基于位置社交网络的上下文感知的兴趣点推荐

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)推荐为基于位置的服务提供了前所未有的机会.兴趣点推荐是一种基于上下文信息的位置感知的个性化推荐.然而用户-兴趣点矩阵的极端稀疏给兴趣点推荐的研究带来严峻挑战.为处理数据稀疏问题,文中利用兴趣点的地理、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,提出一种上下文感知的概率矩阵分解兴趣点推荐算法,称为TGSC-PMF.首先利用潜在狄利克雷分配(Latent Dirichlet Allocation,LDA)模型挖掘兴趣点相关的文本信息学习用户的兴趣话题生成兴趣相关分数;其次提出一种自适应带宽核评估方法构建地理相关性生成地理相关分数;然后通过用户社会关系的幂律分布构建社会相关性生成社会相关分数;另外结合用户的分类偏好与兴趣点的流行度构建分类相关性生成分类相关分数,最后利用概率矩阵分解模型(Probabilistic Matrix Factorization,PMF),将兴趣、地理、社会、分类的相关分数进行有效地融合,从而生成推荐列表推荐给用户感兴趣的兴趣点.该文在一个真实LBSN签到数据集上进行实验,结果表明该算法相比其他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;话题模型;地理相关性;社会相关性;分类相关性;社会媒体

中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2017.00824

Context-Aware Point-of-Interest Recommendation in Location-Based Social Networks

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

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

Abstract The rapid development of location-based social networks (LBSNs) has provided an unprecedented opportunity for better location-based services through Point-of-Interest (POI) recommendation. POI recommendation is a personalized, location-aware, and context depended recommendation. However, extreme sparsity of user-POI matrix creates a severe challenge. In this paper, we propose a context-aware probabilistic matrix factorization method called TGSC-PMF for POI recommendation, exploiting geographical information, text information, social information, categorical information and popularity information, incorporating these factors effectively. First, we exploit an aggregated Latent Dirichlet Allocation (LDA) model to learn the interest topics of users and infer the interest POIs by mining textual information associated with POIs and generate interest relevance score. Second, we propose a kernel estimation method with an adaptive bandwidth to model the geographical correlations and generate geographical relevance score. Third, we build social relevance through the power-law distribution of user social relations to generate social relevance score. Then, we model the categorical correlations which combine the category bias of users and the popularity of POIs into categorical relevance score. Further, we

收稿日期:2016-05-18;在线出版日期:2016-09-28.本课题得到国家科技重点支撑项目(2014BAK15B01)资助.任星怡,女,1983年生,博士研究生,主要研究方向为推荐系统、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.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、ICT关键技术.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

exploit probabilistic matrix factorization model (PMF) to integrate the interest, geographical, social and categorical relevance scores for POI recommendation. Finally, we implement experiments on a real LBSN check-in dataset. Experimental results show that TGSC-PMF achieves significantly superior recommendation quality compare to other state-of-the-art POI recommendation techniques.

Keywords　location-based social network; point-of-interest recommendation; topic model; geographical correlations; social correlations; categorical correlations; social media

1 引　言

随着城市的快速发展,兴趣点(如商场、餐厅、博物馆、娱乐场所、酒店、旅游景点等)的数量也随之增长,它为人们提供了更多体验生活的机会.在日常生活中,人们通常喜欢探索居住城市与邻近的地方,根据自己的个人兴趣选择与自己偏好相关的兴趣点.由于兴趣点与用户偏好的数据中包含大量有价值的信息,可以用于兴趣点推荐中[1].同时在大量的兴趣点中如何有效地帮用户做出满意的决策是一个困难的问题,通常被认为"选择麻痹".为了解决这个问题,兴趣点推荐任务将帮助用户过滤掉不感兴趣的位置并减少决策时间[2].

基于位置服务应用的日益流行,关于空间、时间、社会和内容等方面的兴趣点推荐,基于位置的社交网络为研究人们移动行为提供了前所未有的机会[3].典型的基于位置的社交网站,如国外的 Foursquare、Yelp、Gowalla,国内的街旁、嘀咕等,人们可以使用智能手机、平板电脑等移动设备对当前访问的兴趣点签到,并与好友分享自己的签到信息和体验,导致"W4"的信息布局(即什么人、什么地点、什么时间、什么事件),对应 4 个不同层次的信息[4].的确,兴趣点推荐服务旨在为用户推荐一些新的感兴趣的位置,基于位置社交网络的兴趣点推荐为人们提供更好的定位服务起着重要的作用.

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地点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.首先,用户-兴趣点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只占有非常小的比例,因此兴趣点推荐面临数据稀疏性问题[5].其次,随着不同的时间与不同的地理位置,用户的兴趣是动态变化的.然后,兴趣点推荐包含不同类型的上下文信息,如兴趣点的文本信息、兴趣点的地理坐标、用户的签到时间、用户的社会关系、兴趣点的分类信息、兴趣点的流行度等,与传统推荐不同的是兴趣点相关的文本信息是不完整的且模糊的.

依据上述挑战,本文提出一种上下文感知的概率矩阵分解兴趣点推荐算法,并结合兴趣话题、地理相关性、社会相关性与分类相关性.

本文中我们的贡献总结如下:

(1)本文利用兴趣点的地理、文本、社会、分类与流行度信息,并有效地融合这些因素,提出一种上下文感知的兴趣点推荐算法,称为 TGSC-PMF.

(2)本文利用主题模型挖掘兴趣点相关的文本信息,学习用户的兴趣话题;提出一种自适应带宽核评估方法确定用户的个性化兴趣点签到分布,构建兴趣点之间的地理相关性;通过用户社会关系的幂律分布构建用户之间的社会相关性;结合用户的分类偏好与兴趣点的流行度构建分类相关性.

(3)本文提出一种分数匹配方法,将兴趣相关分数、地理相关分数、社会相关分数以及分类相关分数进行有效地匹配,然后将匹配后的偏好分数融合到概率矩阵分解模型中,从而提出一种新的上下文感知的概率矩阵分解算法进行兴趣点推荐.

(4)本文使用一个真实的 LBSN 签到数据集进行大量的实验评估 TGSC-PMF 的推荐效果,实验结果证明 TGSC-PMF 优于其他先进的兴趣点推荐技术.

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 TGSC-PMF 兴趣点推荐算法;第 4 节介绍实验的方案设计与性能对比,验证该算

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

法的有效性;第 5 节总结并探讨将来的研究工作.

