关系可以用迭代方法得到. 本文模型首先假定参数值来预测 POI 的缺失特征值，然后再用预测值更新模型的参数估计值，反

复进行，直到模型的参数估计最终收敛. 所以本文所提的 GTSCP 模型在一定程度上可以解决 POI 特征缺失问题.

算法 2.　GTSCP 模型的推理框架.

输入：用户签到集合 D，迭代次数 I，老化次数 I_b，采样滞后 I_S，先验 $\alpha,\alpha',\beta,\delta,\gamma,\tau,b,b'$

输出：估计参数 $\hat{\theta},\hat{\theta}',\hat{\phi},\hat{\varsigma},\hat{\vartheta},\hat{\varphi},\hat{\psi},\hat{\mu},\hat{\Sigma},\hat{\lambda}$

1. 创建临时变量 $\theta^{sum},\theta'^{sum},\phi^{sum},\varsigma^{sum},\vartheta^{sum},\varphi^{sum},\psi^{sum},\mu^{sum}\Sigma^{sum}$ 并初始化它们为 0；
2. 创建临时变量 $\hat{\psi},\hat{\mu},\hat{\Sigma}$；
3. 使用 K-Means 方式初始化地理位置聚类；
4. 分别通过式(15)和(16)更新 μ 和 Σ；
5. FOR 每个 $D_u\in D$ DO
6. 　FOR 每个签到记录 $(u,v,l_v,W_v,t)\in D_u$ DO
7. 　　随机配置主题；
8. 　END
9. END
10. 初始化变量计数为零；
11. FOR 迭代＝1 到 I DO
12. 　FOR 每个 $D_u\in D$ DO
13. 　　FOR 每个签到记录 $(u,v,l_v,W_v,t)\in D_u$ do
14. 　　使用式(9)更新社会相关性配置；
15. 　　使用式(10)～(13)更新主题配置；
16. 　　使用式(14)更新区域配置；
17. 　　END
18. 　END
19. 　分别通过式(15)和(16)更新 μ 和 Σ；
20. 　通过式(17)和(18)更新 ψ；
21. 　IF (迭代＞I_b)并且(迭代 mod I_s＝＝0) THEN
22. 　　计数＝计数＋1；
23. 　　更新 $\theta^{sum},\theta'^{sum},\phi^{sum},\varsigma^{sum},\vartheta^{sum},\varphi^{sum},\psi^{sum},\mu^{sum}\Sigma^{sum}$ 如下：
24. 　　通过式(19)～(28)计算 $\theta^{sum}_{u,z}+,\theta'^{sum}_{l,z}+,\phi^{sum}_{z,w}+,\varsigma^{sum}_{u,s}+,\vartheta^{sum}_{u,r}+,\varphi^{sum}_{r,v}+,\psi^{sum}_{z}+,\mu^{sum}_{r}+,\Sigma^{sum}_{r}+,\hat{\lambda}_u$
25. 　END
26. END
27. RETURN 模型参数 $\hat{\theta}=\frac{\theta^{sum}}{count},\hat{\theta}'=\frac{\theta'^{sum}}{count},\hat{\phi}=\frac{\phi^{sum}}{count},\hat{\varsigma}=\frac{\varsigma^{sum}}{count},\hat{\vartheta}=\frac{\vartheta^{sum}}{count},\hat{\varphi}=\frac{\varphi^{sum}}{count},\hat{\psi}=\frac{\psi^{sum}}{count},\hat{\mu}=\frac{\mu^{sum}}{count},\hat{\Sigma}=\frac{\Sigma^{sum}}{count}$

3.3.3　时间复杂度

我们分析以上 GTSCP 模型推理框架的时间复杂度. 假设过程需要 $O(K)$ 次迭代才能达到收敛. 在每次迭代中都需要所有的用户签到记录. 关于每一次签到记录，首先对于采样潜在社会相关性需要 $O(S)$ 操作计算后验分布，然后对于采样潜在主题需要 $O(K)$ 操作计算后验分布，最后关于采样潜在区域需要 $O(R)$ 操作计算后验分布. 因此，总的时间复杂度为 $O(I(S+K+R)\Sigma_u|F_u|\Sigma_u|D_u|)$.

3.4　空间平滑优化

为处理用户兴趣漂移问题，我们构建当地偏好模型. 为进一步处理数据稀疏问题，我们利用树型结构，称为空间金字塔，划分空间项目位置到不用层次不同大小的空间网络中. 更具体地说，空间金字塔分解空间成 H 个水平层. 第 0 层只有一个空间网格. 对于一个给定的第 h 层，空间被划分成 4^h 个面积相等的网格. 因此，空间可以递归地划分成很多不同层次不同粒度的单元格. 空间金字塔的图解如图 3 所示.

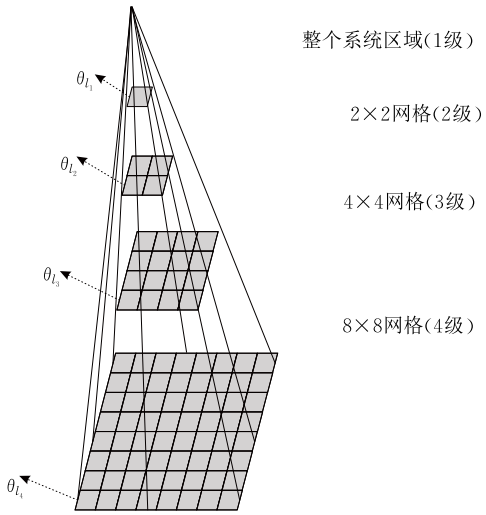


图 3　空间金字塔

在空间数据挖掘中的一个基本假设是一切事物都与其他事物相关联，附近的事物比遥远的事物更加具有相关性，此理论为地理学第一定律[36]. 这个定律被称为“空间自相关”. 空间金子塔可以有效地执行这个定律. 换言之，对于每个位置 l，其可以表示为一条从根节点到其相对应的叶节点的路径. 我们使用一个向量 $(l_1,l_2,\cdots,l_h,\cdots,l_H)$ 来描述路径. 基于位置向量的表示，我们可以很容易地计算出两个位置之间的邻近度. 如果两个位置在空间金子塔共享越多的祖先，那么这两个位置就越邻近. 当在位置 l 的用户活动数据非常稀疏时，则当地偏好 θ'_l 就不能被准确地评估. 为了处理这个问题，我们利用地理相关性来增强模型参数 θ'_l 的先验知识. 在地理空间中，如果两个位置 l 和 l' 是邻近的，则当地偏好 θ'_l 和 $\theta'_{l'}$ 更加相似. 为了融合地理相关性信息到我们的

GTSCP 模型中,基于路径向量来计算位置 l 的当地偏好 θ'_l. 给定一个位置 l,当地偏好表示如下:

$$\theta_l = \sum_{h=1}^{H} \theta_{l_h} \tag{29}$$

根据上面的公式,一个位置的当地偏好取决于它的所有祖先的根. 通过这种表达方式表明邻近位置分享相似的偏好,即通过空间金字塔对当地偏好进行平滑优化. 同时,如果在一个位置有很少或者几乎没有活动时,我们仍可以通过它们的祖先推断它们的偏好. 另外,一旦每个层级的当地偏好被学习后,就不用再重新训练参数,通过改变模型中的最低层级,该模型可以在不同的粒度之间进行快速且方便地转换.

在模型推理的过程中,我们需要优化当地偏好.

因为当地偏好参数的计算为每个层级相对应的参数的和,我们需要计算每个层级参数的粒度. 因此,我们推断当地偏好如下:

$$\rho_{u,l,z} = \frac{n_{u,z}^{-u,v} + \alpha}{\sum_{z'}(n_{u,z'}^{-u,v} + \alpha)} \cdot \frac{n_{l,z}^{-u,v} + \alpha'}{\sum_{z'}(n_{l,z'}^{-u,v} + \alpha')} \tag{30}$$

$$\frac{\partial L}{\partial \theta_{u,z}} = d(u,z) - \sum_{i=1}^{|D_u|} \rho_{u,l_i,z} \tag{31}$$

$$\frac{\partial L}{\partial \theta_{l,z}} = d(l,z) - \sum_{j=1}^{|D_l|} \rho_{u_j,l,z} \tag{32}$$

$$\frac{\partial L}{\partial \theta_{l_h,z}} = d(l_h,z) - \sum_{j=1}^{|D_{l_h}|} \rho_{u_j,l_j,z} \tag{33}$$

其中:$d(u,z)$代表在 D_u 中有多少活动被分配到主题 z;$d(l,z)$代表在位置 l 发生的活动的数量;u_j 代表用户生成的第 j 个活动记录;$d(l_h,z)$代表在位置 l_h 配置到主题 z 的活动记录的数量;D_{l_h} 是在位置 l_h 发生的活动记录集合.