2 相关工作

2.1 兴趣点推荐

随着基于位置社交网络的快速发展,兴趣点推荐可为人们提供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.Ye 等人[2]关于兴趣点推荐采用线性插值的方法结合地理与社会影响应用到基于用户的协同过滤框架中.Levandoski 等人[6]考虑旅行的距离并扩展基于项目的协同过滤方法.用户-用户和项目-项目之间的相似度需要共享的签到数据来计算,并且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.Liu 等人[7]基于贝叶斯非负矩阵分解结合地理影响与文本信息提出一种地理概率因素分析框架.但是在他们的工作中,只考虑了显示反馈.最近,考虑关于签到数据的隐式反馈,Lian 等人[8]提出一种结合地理影响的加权矩阵分解方法.另外,Cao 等人[9]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,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.

2.2 基于文本信息的兴趣点推荐

为了更好地理解 LBSN 的模式并改善其服务,当前更多的研究开始探索文本信息.一些研究采用话题模型或者地理潜在因素获取区域或者 POIs 的潜在特征[10-13].Farrahi 等人[14]应用话题模型挖掘移动手机的文本数据来识别人们日常位置驱动的行程.关于基于位置的社交网络,Ye 等人[15]利用个体位置的显示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研究.Yin 等人[16]利用位置与位置相关的文本提出一种潜在地理话题分析方法有利于发现有意义的地理话题.Ferrari 等人[17]分析 Twitter 上的帖子并利用话题模型提取城市模式,例如热点地区和人群行为.这些关于探索文本信息的兴趣点推荐研究,一种直接的方法是结合话题模型的协同过滤技术.Agarwal 等人[18]利用每个项目相关的词语和用户特征来正则化项目因素和用户间的相关评分.Pennacchiotti 等人[19]利用话题模型研究社会媒体用户的兴趣并推荐给与用户兴趣相似的新朋友.由于兴趣点相关的文本信息通常是不完整的且模糊的,本文利用文本信息并采用话题模型来处理这个问题.

2.3 基于地理信息的兴趣点推荐

事实上地理邻近性显著地影响用户在兴趣点上的签到行为,地理信息被集中用于兴趣点推荐.一种方法是简单地考虑用户当前的位置,过滤离用户较远的 POIs[20-23].另一种方法是应用地理潜在特征或者主题模型推断区域或者 POIs 的潜在特征[8,10-24].更复杂的方法是评估签到过的 POIs 的地理相关性作为所有用户共同的距离分布,即一种多中心高斯分布[25]、一种幂律分布[2,14,26-30]或者一种关于每个用户的个性化非参数分布[31].特别是,关于每个用户的地理经纬度坐标,当前工作[32-33]采用固定带宽核密度评估方法建模兴趣点的地理签到分布.更进一步,本文提出一种自适应带宽核评估方法加强已获取的签到分布的能力,预测一个用户与一个未签到的 POI 之间的相关分数.

2.4 基于社会信息的兴趣点推荐

利用用户之间的社会关系可以提高基于位置社交网络的兴趣点推荐系统的质量.因为在 POIs 上社会朋友比陌生人更有可能分享共同的偏好.当前大部分研究是从用户之间的社会关系中获取相似度,并将其与传统的基于记忆或者基于模型的协同过滤技术相结合.例如,一些文献[21-22, 31-34]将用户的相似性无缝连接到基于用户的协同过滤技术中,然而其他一些研究[25,35-36]利用潜在因素模型的权重或者用户之间的相似性作为正则化项.本文利用用户之间的社会相关性,聚集 POIs 上用户朋友的签到频率或者评价,基于所有用户历史签到数据来评估社会签到频率或者评价的分布,并将其转换成社会相关分数.

2.5 基于分类信息的兴趣点推荐

用户访问过的 POIs 的分类信息隐式地显现了 POIs 上的用户行为.利用 POIs 的分类信息构建用户的特别偏好是有用的.然而,关于兴趣点推荐只有少量研究利用分类信息.Hu 等人[24]利用矩阵分解技术结合每个分类的潜在向量,基于 POI 的分类潜在向量推算用户对 POI 的相关分数.Liu 等人[27]将 POIs 的分类聚类成组,从用户历史签到数据中构建用户-分类转换矩阵替代用户-POI 签到矩阵,应用矩阵分解技术发现下一个用户可能签到的 top-k 分类.Ying 等人[34]通过 POIs 上标注的标签获取 POIs

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的分类权重,基于分类的偏好与权重的内积评估用户与 POIs 间的相关分数. Rahimi 等人[37]通过 POI 的分类信息获取用户的偏好,以此简单识别用户对 POI 的喜好. Zhao 等人[38]聚类用户到各个社区中,每个社区作为加权的分类向量,通过社区中的用户,每个维度代表一个特定的 POI 分类的签到数量,应用基于用户的协同过滤方法并利用用户所在社区的分类向量进一步计算用户间的相似性.

2.6　基于流行度信息的兴趣点推荐

POIs 的流行度反映了 POIs 所提供的服务和产品的质量. 在兴趣点推荐中利用 POIs 的流行度是有用的. 当前大多数研究认为 POIs 的流行度为用户对 POIs 的普遍先验喜好. Ying 等人[34]对于未签到的 POIs,在完全二部图中,利用用户的先验喜好作为用户与 POIs 之间的加权边. 其它一些研究工作[5,7,24,29]利用地理信息获取先验喜好调整后验喜好. 然而,一方面,在这些研究中,先验喜好不是个性化的用户偏好. 因此,在实践中得益于 POIs 的流行度是有限的. 另一方面,当前一些研究[20,24,27,34,37]分开地构建分类的影响和 POIs 的流行度,因此,在这种情况下,关于兴趣点推荐的分类和流行度信息可能不会被充分利用. 本文提出一种融合用户的分类偏好和 POIs 的流行度,并将其转换成用户与 POI 间的分类相关分数.

3　上下文感知的兴趣点推荐

3.1　问题陈述与模型框架

本节定义数据结构、阐述研究问题并展示模型框架. 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 的文本信息、POIs 的地理信息、用户的社会信息、POIs 的分类信息以及POIs的流行度. 表 1 列出本文的关键符号.

表 1　本文中的关键符号

符号	意义
U	在 LBSN 上所有用户的集合
u_i	某用户:$u_i \in U$
L	在 LBSN 上所有 POIs 的集合
l_j	某 POI:$l_j \in L$ 具有一对经纬度地理坐标(x_j, y_j)
C	在 LBSN 上所有 POIs 的分类的集合
c_g	某分类:$c_g \in C$
W	文本相关的唯一词的集合
w_i	某唯一词:$w_i \in W$
$\boldsymbol{R}_{\lvert U\rvert\times\lvert L\rvert}$	签到矩阵
$\boldsymbol{S}_{\lvert U\rvert\times\lvert U\rvert}$	社会关系矩阵
$\boldsymbol{C}_{\lvert U\rvert\times\lvert C\rvert}$	分类偏好矩阵
$\boldsymbol{P}_{\lvert C\rvert\times\lvert L\rvert}$	流行度矩阵

为了便于说明,$U=\{u_1, u_2, \cdots, u_M\}$ 为用户的集合,M 代表用户的数量. $L=\{l_1, l_2, \cdots, l_N\}$ 为 POIs 的集合,N 代表 POIs 的数量. $C=\{c_1, c_2, \cdots, c_B\}$ 为分类的集合,B 代表分类的数量. $W=\{w_1, w_2, \cdots, w_V\}$ 为文本信息相关的所有唯一词的集合,V 是唯一词的数量. r_{u_i,l_j} 为用户 u_i 在 POI l_j 上的签到频率或者评价. d_{l_j} 为与 POI l_j 相关的文本项目. d_{u_i} 为与用户 u_i 签到过的 POIs 相关的文本项目.

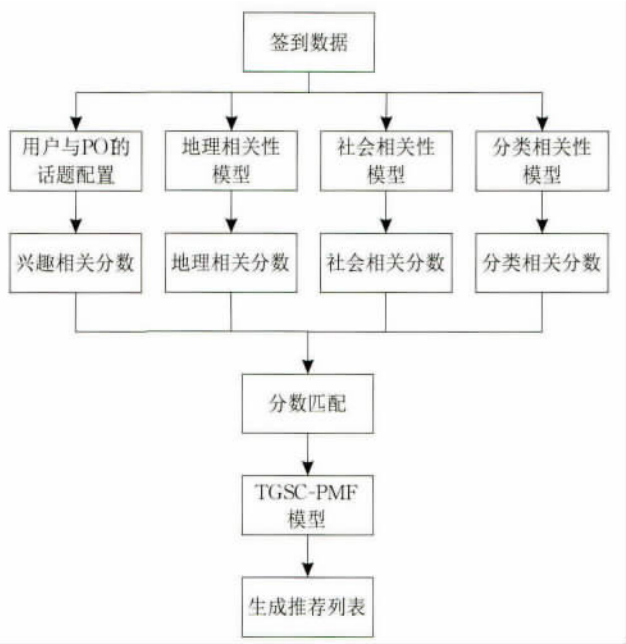


图 1　模型框架

定义 1.　签到矩阵. 给定一个 LBSN 上的 POIs 的用户历史签到数据,构建一个签到矩阵 $\boldsymbol{R}_{|U|\times|L|}$,矩阵中的每个元素 r_{u_i,l_j} 代表用户 $u_i \in U$ 在位置 $l_j \in L$ 上的签到频率或者评价,U 和 L 分别是 LBSN 上的用户和 POIs 的集合.

定义 2.　社会关系矩阵. 给定一个 LBSN 上的用户之间的社会关系,构建一个社会关系矩阵 $\boldsymbol{S}_{|U|\times|U|}$,如果在两个不同的用户 $u_i, u_i' \in U$ 之间存在社会关系,则 $s_{u_i,u_i'}=1$;否则,$s_{u_i,u_i'}=0$.

定义 3.　分类偏好矩阵. 给定一个 LBSN 上的 POIs 的用户历史签到数据与 POIs 的分类信息,构建一个分类偏好矩阵 $\boldsymbol{C}_{|U|\times|C|}$,矩阵中的每个元素 c_{u_i,c_g} 代表用户 u_i 访问属于分类 $c_g \in C$ 的 POIs 的频率,C 是 POIs 的分类集合,通常在 LBSN 上是预先定义的. 请注意一个 POI 可以属于多个分类.

定义 4.　流行度矩阵. 给定一个 LBSN 上的 POIs 的用户历史签到数据,构建一个流行度矩阵 $\boldsymbol{P}_{|C|\times|L|}$,矩阵中的每个元素 p_{c_g,l_j} 代表签到频率或者所有用户在 POI l_j 上的总评价,即在分类 $c_g \in C$ 上

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

的 POI l_j 的流行度.

定义 5. 地理坐标. 一个 POI $l_j \in L$ 是与一对地理经纬度坐标(x_j, y_j)相关的.

本文模型框架如图 1 所示. (1) 用户与 POI 的话题配置. 该方法有效地利用 POI 相关的文本信息及上下文信息,更好地配置用户与 POI 之间的话题模型;(2) 地理相关性模型. 给定一个用户访问过的 POIs,首先评估用户所在位置经纬度坐标的个性化签到分布,然后构建用户已访问的 POIs 和未签到的 POIs 之间的地理相关性,最后计算用户对任一未签到的 POI 的地理相关分数;(3) 社会相关性模型. 给定一个未签到的 POI,首先聚集用户朋友的社会签到频率或者评价,然后基于所有用户历史签到数据评估社会签到频率或者评价的分布,最后将其转化成用户对未签到的 POI 的社会相关分数;(4) 分类相关性模型. 首先从用户访问过的 POIs 的分类标签中获取用户的偏好,其次利用用户的偏好在相对应的分类标签中对未签到的 POI 的流行度进行加权,基于所有用户历史签到数据评估流行度的分布,然后将用户对未签到的 POI 的加权流行度映射成分类相关分数,因此分类相关性即考虑所有 POIs 的流行度又考虑所有 POIs 的分类信息从而有利于兴趣点推荐,其表示 POIs 的质量;(5) 分数匹配. 首先兴趣相关分数是关于话题的 POI 的兴趣匹配用户的个性化兴趣话题;其次,根据用户签到的所有 POIs 的地理坐标,构建 POIs 之间的地理相关性,利用地理相关性生成某用户对某未签到的 POI 的地理相关分数;然后在用户朋友已访问的 POIs 之间,利用社会相关性生成某用户对某未签到的 POI 的社会相关分数;进一步在用户已访问的 POIs 与未签到的 POI 之间的分类与流行度中,利用分类相关性生成某用户对某未签到的 POI 的分类相关分数;最后对兴趣相关分数、地理相关分数、社会相关分数与分类相关分数进行匹配生成偏好分数;(6) TGSC-PMF 模型. 将匹配后的偏好分数融合到概率矩阵分解模型中,从而提出一种新的上下文感知的概率矩阵分解算法进行兴趣点推荐生成推荐列表.

3.2 配置用户与 POI 的话题模型

3.2.1 话题提取

本文话题提取的目的是基于用户签到的 POIs 的文本信息,基于 LDA 算法[39]. 关于兴趣点推荐,我们提出一种聚合 LDA 模型学习用户的兴趣. 通过话题分布提取用户和 POI 的配置文件. 我们构建一个聚合 LDA 模型如图 2 所示. 我们聚集与同一 POI 有关的所有文本评论到一个 POI 文档即 d_{l_j},同样我们聚集同一用户签到过的 POIs 的所有文本评论到一个用户文档即 d_{u_i}. 这样我们获得一个大量的文档集合,每一个文档对应一个 POI 或者一个用户. 此模型有两个潜在变量:(1) 文档-话题分布 $\boldsymbol{\Theta}$;(2) 话题-词语分布 $\boldsymbol{\Phi}$. 我们可以从用户感兴趣的话题以及与这些话题相关的 POIs 的文本评论中获取信息.