4 利用 GTSCP 模型的兴趣点推荐

4.1 兴趣点推荐的查询过程

给定一个查询用户 u_q 以及查询用户的当前位置 l_q 和查询时间 t_q,即查询 $q=(u_q,t_q,l_q)$,当我们已经学习完模型参数集合 $\hat{\Omega}=\{\hat{\theta},\hat{\theta}',\hat{\phi},\hat{\varsigma},\hat{\vartheta},\hat{\varphi},\hat{\psi},\hat{\mu},\hat{\Sigma},\hat{\lambda}\}$后,我们计算用户 u_q 在每个 POIv 签到的概率,然后选择 top-k 列表中的最高概率的 POIs 推荐给查询用户. 特别是,给定查询用户的查询时间 t_q 和当前位置 l_q 以及学习过的模型参数 $\hat{\Omega}$,我们计算用户 u_q 在每个 POIv 签到的概率如下:

$$P(v,l_v,W_v|u_q,t_q,l_q,\hat{\Omega}) \propto \sum_r R(r) \sum_s S(s) \sum_z Z(z) \tag{34}$$

$$P(v,l_v,W_v|u_q,t_q,l_q,\hat{\Omega}) = \frac{P(v,l_v,W_v,t_q|u_q,l_q,\hat{\Omega})}{\sum_{v'} P(v',l_{v'},W_{v'},t_q|u_q,l_q,\hat{\Omega})} \propto P(v,l_v,W_v,t_q|u_q,l_q,\hat{\Omega}) \tag{35}$$

$P(v,l_v,W_v,t_q|u_q,l_q,\hat{\Omega})$计算如下:

$$P(v,l_v,W_v,t_q|u_q,l_q,\hat{\Omega}) = \sum_r P(r|l_q,\hat{\Omega}) P(v,l_v,W_v,t_q|u_q,r,\hat{\Omega}) \tag{36}$$

其中,$P(r|l_q,\hat{\Omega})$代表用户 u 的当前位置 l_q 在区域 r 的概率,并通过贝叶斯法则由式(32)计算,通过式(33)估计潜在区域 r 的先验概率如下:

$$P(r|l_q,\hat{\Omega}) = \frac{P(r)P(l_q|r,\hat{\Omega})}{\sum_{r'} P(r')P(l_q|r',\hat{\Omega})} \propto P(r)P(l_q|r,\hat{\Omega}) \tag{37}$$

$$P(r) = \sum_u P(r|u)P(u) = \sum_u \frac{N_u + \xi}{\sum_{u'}(N_{u'} + \xi)} \hat{\vartheta}_{u',r} \tag{38}$$

其中,N_u 代表有用户 u 所生成的签到数量. 为避免过度拟合,我们引入狄利克雷先验参数 ξ. $P(v,l_v,W_v,t_q|u_q,r,\hat{\Omega})$被定义如下:

$$P(v,l_v,W_v,t_q|u_q,r,\hat{\Omega}) = P(l_v|r,\hat{\Omega}) P(v|r,\hat{\Omega}) \sum_s \Big(\prod_{u \in F_u} P(s|u_q,\hat{\Omega})\Big)^{\frac{1}{|F_u|}} \cdot \sum_z P(z|u_q,\hat{\Omega}) P(z|l_q,\hat{\Omega}) P(t_q|z,\hat{\Omega}) \cdot \Big(\prod_{w \in W_v} P(w|z,\hat{\Omega})\Big)^{\frac{1}{|W_v|}} \tag{39}$$

我们利用关于主题 z 所生成的词集合 W_v 的概率几何平均值,即 $P(W_v|z,\hat{\Omega}) = \prod_{w \in W_v} P(w|z,\hat{\Omega})$,考虑到与不同的 POIs 相关的词的数量可能是不同的.

通过式(35)~(39),初始的式(34)可以表述为式(40)如下:

$$\begin{aligned} P(v,l_v,W_v|u_q,t_q,l_q,\hat{\Omega}) &\propto \sum_r R(r) \sum_s S(s) \sum_z Z(z) \propto \\ &\sum_r \Big[P(r) P(l_q|r,\hat{\Omega}) P(l_v|r,\hat{\Omega}) P(v|r,\hat{\Omega}) \cdot \\ &\sum_s \Big(\prod_{u \in F_u} P(s|u_q,\hat{\Omega})\Big)^{\frac{1}{|F_u|}} \cdot \\ &\sum_z P(z|u_q,\hat{\Omega}) P(z|l_q,\hat{\Omega}) P(t_q|z,\hat{\Omega}) \cdot \\ &\Big(\prod_{w \in W_v} P(w|z,\hat{\Omega})\Big)^{\frac{1}{|W_v|}} \Big] = \end{aligned} \tag{40}$$

$$\sum_r [P(r)P(l_q|\hat{\mu}_r,\hat{\Sigma}_r)P(l_v|\hat{\mu}_r,\hat{\Sigma}_r)\hat{\varphi}_{r,v}] \cdot \sum_s \left[\left(\prod_{u\in F_u} \hat{\varsigma}_{u_q,s} \right)^{\frac{1}{|F_u|}} \right] \cdot \sum_z \left[\hat{\theta}_{u_q,z}\hat{\theta}'_{l_q,z}\hat{\psi}_{z,t_q} \left(\prod_{w\in W_v} \hat{\phi}_{z,w} \right)^{\frac{1}{|W_v|}} \right] \tag{40}$$

4.2　有效在线的兴趣点推荐

我们所提的兴趣点推荐方法是训练离线的GTSCP模型，并获得包含用户兴趣、签到行为和POI属性的必要信息的知识模型，然后在实时查询q时，利用知识模型检索top-k列表中推荐的POIs. 为了加速在线查询过程，基于式(40)，我们提出一个排名框架公式如下：

$$S(q,v)= \sum_r W(q,r)F(v,r)\sum_s W(q,s)\sum_z W(q,z)F(v,z) \tag{41}$$

查询偏好如下：

$$W(q,r) = P(l_q|\hat{\mu}_r,\hat{\Sigma}_r) \tag{42}$$

$$W(q,s) = \left(\prod_{u\in F_u} \hat{\varsigma}_{u_q,s} \right)^{\frac{1}{|F_u|}} \tag{43}$$

$$W(q,z) = \hat{\theta}_{u_q,z}\hat{\theta}'_{l_q,z}\hat{\psi}_{z,t_q} \tag{44}$$

POIv的属性如下：

$$F(v,r) = P(r)P(l_v|\hat{\mu}_r,\hat{\Sigma}_r)\hat{\varphi}_{r,v} \tag{45}$$

$$F(v,z) = \left(\prod_{w\in W_v} \hat{\phi}_{z,w} \right)^{\frac{1}{|W_v|}} \tag{46}$$

关于查询q，需要匹配查询偏好与POIv的属性. $S(q,v)$代表POIv的排名分数. 每个区域、社会相关性或者主题被认为是一个属性(即$a=r,s$或者z)，$W(q,a)$代表q在属性a上的权重，$F(v,a)$代表与属性a相关的POIv的分数. 这个排名框架从在线分数计算中分离了离线分数计算. 因为$F(v,a)$是独立的查询，它可以在离线中被计算. 尽管查询权重$W(q,a)$可以在线计算，但它主要的时间消耗部分(即$\hat{\theta}_{u_q,z}$，$\hat{\theta}'_{l_q,z}$，$\hat{\psi}_{z,t_q}$，$(\hat{\mu}_r,\hat{\Sigma}_r)$)同样是离线计算的，在线计算只是一个简单的组合过程. 这个设计考虑了最大预先计算问题，进而减少查询时间.

简单的方法生成top-k列表中的POIs，需要通过式(41)计算所有POIs的排名分数，并选择排名分数最高的top-k个POIs. 然而这样的计算方法效率低，特别是当POIs的数量或者POIs属性的数量变大时. 基于查询偏好稀疏性的观察，一个查询q只喜欢少量的属性(即3～5个潜在维度)，在大多数属性中查询权重是极其小的. 因此，关于top-k推荐[3,37]，我们采用基于阈值算法(TA)的查询过程技术来提高在线推荐的有效性. 因为$W(q,a)$是非负的，给定一个查询q，我们所提的排名式(36)是单调上升的，满足了基于阈值算法的查询过程技术的要求. 该技术具有良好的性能，不需要扫描所有的POIs，通过检查少量的POIs的数量，能够正确地找到top-k结果，使得GTSCP模型能够扩展应用到大规模的数据集.

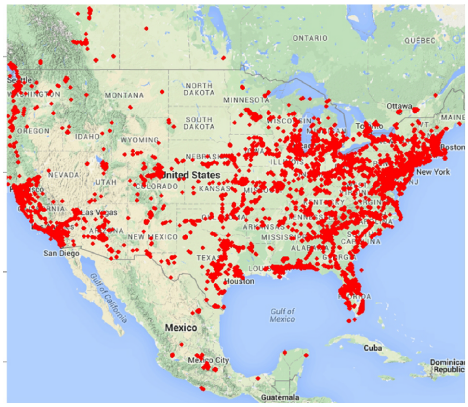
5　实　验

5.1　实验设置

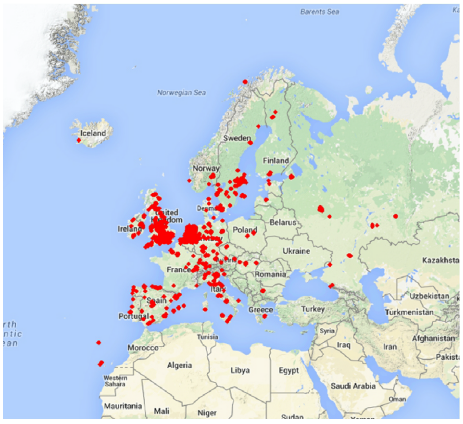
5.1.1　数据集描述

Foursquare数据集. 其是一个大规模的基于位置的社交网络站点. 它允许用户在不同的位置进行签到，通过分析这些签到数据相关的空间、时间、社会、文本与流行度信息，报导人们移动模式的定量评估. 我们使用文献[38]所提供的数据集中的Foursquare的数据集，如图4的地图所示，显示了在我们的实验室中，签到时间跨度从2010年9月到2011年11月，所有美国和欧洲用户的签到记录以及一个特定用户的签到记录. 关于每个用户，该数据集包含了用户的社交网络关系，签到POIs的IDs、每个签到的POI的地理位置(即经纬度坐标)、签到时间和每个POI的内容. 每个签到的存储形式为用户-ID、POI-ID、POI-位置、签到时间、POI-内容. 在社交网络关系中的每个记录的存储形式为用户-ID、朋友-ID和社会关系的总数量为512643. 为了清洗并移除较少发生的异常数据，我们过滤掉少于10次签到的用户，并要求每个POI应该被用户至少访问10次. 此外，我们还限制了用户应至少访问5个不同的POIs. 总结如下，在我们实验中的Foursquare数据集包含5468个用户总共访问了7286个POIs. 数据集的统计数据如表3所示.