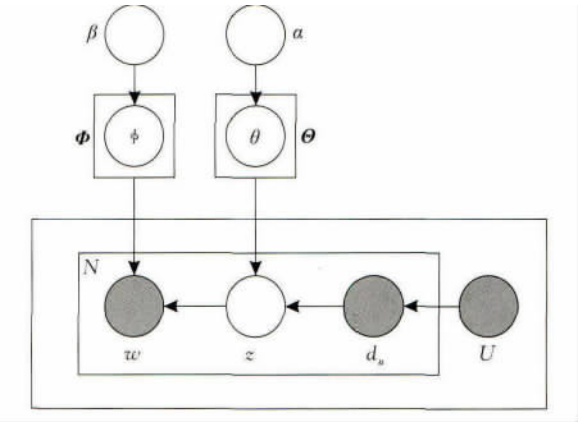


图 2 聚合 LDA 模型

本质上每一个用户或者 POI 是由话题的多项分布所代表,在统一的话题模型框架中,每一个话题与文本词语的多项分布相关. 因此,d_{u_i} 的话题代表用户 u_i 的兴趣话题. 用户 u_i 的兴趣话题多项分布表示为 θ. 每个兴趣话题的文本项目多项分布表示为 ϕ.

聚合 LDA 的生成过程如下:

1. 针对每个话题 $z \in \{1, \cdots, K\}$,提取一个文本项目多项分布,$\phi_z \in Dir(\beta)$.

2. 针对用户 u_i 的文档 d_{u_i}

(1) 提取一个兴趣话题分布 $\theta_{d_{u_i}} \in Dir(\alpha)$

(2) 针对文档 d_{u_i} 的每个词语 $w_{d,n}$:

① 提取一个话题 $z_{d,n} \sim Mult(\theta_{d_{u_i}})$

② 提取一个词语 $w_{d,n} \sim Mult(\phi_{z_{d,n}})$

于是构成了矩阵 $\boldsymbol{\Theta}_{M\times K}$ 和矩阵 $\boldsymbol{\Phi}_{K\times V}$. 基于学习后的矩阵 $\boldsymbol{\Theta}_{M\times K}$ 和矩阵 $\boldsymbol{\Phi}_{K\times V}$,进一步推断用户 u_i 的兴趣话题分布 θ_i 和 POI l_j 的话题分布 π_j. 因此,我们可以计算话题的相似性.

3.2.2 模型参数学习

在聚合 LDA 模型中,有两个未知的兴趣参数集合:文档-话题分布 $\boldsymbol{\Theta}$ 和话题-词语分布 $\boldsymbol{\Phi}$. 潜在变量 z 对应唯一词到话题中的配置. 如图 2 所示,给定两个超参数 α 和 β,所有用户文档的完全似然模型如下:

$$p(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\Phi}|\alpha,\beta)=\prod_{m=1}^{M}\prod_{n=1}^{N_m} p(w_{m,n}|\phi_{z_{m,n}})\cdot p(z_{m,n}|\theta_m)\cdot p(\theta_m|\alpha)\cdot p(\boldsymbol{\Phi}|\beta) \quad (1)$$

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

请注意如式(1)所示,在聚合 LDA 的完全似然模型中,直接评估 $\boldsymbol{\Theta}$ 和 $\boldsymbol{\Phi}$ 是很难计算的. 在参数评估的过程中,我们只需要对矩阵 $\boldsymbol{\Theta}_{M\times K}$ 和矩阵 $\boldsymbol{\Phi}_{K\times V}$ 保持跟踪. 对于这些矩阵,我们使用吉布斯采样[40]来评估话题-词语分布与用户-话题分布. 首先需要采样潜在变量 z 的条件分布如下:

$$p(z_i=k\mid \boldsymbol{w}_i=w_i,\boldsymbol{z}_{-i},\boldsymbol{w})\propto\frac{n_{k,-i}^{(w_i)}+\beta}{n_{k,-i}^{(\cdot)}+V\beta}\cdot\frac{n_{d_i,-i}^{(k)}+\alpha}{n_{d_i,-i}^{(\cdot)}+K\alpha} \tag{2}$$

计数 $n_{\cdot,-i}^{(\cdot)}$ 表明排除项目 i 对应的话题或者文档.

伴随着采样结果,我们使用 $\theta_{ik}=\dfrac{n_i^{(k)}+\alpha}{\sum_{k=1}^{K}n_i^{(k)}+K\alpha}$ 和 $\phi_{kw}=\dfrac{n_k^{(w)}+\beta}{\sum_{w=1}^{V}n_k^{(w)}+V\beta}$ 来评估 θ 和 ϕ,$n_k^{(w)}$ 是关于话题 k 的词配置频率,$n_i^{(k)}$ 是关于用户 u_i 的文档 d_{u_i} 的话题观察计数. V 是唯一词的数量,K 是话题的数量. 这里我们设置 α 和 β 是两个对称的先验.

接下来,给定训练模型 $M:\{\boldsymbol{\Theta},\boldsymbol{\Phi}\}$ 和超参数 α 和 β,根据一个 POI 的文档 d_{l_j} 推导话题分布 $p(\pi_j\mid d_{l_j},M)$. 类似上述的聚合 LDA 模型的参数评估,我们同样使用吉布斯采样方法提取每个 POI 的话题分布. 吉布斯采样的完全条件分布如下:

$$p(z_{d_{l_j}}=k\mid \boldsymbol{w}_i=w_i,\boldsymbol{z}_{-i},\boldsymbol{w}_{-i},M)\propto(n_{d_{l_j},-i}^{(k)}+\alpha) \tag{3}$$

然后,POI l_j 的文档 d_{l_j} 的话题分布是 $\pi_{jk}=\dfrac{n_j^{(k)}+\alpha}{\sum_{k=1}^{K}n_j^{(k)}+K\alpha}$,$n_j^{(k)}$ 是文档 d_{l_j} 的话题观察计数.

3.2.3　兴趣相关分数

我们定义用户 u_i 和 POI l_j 之间的兴趣相关分数作为用户话题分布 θ_i 和 POI 话题分布 π_j 的相似性. 通过兴趣相关分数计算 POI 的兴趣与用户的个性化兴趣的匹配程度. 我们使用 Jensen-Shannon 散度测量上述两个多项话题分布之间的相似性. 用户 u_i 和 POI l_j 之间的对称 Jensen-Shannon 散度如下:

$$D_{JS}(u_i,l_j)=\frac{1}{2}D(\theta_i\parallel M)+\frac{1}{2}D(\pi_j\parallel M) \tag{4}$$

$M=\dfrac{1}{2}(\theta_i+\pi_j)$ 和 $D(\cdot\parallel\cdot)$ 是 Kullback-Leibler 距离. 兴趣相关分数定义如下:

$$S(u_i,l_j)=1-D_{JS}(u_i,l_j) \tag{5}$$

我们在配置用户与 POI 的话题模型中采用兴趣相关分数模型的目的是通过话题提取与参数学习的过程,获取用户对 POIs 的兴趣偏好生成兴趣相关分数,为了能够更好地与本文接下来所提的地理相关分数、社会相关分数与分类相关分数进行分数匹配生成偏好分数,从而更加有效地融合兴趣点推荐的文本、地理、社会、分类与流行度信息.

3.3　兴趣点推荐的地理相关性模型

POIs 的地理邻近性在用户签到行为中起着重要的作用. 换言之,邻近的 POIs 比偏远的 POIs 的地理相关性要强. 因此,我们利用用户已签到的 POIs 和用户未签到的 POIs 之间的地理相关性来评估用户对未签到的 POI 的地理相关分数. 为构建 POIs 之间的地理相关性,基于每个用户签到过的 POIs,我们在地理坐标上评估个性化签到分布. 我们采用核带宽到每个签到数据点,并且从底层的签到数据中可以学习出自适应带宽. 自适应核评估方法包括 3 个步骤:试点估计、当地带宽决策、自适应核评估地理相关分数.

3.3.1　试点估计

首先,我们基于固定带宽核密度估计发现一个试点估计. 让 $L_u=\{l_1,l_2,\cdots,l_n\}$ 为用户 u_i 签到过的 POIs 的集合. L_u 中的每个 POI l_j 都与一对经纬度坐标 (x_i,y_i) 相关. 特别是,我们利用用户 u_i 在 POI l_i 上的签到频率或者评价(即 r_{u_i,l_i}),作为 l_j 的权重,因为一个 POI 的签到频率或者评价高就暗示着此 POI 对用户更重要. 用户 u_i 在一个未签到的 POI l_j 上的签到分布的试点估计给定如下:

$$f_{Geo}(l_j\mid u_i)=\frac{1}{2}\sum_{i=1}^{n}(r_{u_i,l_i}\cdot Q_H(l_j-l_i)) \tag{6}$$

$$A=\sum_{i=1}^{n}r_{u_i,l_i} \tag{7}$$

$$Q_H(l_j-l_i)=\frac{1}{2\pi H_1H_2}\exp\left(-\frac{(x_j-x_i)^2}{2H_1^2}-\frac{(y_j-y_i)^2}{2H_2^2}\right) \tag{8}$$

$Q_H(l_j-l_i)$ 是包含两个全局带宽 (H_1,H_2) 的固定带宽 H 的标准内核函数,两个全局带宽 (H_1,H_2) 给定如下:

$$H_1=1.08n^{-\frac{1}{5}}\sqrt{\frac{1}{A}\sum_{i=1}^{n}\left(r_{u_i,l_i}\cdot x_i-\frac{1}{A}\sum_{k=1}^{n}r_{u_i,l_k}\cdot x_k\right)^2} \tag{9}$$

$$H_2=1.08n^{-\frac{1}{5}}\sqrt{\frac{1}{A}\sum_{i=1}^{n}\left(r_{u_i,l_i}\cdot y_i-\frac{1}{A}\sum_{k=1}^{n}r_{u_i,l_k}\cdot y_k\right)^2} \tag{10}$$

(H_1,H_2) 从用户 u_i 的签到数据中被分别计算成经度值与纬度值的标准偏差.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

3.3.2 当地带宽决策

进一步，不是直接使用式(6)中的试点估计 $f_{Geo}(l_j|u_i)$来预测用户 u_i 到 POI l_j 的相关分数，而是我们利用试点估计来决策每个签到的 POI l_i 的自适应当地带宽 h_i 给定如下：

$$h_i=(d^{-1}\cdot f_{Geo}(l_i|u_i))^{-\tau} \tag{11}$$

τ 是敏感参数 $0\leq\tau\leq1$，即参数 τ 越大，自适应当地带宽 h_i 对试点估计 $f_{Geo}(l_i|u_i)$ 越敏感，d 是几何平均值如下：

$$d=\sqrt[n]{\prod_{i=1}^{n} f_{Geo}(l_i|u_i)} \tag{12}$$

强制约束 $h_i(i=1,2,\cdots,n)$的几何平均值为 1.

3.3.3 自适应核评估地理相关分数

最后，根据式(9)和式(10)的全局带宽 $H=(H_1,H_2)$和式(11)的自适应当地带宽 h_i，则用户 u_i 在一个未签到的 POI l_j 上的签到分布的自适应核评估 $F_{Geo}(l_j|u_i)$的计算如下：

$$F_{Geo}(l_j|u_i)=\frac{1}{A}\sum_{i=1}^{n}(r_{u_i,l_i}\cdot Q_{Hh_i}(l_j-l_i)) \tag{13}$$

$$Q_{Hh_i}(l_j-l_i)=\frac{1}{2\pi H_1H_2h_i^2}\exp\left(-\frac{(x_j-x_i)^2}{2H_1^2h_i^2}-\frac{(y_j-y_i)^2}{2H_2^2h_i^2}\right) \tag{14}$$

请注意的是：当一个签到的 POI l_i 在一个高签到密度区域时，则试点估计 $f_{Geo}(l_j|u_i)$越大，自适应当地带宽 h_i 越小，生成 l_i 附近的一个峰值自适应核评估 $F_{Geo}(l_j|u_i)$；当一个签到的 POI l_i 在一个低签到密度区域时，则试点估计 $f_{Geo}(l_j|u_i)$越小，自适应当地带宽 h_i 越大，生成 l_i附近的一个平滑自适应核评估 $F_{Geo}(l_j|u_i)$. 因此，关于用户 u_i 在一个未签到 POI l_j 上的地理相关分数，我们的自适应核评估方法可以提高评估签到分布的预测能力.

3.4 兴趣点推荐的社会相关性模型

在真实的基于位置的社交网络中，用户之间的社会关系很大程度影响用户对 POIs 的签到行为. 在 LBSNs 上用户创建社会联系意味他们之间存在着社会关系. 因此，利用用户和用户朋友之间的社会关系，根据用户朋友签到过的 POI 来推算用户与未签到的 POI 间的相关分数. 这个过程包括 3 个步骤：社会聚合、社会签到频率或评价的分布估计，社会相关分数.

3.4.1 社会聚合

形式上，给定一个用户 u_i 和一个未签到的 POI l_j，我们聚集用户 u_i 的朋友在 POI l_j 上的签到频率或者评价 x_{u_i,l_j}(即如果在两个不同的用户 $u_i,u_i'\in U$ 之间存在社会联系，则 $s_{u_i,u_i'}=1$；否则，$s_{u_i,u_i'}=0$)如下：

$$x_{u_i,l_j}=\sum_{u_i'\in U}s_{u_i,u_i'}\cdot r_{u_i',l_j} \tag{15}$$

$r_{u_i',l}$是用户 u_i'在 POI l_j 上的签到频率或者评价. $s_{u_i,u_i'}$是用户 u_i 和 u_i'之间的社会联系.