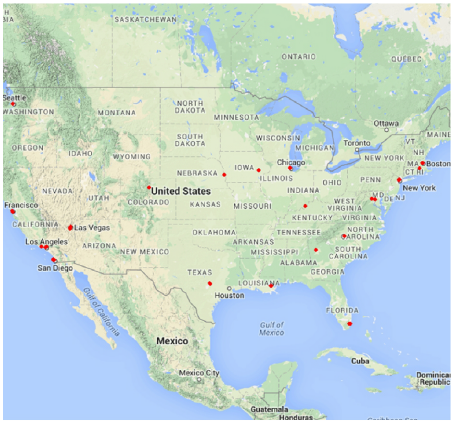
Twitter数据集. Twitter支持第3方的位置共享服务如Foursquare，Gowalla，这些服务的用户可以在Twitter上分享他们的签到. 我们使用文献[38]所提供的数据集中的Twitter的数据集. 但是原始的数据集不包含每个POI的分类与标签信息，所以我们通过Foursquare的公共API，获取每个POI的内容信息. 每个签到记录的存储形式与上述的Foursquare数据集的每个签到记录存储形式相同. Twitter数据集包含114508个用户总共访问了62462个POIs. 数据集的统计数据如表3所示.



(a) 在美国的所有签到



(b) 在欧洲的所有签到



(c) 在美国某一用户的签到

图 4 本文 Foursquare 数据集的签到地域分布

表 3 数据集的统计

数据集	用户的数量	POIs 的数量	分类的数量	社会关系的数量	签到的数量	时间跨度
Foursquare	5468	7286	60	35216	512643	2010.9～2011.11
Twitter	114508	62462	100	50566	1434668	2010.9～2011.1

豆瓣数据集. 豆瓣网是中国最大的基于事件的社交网站,用户可以发布和参与社交活动. 我们使用文献[3]所提供的豆瓣数据集. 在豆瓣网上,某个用户可以通过指定事件的内容、时间和地点来创建一个社会事件,其他用户可以通过在线回复表达他们的意愿来参与到事件当中. 该数据集包含 100000 个用户,300000 个事件和 3500000 个正确定义的发请帖. 记录采集的数据集的存储形式如下:(1) 用户信息,包括用户-ID,用户-名字,和用户-所在城市;(2) 事件信息,包括事件-ID,事件-名称,事件-经度,事件-维度,事件-总结及分类;(3) 用户反馈信息,包括用户-ID,事件-ID.

本论文所提的方法是一个灵活通用的算法,同样适用于文献[15]中所提及的点评网数据如 Yelp, Douban 等数据集来进行实验评估.

5.1.2 评估方法

因为我们所提的 GTSCP 模型的设计支持本地和异地两个推荐场景,我们分别在两个场景中对 GTSCP 模型进行有效地推荐评估. 依据用户签到行为的集合,给定一个用户配置文件,我们将用户签到行为数据划分成训练集和测试集. 关于异地推荐场景,当用户到异地旅行时,我们随机选择用户生成的签到数据的 20%作为测试集,其余用户生成的签到数据作为训练集. 同样,关于本地推荐场景,我们随机选择在本地用户生成的签到数据的 20%作为测试集,其余用户生成的签到数据作为训练集. 为了决定一个活动记录发生在本地还是异地,我们测量 POI 的位置与用户本地位置的距离(即$|l_q - l_u|$). 如果距离大于阈值 d,我们假设用户的活动发生在异地,否则,用户的活动发生在本地. 另外,为了模仿更

真实的兴趣点推荐情景，我们选择一个位置坐标作为目标用户在访问 POIv 之前的当前所处位置. 特别是，关于每个(u,v,l_v,W_v,t)的测试情况，我们使用一个高斯函数生成一个以 l_v 为中心、圆半径为 d 的地理坐标 l. 它代表目标用户 u 的当前所处位置. 因此，根据测试情况，生成一个查询 $q=(u_q,t_q,l_q)$.

根据以上划分方法，我们划分用户签到行为数据集 D 为训练集 D_{train} 和测试集 D_{test}. 我们利用评估方法和测量 $Precision@K$ 和 $Recall@k$ 来评估推荐方法的效果. 特别是，关于在测试集 D_{test} 中的每个签到行为记录(u,v,l_v,W_v,t)以及相关的查询 q，首先，我们计算关于 POIv 以及其他未被用户 u 之前访问的并在以 l_v 为中心、圆半径为 d 范围内的 POIs 的排名分数. 然后，我们通过他们的排名分数来排序这些 POIs 生成一个排名列表. 让 p 代表 POIv 在这个列表中的位置. 在这种情况下，最好的结果是 POIv 优于其他未被访问的 POIs，则 $p=1$. 最后，我们从列表中挑选 k 个最高排名的 POIs 生成 top-k 推荐列表. 如果 $p\leq k$，我们就有一次命中（即真实的 POIv 推荐给用户）. 否则，我们就有一次丢失.

在我们的方法对比试验中，采用两种指标准确率 $Precision@K$ 和召回率 $Recall@k$ 来评估兴趣点推荐的质量. $Precision@K$ 定义如下：

$$Precision@K=\frac{\sum_u|R(u)\cap T(u)|}{K} \tag{46}$$

其中：K 是推荐给用户的 POIs 的数量；$R(u)$是推荐给用户 POIs 的 top-k 列表；$T(u)$是用户实际访问的 POIs 的数量.

关于一次测试情况，如果真实的 POIv 出现在 top-k 列表中，我们就定义 $hit@k$ 的值为 1，否则，$hit@k$ 的值为 0. 平均所有测试情况总的 $Recall@k$ 定义如下：

$$Recall@k=\frac{\#hit@k}{|D_{\text{test}}|} \tag{47}$$

其中：$\#hit@k$ 代表测试集中的命中的次数；$|D_{\text{test}}|$为所有测试集的数量. 定义通过一个标准的 5 倍交叉验证证明.

5.1.3　对比方法

我们所提的 GTSCP 模型和其他先进的兴趣点推荐技术的对比如下：

(1) TACF. TACF 融合了时间效应，是先进的时间感知的协同过滤兴趣点推荐方法[22]. 特别是，TACF 分裂时间成基于小时的时间片段，并在一个时间片段内通过已访问的 POIs 建模用户对 POIs 的时间偏好. 给定一个用户和一个具体的时间，TACF 会找到与用户分享相似的时间偏好.

(2) GeoMF. GeoMF 是一种矩阵分解模型，利用用户签到行为的空间聚类现象[6]. 特别是，GeoMF 增强了用户和 POIs 的潜在因素以及用户的活动区域向量和 POIs 的影响区域向量，从二维核密度评估方面捕捉空间聚类现象并将其融合到矩阵分解中.

(3) UPS-CF. UPS-CF 是一个协同推荐框架，特别为异地用户而设计的[2]. 这个框架考虑了用户偏好、社会关系、地理相似性，并结合基于用户的协同过滤和基于社会的协同过滤方法，通过相似的用户或者用户的朋友的签到记录推荐 POIs 给目标用户.

(4) LCA-LDA. LCA-LDA 是位置-内容感知的推荐模型，支持用户到新城市旅行场景的兴趣点推荐[3]. 此模型通过 POIs 的内容和 POI 的共访问模式考虑了个人兴趣和每个城市的当地偏好. 与我们所提的 GTSCP 模型相比，该模型忽略了地理影响、社会关系和时间效应.

(5) TGSC-PMF. TGSC-PMF 是一种上下文感知的概率矩阵分解兴趣点推荐算法. 此模型利用兴趣点的地理、文本、社会、分类与流行度信息，并将这些因素进行有效地融合. 该模型是本文作者之前研究工作所提出的模型. 与现在该文所提的 GTSCP 模型相比，该模型忽略了时间效应.

(6) SVDFeature. SVDFeature 是一个机器学习工具包旨在解决基于特征的矩阵分解[39]. 基于这个工具包，构建一个融合更多的边信息的分解模型，即融合 POI 的地理位置，POI 的内容信息，社会相关性，时间动态（即签到时间）和 POI 的流行度，与我们所提的 GTSCP 模型进行公平地对比. SVDFeature 的局限性是它不能处理连续的位置和时间，因此我们利用文献[22,40]中的离散方法把它们分割成箱子和网格方块.

我们设计 5 种基线方法进一步验证分别利用地理影响、时间效应、社会相关性、内容信息和流行度影响所带来的好处. TSCP 是第 1 个 GTSCP 模型的简化版即没有考虑地理影响. GSCP 是第 2 个 GTSCP 模型的简化版移除了签到时间信息，关于每个签到记录潜在变量 z 只生成词集合 W_v. GTCP 是第 3 个 GTSCP 模型的简化版即没有考虑社会相关性. GTSP 是第 4 个 GTSCP 模型的简化版移除了 POIs 的内容信息，潜在变量 z 只生成签到时间. GTSC 是第 5 个 GTSCP 模型的简化版即没有考虑流行度影响，所以潜在变量 r 只生成 POIv 的 ID，则流行度感知用

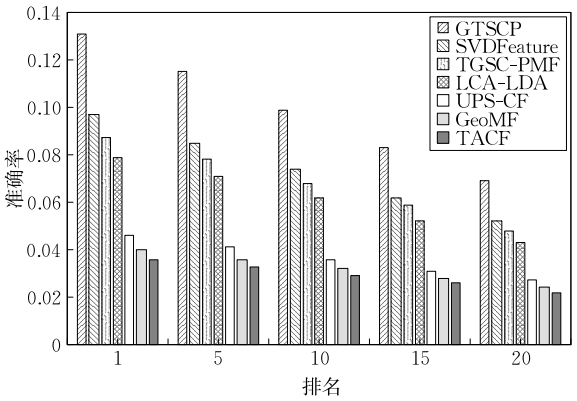
户移动模型退化为一个基于模型的协同过滤方法.