依据上述可以简单地认为社会签到频率或者评价 x_{u_i,l_j}是用户 u_i 和 POI l_j 之间的相关分数. 在传统协同过滤技术中，通过用户 u_i 的朋友数量可以简单划分 x_{u_i,l_j}. 但更复杂的是，本文基于社会签到频率或者评价的分布，对所有用户历史签到数据进行学习后，将社会签到频率或者评价转换成正则化的相关分数.

3.4.2 社会签到频率或者评价的分布估计

在真实世界的数据集中，社会签到频率或者评价的随机变量 x 遵循幂律分布，概率密度函数定义如下：

$$f_{So}(x)=(\gamma-1)(1+x)^{-\gamma} \tag{16}$$

γ 由签到矩阵 $\boldsymbol{R}_{|U|\times|L|}$ 和社会关系矩阵 $\boldsymbol{S}_{|U|\times|U|}$ 所评估：

$$\gamma=1+|U\parallel L|\left[\sum_{u_i'\in U}\sum_{l_j'\in L}\ln\left(1+\sum_{u_i''\in U}s_{u_i',u_i''}\cdot r_{u_i'',l_j'}\right)\right]^{-1} \tag{17}$$

$\sum_{u_i''\in U}s_{u_i',u_i''}\cdot r_{u_i'',l_j'}$是用户 u_i'的朋友 u_i''在 POI l_j'上的社会签到频率或者评价.

3.4.3 社会相关分数

估计概率密度函数 f_{So}相对于社会签到频率或者评价 x 是单调递减的，但是社会相关分数相对于社会签到频率或者评价应该是单调递增的，因为在 POIs 上朋友之间会分享更多的共同兴趣. 因此，基于 f_{So}的累积分布函数，我们定义 x_{u_i,l_j} 的社会相关分数如下：

$$F_{So}(x_{u_i,l_j})=\int_0^{x_{u_i,l_j}}f_{So}(a)\mathrm{d}a=1-(1+x_{u_i,l_j})^{1-\gamma} \tag{18}$$

由于 $1-\gamma<0$，则 F_{So}相对于社会签到频率或者评价 x_{u_i,l_j}是一个递增函数. 另外，基于累积分布函数 F_{So}，社会签到频率或者评价 x_{u_i,l_j} 转换成社会相关分数反映了用户在 POIs 上的所有社会签到频率或者评价 x_{u_i,l_j}的相对位置.

3.5 兴趣点推荐的分类相关性模型

在基于位置的社交网络中，每个 POI 被附加一些分类. POI 的分类明显地表示了在 POI 上发生什

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

么活动或者提供什么服务和产品.因此,利用用户签到的 POIs 和未签到的 POI 之间的分类相关性推算一个用户对一个未签到的 POI 的相关分数.另外,POI 的流行度反映了 POI 所提供的产品和服务的质量.因此,在兴趣点推荐中利用流行度是有用的.特别是,我们提出一种结合用户的分类偏好和 POI 的流行度的新方法.这个过程包括 3 个步骤:分类偏好的加权流行度、分类流行度的分布估计和分类相关分数计算.

3.5.1　利用分类偏好加权流行度

首先,我们定义 c_{u_i,c_g} 为用户 u_i 对分类 c_g 的偏好,即用户 u_i 签到属于分类 c_g 的 POIs 的频率.然后,使用偏好 c_{u_i,c_g} 加权分类 c_g 上未签到的 POI l_j 的流行度,即 p_{c_g,l_j}.相应地,我们获取了用户 u_i 在 POI l_j 上的分类流行度 y_{u_i,l_j} 如下:

$$y_{u_i,l_j}=\sum_{c_g\in C}c_{u_i,c_g}\cdot p_{c_g,l_j} \tag{19}$$

C 是在 LBSN 上预先定义的分类的集合.y_{u_i,l_j} 的值越大意味着 POI l_j 的分类越满足用户 u_i 的偏好并且意味着 POI l_j 越受公众的欢迎.

依据上述可以简单地认为分类流行度 y_{u_i,l_j} 是用户 u_i 和 POI l_j 之间的相关分数.但更复杂的是,本文基于分类流行度的分布,对所有用户历史签到数据进行学习后,将某用户对某未签到的 POI 的分类流行度映射成正则化的相关分数.

3.5.2　分类流行度的分布估计

形式上,分类流行度随机变量 y 遵循幂律分布,概率密度函数定义如下:

$$f_{Ca}(y)=(\delta-1)(1+y)^{-\delta},\ y\geq 0,\delta>1 \tag{20}$$

δ 由分类偏好矩阵 $\boldsymbol{C}_{|U|\times|C|}$ 和流行度矩阵 $\boldsymbol{P}_{|C|\times|L|}$ 所评估:

$$\delta=1+|U\parallel L|\Big[\sum_{u_i'\in U}\sum_{l_j'\in L}\ln\Big(1+\sum_{c_g\in C}c_{u_i',c_g}\cdot p_{c_g,l_j'}\Big)\Big]^{-1} \tag{21}$$

$\sum_{c_g\in C}c_{u_i',c_g}\cdot p_{c_g,l_j'}$ 是 POI l_j' 上的用户 u_i' 的分类流行度.

3.5.3　分类相关分数

估计概率密度函数 f_{Ca} 相对于分类流行度 y 是单调递减的,但是分类相关分数相对于分类流行度应该是单调递增的,因为人们偏好的流行 POIs 也满足人们的分类偏好.因此,基于 f_{Ca} 的累积分布函数,我们定义 y_{u_i,l_j} 的分类相关分数如下:

$$F_{Ca}(y_{u_i,l_j})=\int_0^{y_{u_i,l_j}}f_{Ca}(a)\mathrm{d}a=1-(1+y_{u_i,l_j})^{1-\delta} \tag{22}$$

由于 $1-\delta<0$,F_{Ca} 相对于分类流行度 y_{u_i,l_j} 是一个递增函数.另外,基于累积分布函数 F_{Ca},分类流行度 y_{u_i,l_j} 正则化成分类相关分数反映了相比于用户在 POIs 上的其他分类流行度 y_{u_i,l_j} 的相对位置.

3.6　上下文感知的概率矩阵分解模型

因为兴趣点推荐是一个与文本、地理、社会、分类、流行度相关的上下文感知的个性化推荐.因此我们提出一个上下文感知的概率矩阵分解兴趣点推荐算法.