5.2 实验结果

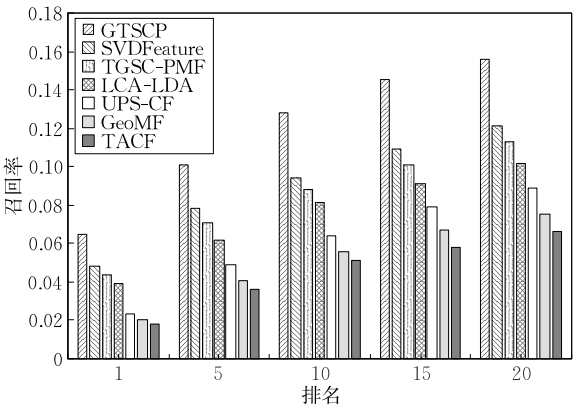
5.2.1 推荐效果

在这部分中,在调优参数的情况下,我们呈现了对比推荐方法的实验结果.依据 top-k 的准确率和召回率,明显显示了对比推荐方法的不同性能.从图中所示,我们可以观察到准确率的值是随着 k 值的增加而下降的,召回率的值是随着 k 值的增加而上升的.这是因为返回越多的用户签到行为,就越有可能发现更多的用户可能愿意访问的 POIs.请注意我们只显示当 k 值设置为 1,5,10,15,20 时的推荐性能.关于 top-k 推荐任务,更大的 k 值通常被忽略.

Foursquare 数据集上的异地推荐场景.图 5(a)显示了在 Foursquare 数据集上的异地推荐场景下的准确率.当 $k=10$ 时,GTSCP 的准确率大约为 0.099,当 $k=20$ 时,其准确略大约为 0.069.这就意味着当推荐一个有吸引力的 POI 分别在 Top-10 和 Top-20 时 GTSCP 模型分别有 9.9%和 6.9%的准确率.图 5(b)显示了在 Foursquare 数据集上的异地推荐场景下的召回率.当 $k=10$ 时,GTSCP 的召回率大约为 0.128,当 $k=20$ 时,其召回率大约为 0.156.这就意味着当推荐一个有吸引力的 POI 分别在 Top-10 和 Top-20 时 GTSCP 模型分别有 12.8%和 15.6%的召回率.很明显,我们所提的 GTSCP 模型明显优于其他的对比推荐方法,证明 GTSCP 的优势超过其他对比方法.一些观察结果如下:(1) TACF,GeoMF 和 UPS-CF 的性能落后于 GTSCP,SVDFeature,TGSC-PMF 和 LCA-LDA 的性能,显示了潜在分类模型融合签到的 POIs 的内容信息的优势.这是因为用户在异地区域有很少的签到行为记录,因此在异地推荐场景中这些基于 CF 或者基于矩阵分解的方法会遭受到严重的数据稀疏问题.而 GTSCP,SVDFeature,TGSC-PMF 和 LCA-LDA 是潜在分类模型并融合了 POIs 的内容信息,可以很大程度地缓解数据稀疏问题;(2) UPS-CF 的性能优于 TACF 和 GeoMF 的性能,显示了利用社会影响的好处.在异地推荐场景中通过社会朋友的签到记录可以较小程度地缓解数据稀疏问题,因为我们有时候去异地旅行时会看望我们的朋友;(3) GTSCP 和 SVDFeature 的性能优于 LCA-LDA 的性能.这是因为 LCA-LDA 相比于 GTSCP 和 SVDFeature 忽略了地理影响、时间效应和社会相关性;(4) GTSCP 和 SVDFeature 的性能优于 TGSC-PMF 的性能.这是因为 TGSC-PMF 相比于 GTSCP 和 SVDFeature 忽略了时间效应;(5) TGSC-PMF 的性能优于 LCA-LDA 的性能.这是因为 LCA-LDA 相比于 TGSC-PMF 忽略了地理影响和社会相关性;(6) GTSCP 所达到的推荐精确度远高于 SVDFeature,尽管它们使用相同类型的特征和信息,显示了概率生成模型的优势胜过基于特征的矩阵分解.其中可能的原因是当离散化时间和地理位置时会丢失一些重要的信息.



(a) 异地推荐的top-k准确率

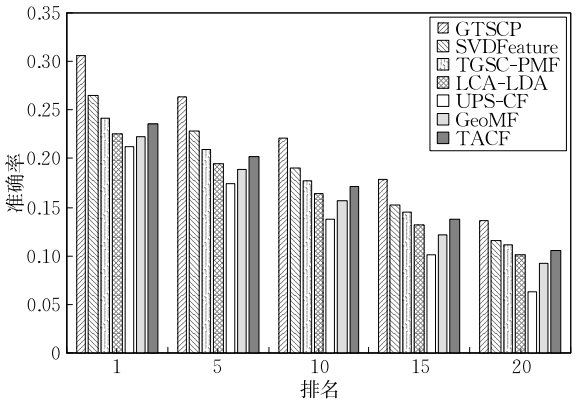


(b) 异地推荐的top-k召回率

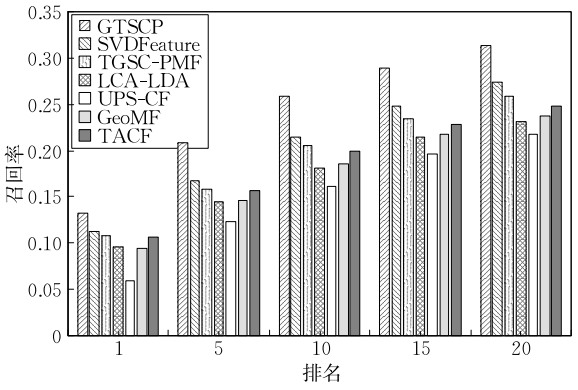
图 5 Foursquare 数据集上的异地推荐 top-k 性能

Foursquare 数据集上的本地推荐场景.图 6 显示了在本地推荐场景中所有对比的推荐模型的性能.一些观察结果如下:(1) 从结果中可以观察到图 6 中的所有方法的推荐精确度都高于图 5 中所示;(2) 另外,图 5 中 LCA-LDA 的性能胜过 TACF 的性能,然而图 6 中 TACF 的性能略超过 LCA-LDA,UPS-CF 和 GeoMF 的性能,由于其融合了时间模式,显示了时间感知的协同过滤方法更适合于用户-POI 矩阵不稀疏的时候,然而基于模型的方法,特别是结合用户签到行为的语义模式更有能力克服异地推荐场景的数据稀疏问题;(3) 图 5 与图 6 中比较的 LCA-LDA 和 GeoMF,我们可以观察到在异地推荐场景中利用语义模式更有利,然而在本地

推荐场景中利用空间模式更有利；(4) 另一个观察是图 6 中的 GTSCP 模型和其它对比的推荐方法之间的性能差距小于图 5 中所示，显示了当数据稀疏问题不严重的时候，推荐方法之间的性能差异就变得不那么明显. 图 5 和图 6 之间的比较显示这两个推荐场景本质上是不同的，应该被分开单独评估.



(a) 本地推荐的top-k准确率

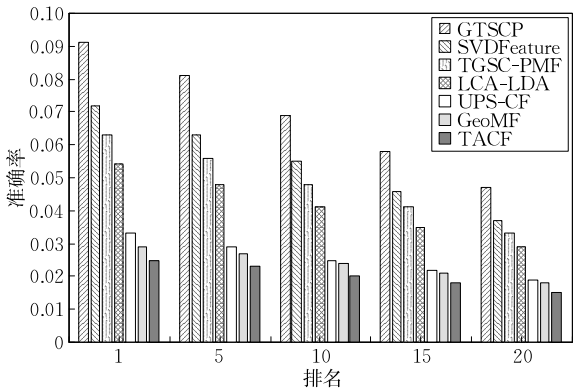


(b) 本地推荐的top-k召回率

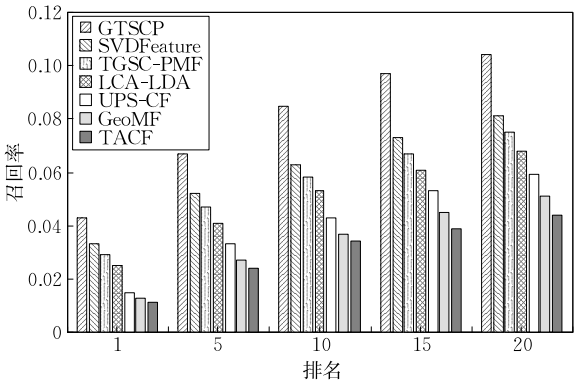
图 6　Foursquare 数据集上的本地推荐 top-k 性能

Twitter 数据集上的异地与本地推荐场景. 图 7 和图 8 显示了在 Twitter 数据集上的对比推荐模型的性能. 从图 7 和图 8 中，观察的结果与图 5 和图 6 的趋势相似. 主要的区别为在本文实验的 Twitter 数据集上的所有推荐方法的精确度比在 Foursquare 数据集上的所有推荐方法的精确度低. 这可能是由于用户在 Foursquare 数据集上的平均签到记录超过用户在 Twitter 数据集上的平均签到记录，能够使模型更加准确地捕捉用户的兴趣.

豆瓣数据集上的异地与本地推荐场景. 图 9 和图 10 显示了在豆瓣数据集上的对比推荐模型的性能. 从图 9 和图 10 中，观察的结果与图 5 和图 6 的趋势相似. 主要的区别为在本文实验的豆瓣数据集上的所有推荐方法的精确度比在 Foursquare 数据集上的所有推荐方法的精确度高. 这可能是由于用户在豆瓣数据集上的平均签到记录超过用户在

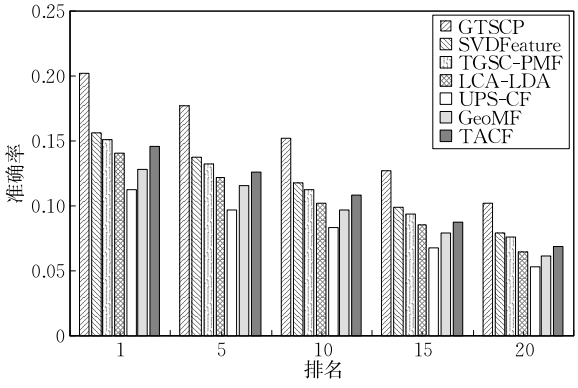


(a) 异地推荐的top-k准确率

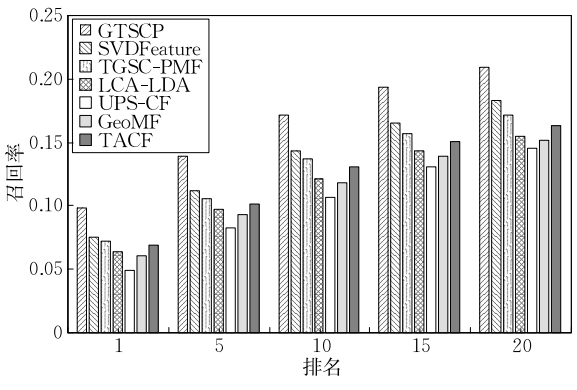


(b) 异地推荐的top-k召回率

图 7　Twitter 数据集上的异地推荐 top-k 性能

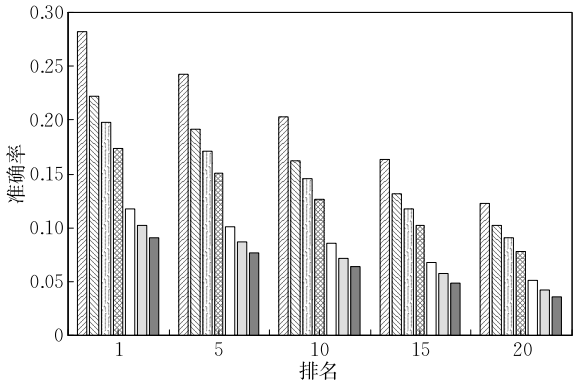


(a) 本地推荐的top-k准确率

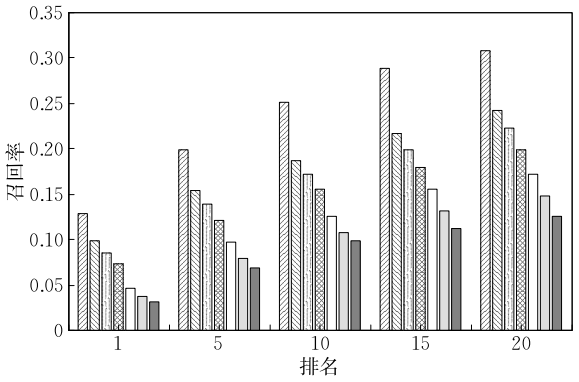


(b) 本地推荐的top-k召回率

图 8　Twitter 数据集上的本地推荐 top-k 性能

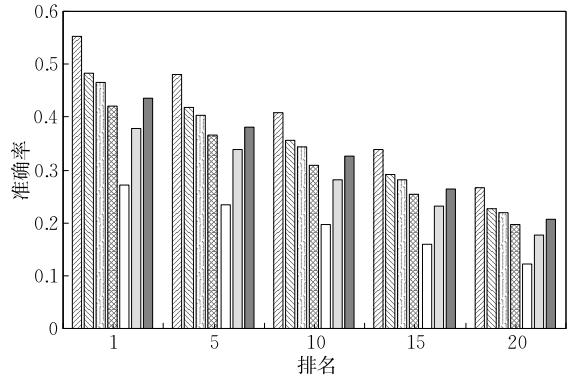


(a) 异地推荐的top-k准确率

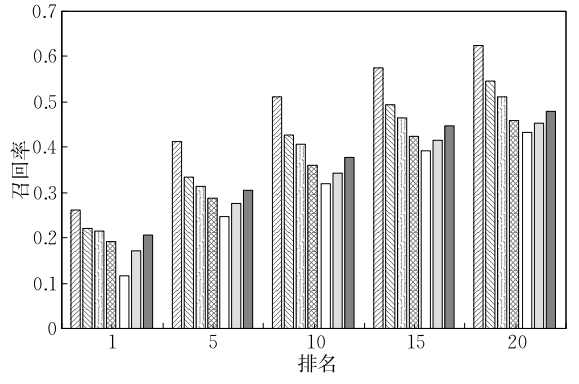


(b) 异地推荐的top-k召回率

图 9 豆瓣数据集上的异地推荐 top-k 性能



(a) 本地推荐的top-k准确率



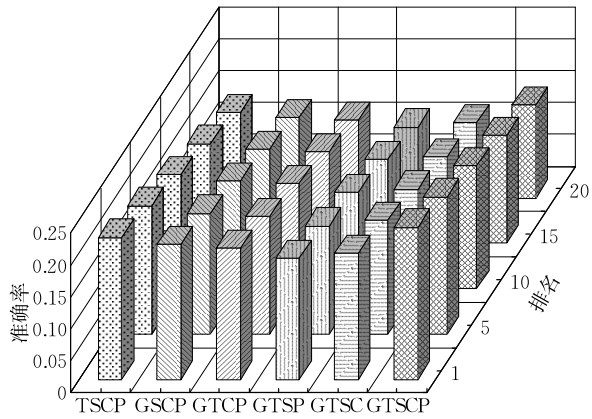
(b) 本地推荐的top-k召回率

图 10 豆瓣数据集上的本地推荐 top-k 性能

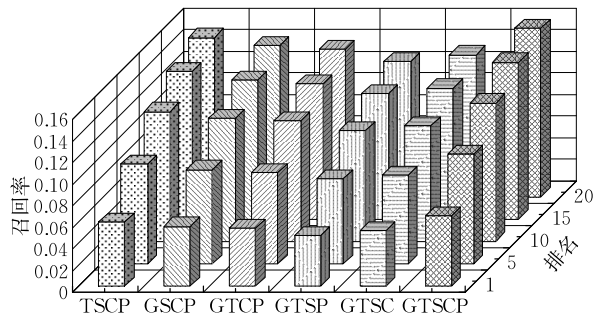
Foursquare 数据集上的平均签到记录,能够使模型更加准确地捕捉用户的兴趣.

5.2.2 不同因素的影响

为了分别研究 GTSCP 模型融合地理影响、时间效应、社会相关性、内容信息和流行度影响对我们所提的兴趣点推荐模型性能的影响.我们对比了我们所提的 GTSCP 模型的 5 个基线方法,TSCP、GSCP、GTCP、GTSP 和 GTSC.对比的结果如图 11 和图 12 所示.从结果中,我们首先观察到在异地和本地两个推荐场景中 GTSCP 始终优于 5 个基线方法,并表明在两个推荐场景中 GTSCP 受益于同时考虑 5 个因素来提高推荐精确度的程度是不同的.此外,另一个观察是同一个因素在两个不同的推荐场景中的影响程度是不同的.



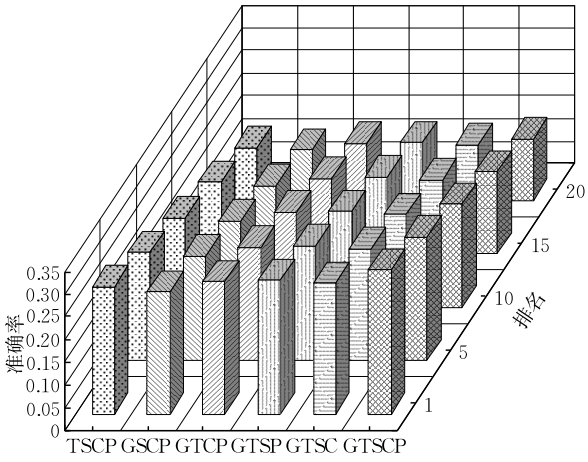
(a) 异地推荐的准确率



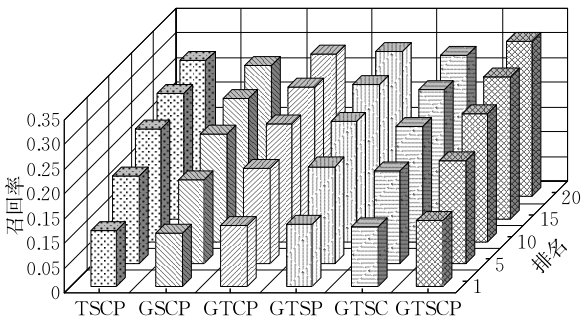
(b) 异地推荐的召回率

图 11 Foursquare 数据集上的不同因素影响异地推荐精度

特别是,在异地推荐场景的情况下,根据 5 个因素的重要性,它们的影响程度排列如下:内容影响>流行度影响>社会相关性>时间效应>地理影响;然而在本地推荐场景的情况下,它们的影响程度排列如下:时间效应>地理影响>流行度影响>社会相关性>内容影响.显然,内容信息在异地推荐场景中对于克服数据稀疏问题起着主导的作用,然而时间效应在本地推荐场景中却起着最重要的作用.这