3.6.1　分数匹配

本文认为用户历史签到矩阵 $\boldsymbol{R}_{|U|\times|L|}$ 中的 y_{u_i,l_j} 为用户 u_i 在 POI l_j 上的签到频率或者评价(如果用户 u_i 对 POI l_j 感兴趣,则 $y_{u_i,l_j}=1$;否则 $y_{u_i,l_j}=0$).此外,关于兴趣点推荐我们利用了文本信息、地理信息、社会信息、分类信息与流行度信息.

关于分数匹配,我们需要考虑 4 个部分:(1) POI 的兴趣话题匹配用户的个性化兴趣话题,从而推导兴趣相关分数;(2) 评估用户所在位置经纬度坐标的个性化签到分布,基于地理相关性,推导用户对未签到的 POI 的地理相关分数;(3) 根据用户朋友已签到的 POI,利用用户与朋友之间的社会关系,推导用户对未签到的 POI 的社会相关分数;(4) 根据用户已签到的 POIs 与未签到的 POI 的分类与流行度,基于分类相关性,推导用户对未签到的 POI 的分类相关分数.用户 u_i 在 POI l_j 上的签到频率或者评价由用户和 POI 这两种因素所决定.一方面,评价 r_{u_i,l_j} 反映用户的兴趣话题和 POI 的话题之间的匹配程度.两个话题分布匹配得越好,则评价 r_{u_i,l_j} 越高.另一方面,评价 r_{u_i,l_j} 反映用户与 POIs 之间的地理、社会与分类相关性.地理、社会与分类相关分数越高,则评价 r_{u_i,l_j} 越高.

我们融合兴趣、地理、社会与分类相关分数,由式(5)、式(13)、式(18)和式(22)给定的相关分数,关于用户 u_i 对 POI l_j 的偏好,基于乘法法则,我们把这些相关分数整合到一个统一的偏好分数 $TGSC_{ij}$ 中,定义如下:

$$TGSC_{ij}=S(u_i,l_j)F_{Geo}(l_j\mid u_i)F_{So}(x_{u_i,l_j})F_{Ca}(y_{u_i,l_j}) \tag{23}$$

$S(u_i,l_j)$ 是用户 u_i 和 POI l_j 之间的用户兴趣话题分布 θ_i 与 POI 话题分布 π_j 的兴趣相关分数.$F_{Geo}(l_j\mid u_i)$ 是自适应核评估地理相关分数,$F_{So}(x_{u_i,l_j})$ 是社会相关分数,$F_{Ca}(y_{u_i,l_j})$ 是分类相关分数.值得一提的是,关于兴趣点推荐,在之前的研究工作

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

中[7,25,31]，乘法法则被广泛应用于融合不同的因素，并显示了高鲁棒性. 因此，本文采用乘法法则融合兴趣、地理、社会与分类相关分数，从而有效地进行分数匹配.

3.6.2 TGSC-PMF 模型

利用兴趣话题、地理相关性、社会相关性与分类相关性，我们将这些因素融合到概率矩阵分解模型中. 所提的 TGSC-PMF 模型的图解如图 3 所示.

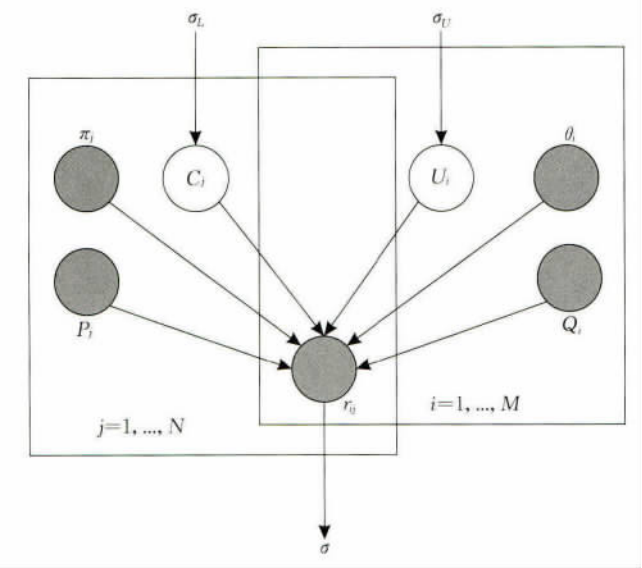


图 3 TGSC-PMF 模型

用户-POI 的签到矩阵 $\boldsymbol{R}_{|U|\times|L|}$，在签到矩阵里的每个元素 r_{u_i,l_j} 代表某用户 u_i 对某 POI l_j 的签到频率或者评价. 总共有 M 个用户和 N 个 POI. U_i 和 L_j 是用户和 POI 的潜在特征向量. 我们定义了被观察的签到频率或者评价的条件分布如下：

$$P(R|U,L,TGSC,\sigma^2)=\prod_{i=1}^{M}\prod_{j=1}^{N}[N(r_{u_i,l_j}|f(U_i,L_j,TGSC_{ij}),\sigma^2)]^{I_{ij}} \quad (24)$$

$N(\cdot|\mu,\sigma^2)$ 是具有均值 μ 和方差 σ^2 的高斯分布概率密度函数，I_{ij} 是指示函数(如果用户 u_i 访问 POI l_j 则 $I_{ij}=1$，否则，$I_{ij}=0$). 我们使用函数 $f(U_i,L_j,TGSC_{ij})$ 近似表示用户 u_i 到 POI l_j 的签到频率.

关于用户 u_i 对 POI l_j 的偏好，我们考虑了将兴趣话题、地理相关性、社会相关性与分类相关性进行有效地融合，融合函数定义如下：

$$f(U_i,L_j,TGSC_{ij})=TGSC_{ij}\cdot U_i^{\mathrm{T}}L_j \quad (25)$$

$TGSC_{ij}$ 由式(23)所计算. U_i 和 L_j 分别是用户 u_i 和 POI l_j 的 D 维潜在因素，$TGSC_{ij}$ 是用户 u_i 对 POI l_j 的话题、地理、社会与分类指数，我们利用用户潜在因素 U_i 与 POI 潜在因素 L_j 的加权内积并与话题、地理、社会和分类指数 $TGSC_{ij}$ 相乘从而能改进 PMF 模型.