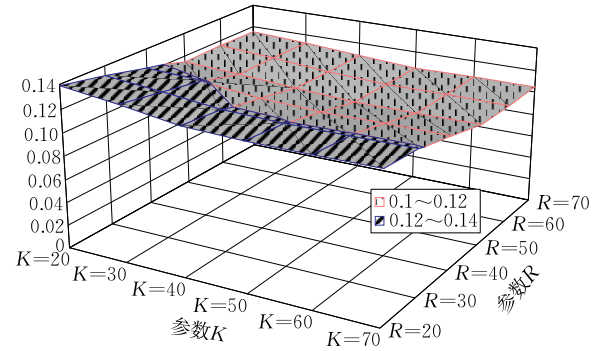


(a) 本地推荐的准确率

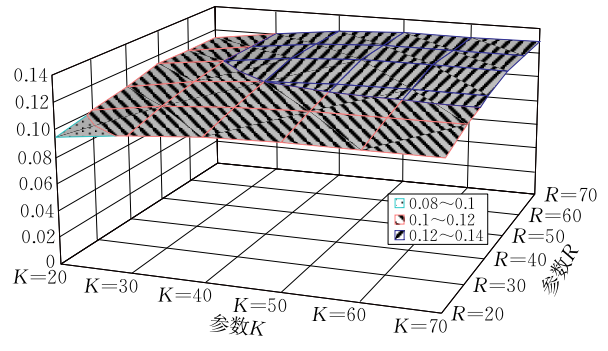


(b) 本地推荐的召回率

图 12　Foursquare 数据集上的不同因素影响本地推荐精度



(a) 异地推荐的准确率



(b) 异地推荐的召回率

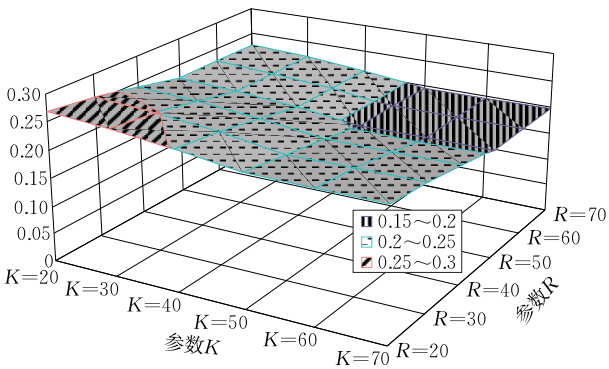
图 13　Foursquare 数据集上当 *Precision*@10 和 *Recall*@10 时异地推荐场景中参数 K 和 R 的影响

是因为两个推荐场景有不同的特征：(1) 在本地推荐场景中有足够的用户签到记录，然而在异地推荐场景中却有为数不多的用户签到记录；(2) 在本地推荐场景中的限制没有在异地推荐场景中的限制那么多；(3) 当用户在异地旅行时，用户的日常行为有可能会改变.

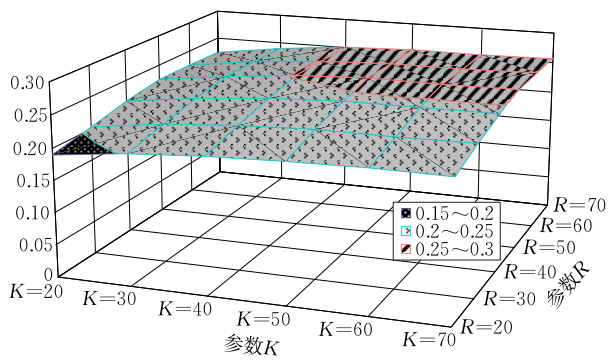
5.2.3　模型参数的影响

调整模型参数，即区域的数量 R 和主题的数量 K 对于 GTSCP 模型的性能是至关重要的. 因此，在这一部分我们在 Foursquare 数据集上研究模型参数的影响. 关于超参数 $\alpha,\alpha',\beta,\delta,\gamma,\tau,b$ 和 b'，我们设置固定值 $\alpha=\alpha'=50/K,\gamma=50/R,\beta=\delta=\tau=0.01,b=b'=0.5$. 我们尝试不同的设置，发现 GTSCP 模型的性能对这些超参数是不敏感的，但是 GTSCP 模型的性能对区域和主题的数量是敏感的. 因此，我们通过设置区域和主题的数量来测试 GTSCP 模型的性能，结果如图 13 和图 14 所示.

从结果中，首先我们观察到 GTSCP 的推荐准确率值随着区域的数量 R 增加而下降，而推荐召回率值随着区域的数量 R 增加而增加，然后当区域的数量大于 50 时推荐的准确率值和召回率值的变化



(a) 本地推荐的准确率

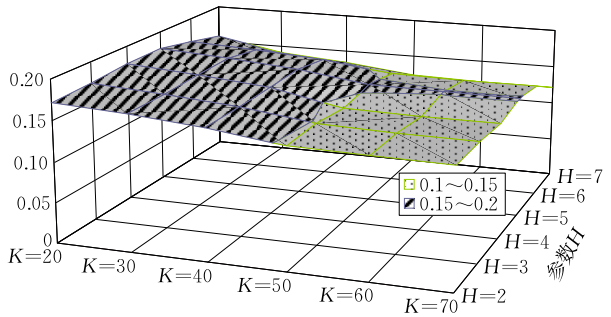


(b) 本地推荐的召回率

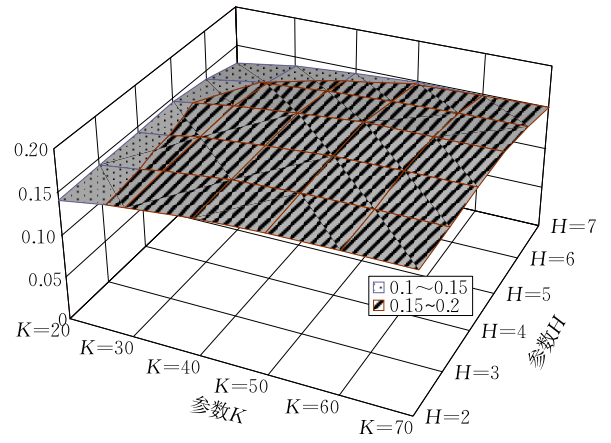
图 14　Foursquare 数据集上当 *Precision*@10 和 *Recall*@10 时本地推荐场景中参数 K 和 R 的影响

就不明显了.类似的观察,随着主题的数量 K 增加,GTSCP 的推荐准确率值下降而召回率值增加,当主题的数量大于 50 时推荐准确率值和召回率值的变化也不明显.原因是 R 和 K 代表模型的复杂度.因此,一方面,当 R 和 K 的值太小时,模型描述数据的能力受限;另一方面,当 R 和 K 的值超过阈值时,模型的复杂度足以处理数据.在这一点上,增加 R 和 K 的值对提高模型性能的帮助就不明显了.请注意的是,在我们实验的性能报告中已达到 50 个潜在区域(即 $R=50$)和 50 个潜在主题(即 $K=50$).

为了研究在空间平滑优化中的两个参数 H 和 K 的影响,我们尝试了这两个参数的不同设置.由于空间限制,我们只显示异地推荐场景中在 Foursquare 数据集上的实验结果.我们通过改变金子塔的高度 H 的值从 2 到 7 和主题的数量 K 的值从 20 到 70.结果如图 15 所示.



(a) 异地推荐的准确率



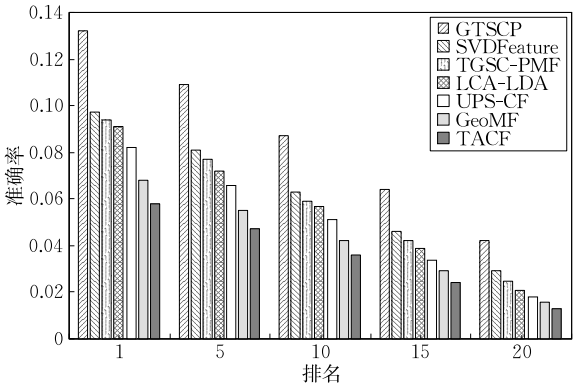
(b) 异地推荐的召回率

图 15 Foursquare 数据集上当 *Precision*@10 和 *Recall*@10 时异地推荐场景中参数 K 和 H 的影响

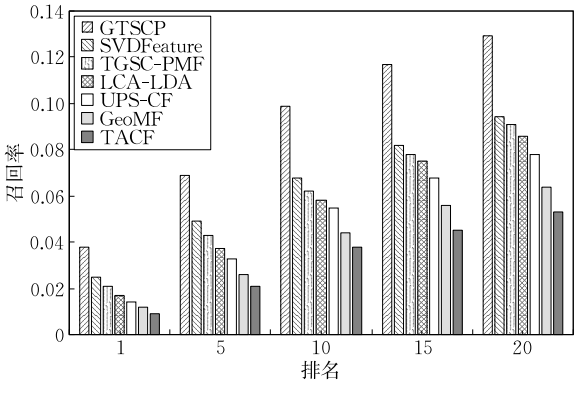
从结果中,我们观察到随着 H 值的增加,GTSCP 模型的准确率值和召回率值,都是先增加然后再下降.一个可能的原因是,早期准确率值和召回率值随着高度 H 值的增加而增加,是由于利用了空间影响,使得当地偏好的推断更具有预知性.然后,召回率值随着高度 H 值的增加而下降,是由于高度的增加使得用户活动数据在区域单元格中变得稀疏.当空间金子塔的高度设置为 5 时,为两个因素的权衡,GTSCP 模型达到最好的性能.另外,我们也观察到准确率值随着主题数量 K 值的增加而下降,而召回率值随着主题数量 K 值的增加而增加,当主题的数量大于 50 时推荐准确率值和召回率值的变化就不明显了.原因是 K 值代表模型的复杂度.因此,当 K 值太小时,模型描述数据的能力受限;然而当 K 值超过阈值(即 $K=50$)时,模型的复杂度足以处理数据.在这一点上,增加 K 值对提高模型性能的帮助就不明显了.因此,在 Foursquare 数据集上,我们选择 $H=5$,$K=50$ 作为模型效果与准确度之间最好的权衡.

5.2.4 冷启动问题测试

我们进一步进行实验,研究不同的推荐算法对解决冷启动问题的有效性.冷启动是推荐领域里的一个关键问题,即用户没有访问任何 POIs 或者只访问了极少数的 POIs.关于冷启动用户,我们在 Foursquare 数据集上测试了推荐效果,显示结果如图 16 和图 17 所示.图 16 显示了用户少于 5 次活动

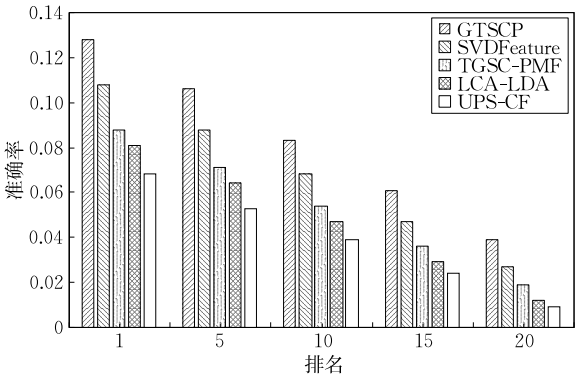


(a) 冷启动推荐准确率

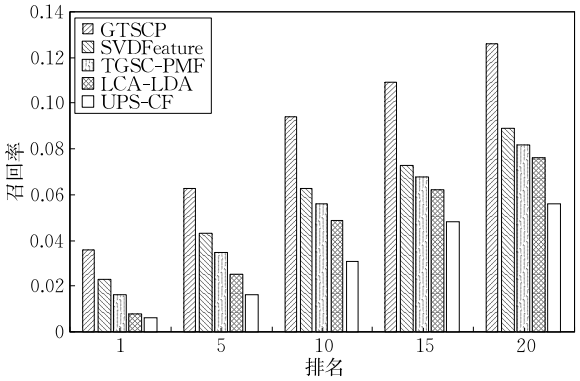


(b) 冷启动推荐召回率

图 16 Foursquare 数据集上关于冷启动用户少于 5 次活动记录的推荐效果



(a) 冷启动推荐准确率



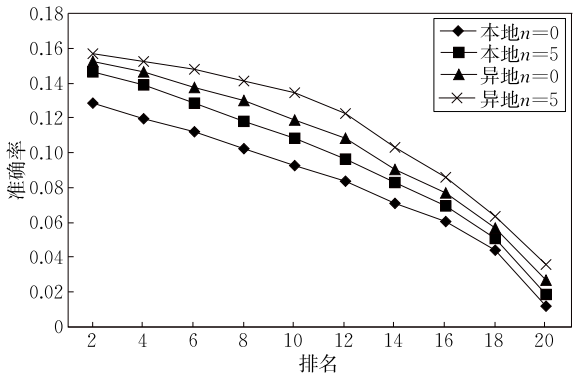
(b) 冷启动推荐召回率

图 17　Foursquare 数据集上关于冷启动用户没有任何活动记录的推荐效果

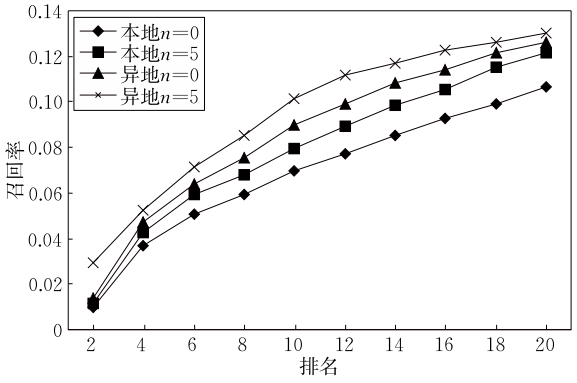
记录的结果，图 17 显示了用户没有任何活动记录的结果.

我们对比图 5、图 6 和图 16、图 17 的结果. 一些观察结果如下：(1) 尽管所有对比推荐方法的推荐效果不同程度地降低，但我们所提的 GTSCP 模型的性能依然是最好的. 值得注意的是，从目标用户访问数量极少的 POIs 中，很难捕捉冷启动用户的偏好. 这就是为什么只使用用户历史活动记录的 GeoMF 和 TACF 的性能最差. GTSCP，SVDFeature，TGSC-PMF，LCA-LDA 和 UPS-CF 分别利用群体的智慧和用户社会朋友克服了用户历史活动记录的数据稀疏问题. 群体的智慧和用户社会朋友为目标区域提供许多有价值的参考和线索，对兴趣点推荐起着潜在的作用；(2) GTSCP 的性能优于 TGSC-PMF，LCA-LDA 的性能是因为其区分了当地偏好和用户偏好，并且在目标区域中目标用户更有可能访问与其有相同角色的群体偏好的 POIs. 值得注意的是，在 Foursquare 数据集上的所有用户的本地信息是已知的. 因此，对于没有任何签到记录的用户，在目标区域中，GTSCP 模型同样可以识别目标用户是旅游者还是本地人，然后通过其他旅行者或者本地人偏好的 POIs 推荐给目标用户；(3) 另一个观察是 GTSCP，SVDFeature，TGSC-PMF 和 LCA-LDA 的性能优于 UPS-CF 的性能. 这是因为 UPS-CF 是一个基于记忆的方法，遭受了严重的数据稀疏问题，然而 GTSCP，SVDFeature，TGSC-PMF 和 LCA-LDA 是潜在分类模型通过减少维数与融合内容信息可以很大程度地处理数据稀疏问题；(4) GTSCP 达到的推荐精确度远高于 SVDFeature，尽管他们使用相同类型的特征和信息，显示了在冷启动问题的情况下概率生成模型同样优于基于特征的矩阵分解模型.

在本地和异地推荐场景中，关于冷启动用户，图 18 说明了 GTSCP 的性能. n 代表每个冷启动用户在训练集中的签到活动记录的最大数值. 从结果中，我们观察到关于冷启动用户，GTSCP 在异地推荐场景中比在本地推荐场中所达到的性能更好. 另外，当 n 值从 0 到 5 时，GTSCP 模型在本地推荐场景中比在异地推荐场景中得到一个更大的性能提升. 结果表明关于冷启动用户，在异地推荐场景中当地偏好对提高推荐效果所起的作用比在本地推荐场景中用户偏好对提高推荐效果所起的作用更重要. 这是因为用户的访问行为在本地时更个性化与多样



(a) 冷启动推荐准确率



(b) 冷启动推荐召回率

图 18　Foursquare 数据集上关于冷启动用户在本地和异地推荐场景的推荐效果

化,而用户的访问行为在外地时往往趋向于一致性.也就是说,一方面,不同的用户当在相同的异地区域旅行时趋向访问相同的 POIs;另一方面,结果显示:用户的个性化兴趣在本地推荐场景中比在异地推荐场景中的影响要大.这就是为什么本地推荐场景对于用户历史活动的数量更加敏感.

5.2.5 推荐效率

这部分实验是评估在线推荐的效率.关于 GTSCP 的在线推荐,我们利用两种方法.第 1 种方法称为 GTSCP-TA,利用基于 TA 的查询过程技术生成在线推荐.第 2 种方法称为 GTSCP-BF,利用一个朴素蛮力算法通过线性扫描生成一个 top-k 推荐.由于基于特征的矩阵分解模型 SVDFeature 可以无缝地融合文献[41]中所提的度量树检索算法,我们对比 GTSCP-TA 与 SVDFeature-MT.关于 SVDFeature 和 GTSCP 的潜在因素的数量设置为 120(即 $K=50$ 和 $R=70$).请注意在 SVDFeature 中的排名函数不是单调的,因此基于 TA 的查询过程技术不能应用于 SVDFeature 算法中.

关于在 Foursquare 数据集上的测试集 D_{test} 创建的所有查询,表 4 给出了 3 种不同方法的平均在线有效性.我们显示了当 k 设置为 1,5,10,15,20 时的性能.在 top-10 推荐时,GTSCP-TA 平均在 30.59ms.从结果中,我们观察到:(1) GTSCP-TA 显著优于 GTSCP-BF,证明了基于 TA 的查询处理技术带来的好处.在 top-10 推荐时,它只需要平均访问大约所有 POIs 的 18%,因为在查询偏好时它充分利用了稀疏性;(2) GTSCP-TA 的时间消耗成本随着推荐的数量 k 值的增加而增加,在推荐任务中,即使当 $k=20$ 时,它的时间消耗成本依然远低于 GTSCP-BF;(3) 在我们的推荐任务中,SVDFeature-MT 的时间成本高于朴素线性扫描方法 GTSCP-BF,尽管在设置 50 个潜在因素数量时,Koenigstein 等人[41]报导度量树检索算法的优良性能,但当 $k=10$ 时,它的时间消耗成本仍是朴素线性扫描方法的 1.9 倍.这个结果表明,当 POIs 有数以百计的维度或者属性时,度量树检索算法仍然不能克服维数多的困难.

表 4 在 Foursquare 数据集上的推荐效率

方法	在线推荐时间成本/ms				
	$k=1$	$k=5$	$k=10$	$k=15$	$k=20$
GTSCP-TA	10.22	19.02	30.59	38.39	45.63
GTSCP-BF	149.84	150.26	150.41	149.63	149.77
SVDFeature-MT	178.47	230.37	283.28	323.77	359.97

6 总 结

本文我们提出一个联合概率生成模型 GTSCP,建模基于位置的社交网络的用户签到行为,有效地融合地理影响、时间效应、社会相关性、内容信息和流行度影响这些方面的因素,有效地克服了数据稀疏问题和用户兴趣漂移问题,特别是在异地推荐场景中.然后通过空间金字塔的方法利用地理相关性对当地偏好进行平滑优化,进一步缓解了数据稀疏和用户兴趣漂移问题.我们所提的兴趣点推荐由离线模型和在线推荐两部分组成.为证明 GTSCP 的灵活性和适用性,我们研究了在一个联合的模型中是如何支持本地和异地两个推荐场景的.在真实的大规模数据集上,我们进行了大量的实验,在效果和效率两个方面对 GTSCP 模型的性能进行了评估.结果表明 GTSCP 的推荐精确度相比于其他当前先进的兴趣点推荐方法有了明显地提高.此外,我们在同一框架下,关于本地和异地两种推荐场景,研究了每个因素对改善推荐性能所起到的重要性,发现内容信息在异地推荐场景中起主导作用,而时间效应在本地推荐场景中却是最重要的.

本文做为我们未来工作的一部分,我们将进一步完善解决兴趣点特征缺失的问题,从而进一步提高兴趣点推荐的性能.

致 谢 特别感谢本文审稿专家和编辑所提出的宝贵意见和建议!

参 考 文 献

[1] Levandoski J J, Sarwat M, Eldawy A, Mokbel M F. LARS: A location-aware recommender system//Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. Vriginia, USA, 2012: 450-461

[2] Ference G, Ye M, Lee W C. Location recommendation for out-of-town users in location-based social networks//Proceedings of the 22nd ACM international conference on Information and Knowledge Management. Burlingame, USA, 2013: 721-726

[3] Yin H Z, Sun Y Z, Cui B, et al. Lcars: A location-content-aware recommender system//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 221-229

[4] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Journal of Knowledge and Data Engineering, 2005, 17(6): 734-749

[5] Ye M, Yin P, Lee W C. Location recommendation for location-based social networks//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose, USA, 2010: 458-461

[6] Lian D F, Zhao C, Xie X, et al. GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 831-840

[7] Gao H J, Tang J L, Hu X, Liu H. Exploring temporal effects for location recommendation on location-based social networks//Proceedings of the 7th ACM Conference on Recommender Systems. Hong Kong, China, 2013: 93-100

[8] Zheng Y, Xie X. Learning travel recommendations from user-generated GPS traces. ACM Transactions on Intelligent Systems and Technology, 2011, 2(1): 2:1-2:29

[9] Zheng Y, Zhang L Z, Xie X, Ma W Y. Mining interesting locations and travel sequences from GPS trajectories// Proceedings of the 18th International Conference on World Wide Web. Raleigh, USA, 2009: 791-800

[10] Ye M, Shou D, Lee W C, et al. On the semantic annotation of places in location-based social networks//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 520-528

[11] Bao J, Zheng Y, Wilkie D, Mokbel M. Recommendations in location-based social networks: A survey. GeoInformatica, 2015, 19(3): 525-565

[12] Cao Jiu-Xin, Dong Yi, Yang Peng-Wei, et al. POI recommendation based on meta-path in LBSN. Chinese Journal of Computers, 2016, 39(4): 675-684(in Chinese)
(曹玖新，董羿，杨鹏伟等. LBSN 中基于元路径的兴趣点推荐. 计算机学报，2016，39(4)：675-684)

[13] Jiang S H, Qian X M, Shen J L, et al. Author topic model-based collaborative filtering for personalized POI recommendations. IEEE Transactions on Multimedia, 2015, 17(6): 907-918

[14] Jiang S H, Qian X M, Mei T, Fu Y. Personalized travel sequence recommendation on multi-source big social media. IEEE Transactions on Big Data, 2016, 2(1): 43-56

[15] Qian X M, Feng H, Zhao G S, Mei T. Personalized recommendation combining user interest and social circle. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7): 1763-1777

[16] Zhao G S, Qian X M, Kang C. Service rating prediction by exploring social mobile users' geographical locations. IEEE Transactions on Big Data, 2016

[17] Zhao G S, Qian X M, Xie X. User-service rating prediction by exploring social users' rating behaviors. IEEE Transactions on Multimedia, 2016, 18(3): 496-506

[18] Ye M, Yin P F, Lee W C, Lee D L. Exploiting geographical influence for collaborative point-of-interest recommendation //Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 325-334

[19] Cheng C, Yang H Q, King I, Lyu M R. Fused matrix factorization with geographical and social influence in location-based social networks//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 17-23

[20] Cho E, Myers S A, Leskovec J. Friendship and mobility: User movement in location-based social networks//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 1082-1090

[21] Cheng C, Yang H, Lyu M R, King I. Where you like to go next: Successive point-of-interest recommendation//Proceedings of the 23th International Conference on Artificial Intelligence. Beijing, China, 2013: 2605-2611

[22] Yuan Q, Cong G, Ma Z Y, et al. Time-aware point-of-interest recommendation//Proceedings of the 36th ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 2013: 363-372

[23] Ferrari L, Rosi A, Mamei M, Zambonelli F. Extracting urban patterns from location-based social networks// Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. Chicago, USA, 2011: 9-16

[24] Gao H J, Tang J L, Hu X, Liu H. Content-aware point of interest recommendation on location-based social networks// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Astin, USA, 2015: 1721-1727

[25] Zhao K, Cong G, Yuan Q, Zhu K. Sar: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews//Proceedings of the 2015 IEEE 31st International Conference on Data Engineering. Seoul, South Korea, 2015: 675-686

[26] Ying J J C, Kuo W N, Tseng V S, Lu E H C. Mining user check-in behavior with a random walk for urban point-of-interest recommendations. ACM Transactions on Intelligent Systems and Technology, 2014, 5(3): 40:1-40:26

[27] Li X T, Cong G, Li X L, et al. Rank-GeoFM: A ranking based geographical factorization method for point of interest recommendation//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 433-442

[28] Liu B, Fu Y J, Yao Z J, Xiong H. Learning geographical preferences for point-of-interest recommendation//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 1043-1051

[29] Hu L K, Sun A X, Liu Y. Your neighbors affect your ratings: On geographical neighborhood influence to rating prediction//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast, Australia, 2014: 345-354

[30] Li R, Wang S J, Deng H B, et al. Towards social user profiling: Unified and discriminative influence model for inferring home locations//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 1023-1031

[31] Scellato S, Noulas A, Lambiotte R, Mascolo C. Socio-spatial properties of online location-based social networks//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 329-336

[32] Liu B, Xiong H. Point-of-interest recommendation in location based social networks with topic and location awareness//Proceedings of the SIAM International Conference on Data Mining. Austin, USA, 2013: 396-404

[33] Hu B, Ester M. Social topic modeling for point-of-interest recommendation in location-based social networks//Proceedings of the 2014 IEEE International Conference on Data Mining. Shenzhen, China, 2014: 846-850

[34] Lichman M, Smyth P. Modeling human location data with mixtures of kernel densities//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 35-44

[35] Wang X R, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Ljubljana, Slovenia, 2006: 424-433

[36] Shekhar S, Zhang P, Huang Y, Vatsavai R. Data Mining: Next Generation Challenges and Future Directions. USA: AAAI Press, 2004

[37] Yin H, Cui B, Chen L, et al. A temporal context-aware model for user behavior modeling in social media systems//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. Snowbird, USA, 2014: 1543-1554

[38] Cheng Z Y, Caverlee J, Lee K, Sui D Z. Exploring millions of footprints in location sharing services//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 81-88

[39] Chen T Q, Zhang W N, Lu Q X, et al. SVDFeature: A toolkit for feature-based collaborative filtering. Journal of Machine Learning Research, 2012, 13: 3619-3622

[40] Yin H Z, Cui B, Chen L, et al. Modeling location-based user rating profiles for personalized recommendation. ACM Transactions on Knowledge Discovery from Data, 2015, 9(3): 19:1-19:41

[41] Koenigstein K, Ram P, Shavitt Y. Efficient retrieval of recommendations in a matrix factorization framework//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui Hawaii, USA, 2012: 535-544

REN Xing-Yi, born in 1983, Ph. D. candidate. Her current research interests include data mining, recommendation system and big data.

SONG Mei-Na, born in 1974, Ph. D., professor. Her current research interests include service computing, cloud computing, very large scale information service system.

SONG Jun-De, born in 1938, Ph. D., professor. His current research interests include service science and engineering, cloud computing, big data, the Internet of Things and ICT key technologies.

Background

This paper belongs to social network and recommendation system area. With the rapid development of mobile devices, global position system (GPS) and Web 2.0 technologies, location-based social networks (LBSNs) have attracted millions of users to share rich information, such as geographical location, experiences and tips. Point-of-Interest (POI) recommender system plays an important role in LBSNs since it can help users explore attractive locations as well as help social network service providers design location-aware advertisements for Point-of-Interest. Memory-based and model-based collaborative filtering algorithms are normally very effective in traditional recommendation systems, such as movie recommendation and

goods recommendation. However, as a user can only visit a few POIs in LBSN, the check-ins of POI recommendation are scarce, traditional collaborative filtering recommendation methods easily suffer from the data sparsity problem. Thus, collaborative filtering (CF) techniques are unreliable for making effective POI recommendations.

Therefore, we propose a joint probabilistic generative model GTSCP to model users' check-in behaviors in LBSNs, which strategically integrates the factors of geographical influence, temporal effect, social correlation, content information and popularity impact to effectively overcome the problem of data sparsity and user interest drift, especially in out-of-town recommendation scenario. Then we exploit the geographical correlation by smoothing local preference through spatial pyramid to further alleviate the data sparsity and user interest drift. The POI recommendation method consists of two components offline modeling and online recommendation. To demonstrate the flexibility and applicability of GTSCP, we studied how it supports two recommendation scenarios in a joint model, home-town recommendation and out-of-recommendation.

This work is supported by Chinese Key Projects in the National Science and Technology Pillar Program No. 2014BAH26F02. We study on the recommendation system of insurance platform, which researches the construction and model recommender system for the insurance recommended business.