并且我们在用户和 POI 的潜在特征向量中设置零均值高斯球面先验如下：

$$P(U|\sigma_U^2)=\prod_{i=1}^{M}N(U_i|0,\sigma_U^2I) \quad (26)$$

$$P(L|\sigma_L^2)=\prod_{j=1}^{N}N(L_j|0,\sigma_L^2I) \quad (27)$$

因此，通过一个简单的贝叶斯推理，式(24)的后验分布给定如下：

$$\begin{aligned}&P(U,L|R,\sigma^2,TGSC,\sigma_U^2,\sigma_L^2)\propto\\&P(R|U,L,\sigma^2,TGSC,\sigma_U^2,\sigma_L^2)P(U|\sigma_U^2)P(L|\sigma_L^2)=\\&\prod_{i=1}^{M}\prod_{j=1}^{N}[N(r_{u_i,l_j}|f(U_i,L_j,TGSC_{ij}),\sigma^2)]^{I_{ij}}\times\\&\prod_{i=1}^{M}N(U_i|0,\sigma_U^2I)\times\prod_{j=1}^{N}N(L_j|0,\sigma_L^2I)\end{aligned} \quad (28)$$

用户和 POI 的潜在特征的后验分布的对数函数如下：

$$\begin{aligned}&\ln P(U,L|R,\sigma^2,TGSC,\sigma_U^2,\sigma_L^2)=\\&-\frac{1}{2\sigma^2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}(r_{u_i,l_j}-f(U_i,L_j,TGSC_{ij}))^2-\\&\frac{1}{2\sigma_U^2}\sum_{i=1}^{M}U_i^{\mathrm{T}}U_i-\frac{1}{2\sigma_L^2}\sum_{j=1}^{N}L_j^{\mathrm{T}}L_j-\\&\frac{1}{2}\Big[\Big(\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}\Big)\ln\sigma^2+MD\ln\sigma_U^2+ND\ln\sigma_L^2\Big]+P\end{aligned} \quad (29)$$

D 是潜在因素的维数，P 是一个不依赖于参数的常量. 与用户和 POI 的潜在向量的超参数保持固定的最大化对数后验，等同于最小化平方误差和目标函数的二次正则化项：

$$\begin{aligned}E=&\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}(r_{u_i,l_j}-TGSC_{ij}\cdot U_i^{\mathrm{T}}L_j)^2+\\&\frac{\lambda_U}{2}\sum_{i=1}^{M}\|U_i\|_F^2+\frac{\lambda_L}{2}\sum_{j=1}^{N}\|L_j\|_F^2\end{aligned} \quad (30)$$

$\lambda_U=\sigma^2/\sigma_U^2$，$\lambda_L=\sigma^2/\sigma_L^2$，$\|\cdot\|_F^2$ 是 Frobenius 范数. 针对式(21)的目标函数的局部最小化，对 U 和 L 进行梯度下降算法如下：

$$\begin{aligned}\frac{\partial E}{\partial U_i}=&-\sum_{j=1}^{N}I_{ij}(r_{u_i,l_j}-TGSC_{ij}\cdot U_i^{\mathrm{T}}L_j)\cdot\\&TGSC_{ij}L_j+\lambda_U U_i\end{aligned} \quad (31)$$

$$\begin{aligned}\frac{\partial E}{\partial L_j}=&-\sum_{i=1}^{M}I_{ij}(r_{u_i,l_j}-TGSC_{ij}\cdot U_i^{\mathrm{T}}L_j)\cdot\\&TGSC_{ij}U_i+\lambda_L L_j\end{aligned} \quad (32)$$

3.6.3 预测和推荐

对用户与 POI 之间的兴趣话题、地理相关性、社会相关性、分类相关性以及参数 U 和 L 研究之

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

后,对于一个给定的 POI,TGSC-PMF 模型预测用户对其评价被 $E(r_{u_i,l_j}\mid u_i,l_j)=TGSC_{ij}\cdot U_i^{\mathrm{T}}L_j$ 所评估.由于基于位置社交网络的兴趣点推荐是高位置感知的,所以推荐列表推荐的 POIs 应该接近用户的当前区域,因此推荐用户所在地理位置附近的 POIs 给用户是可取的. TGSC-PMF 模型可以预测用户偏好分数.在真实的世界里,我们需要考虑文本、位置、社会、分类与流行度信息做出合理的个性化兴趣点推荐.给定用户当前的地理位置,可能的方法是在一定范围内对应 Top-K 的预测分数推荐给用户 K 个 POIs.

4　实　验

4.1　实验设置

4.1.1　数据集描述

Foursquare 是一个大规模的基于位置的社交网站.允许用户在不同的位置进行签到,通过分析空间、时间、社会与文本等方面的签到数据,定量评估人们的移动模式.我们使用文献[41]所提供的数据集的一部分.数据集的统计如表 2 所示.

表 2　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	签到或者评价的数量	用户-POI 矩阵密度/%
Foursquare	5468	7286	60	35216	512643	1.287

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 应该被至少访问过 10 次.此外,我们要求一个用户应该至少访问 5 个不同的 POIs.在我们的实验中,随机选择数据集的 20%作为测试数据集,其余 80%的数据集作为训练数据集.在我们实验数据集中,有 5468 个用户总共访问 7286 个 POIs.

4.1.2　评价指标

我们使用隐式评分,即一个 POI 的签到频率作为 POI 的评分.传统的电影推荐评分范围是从 1～5.不同于电影推荐评分,我们需要通过函数 $f(x)=(x-1)/(K-1)$ 将离散的评分转化到[0,1]范围内,K 是最大的评分值[42].

在我们的性能对比实验中,采用两种预测指标来评估兴趣点推荐系统的性能:均方根误差($RMSE$)和平均绝对误差(MAE).定义如下:

$$RMSE=\sqrt{\frac{1}{B}\sum_{r=1}^{B}[(R_r-\hat{R}_r)/R_r]^2} \tag{33}$$

$$MAE=\frac{\sum_{r=1}^{B}|(R_r-\hat{R}_r)/R_r|}{B} \tag{34}$$

B 是预测的总数,R_r 是用户 u_i 对一个 POI l_j 的真实评价,$\hat{R}_r$ 是预测评分.$RMSE$ 和 MAE 的值越低,则对应的预测分析精度越高.

根据预测值的排序我们给用户推荐 K 个 POIs 并且基于这些被用户签到的 POIs 进行评估.在我们的方法对比试验中,采用两种 Top-K 指标来评估兴趣点推荐的质量:$Precision@K$ 和 $Recall@K$.

定义如下:

$$Precision@K=\frac{\sum_{u}|R(u)\bigcap T(u)|}{K} \tag{35}$$

$$Recall@K=\frac{\sum_{u}|R(u)\bigcap T(u)|}{T(u)} \tag{36}$$

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.

4.1.3　参数调整

关于 TGSC-PMF 模型,融合了话题、地理、社会与分类相关性.在聚合 LDA 模型中,我们设置 $\alpha=50/T$ 和 $\beta=0.1$.在地理相关性模型中,我们发现地理相关性对参数 τ 是敏感的,当 $\tau=0.5$ 时,地理相关分数达到最佳的效果.在社会相关性模型中,参数 γ 不是自由参数,是从签到数据中学习后,由式(17)所计算得出的.在分类相关性模型中,参数 δ 不是自由参数,是从签到数据中学习后,由式(21)所计算得出的.在 TGSC-PMF 模型中,我们设置 $\lambda_U=0.01$ 和 $\lambda_L=0.01$.

4.1.4　对比方法

我们所提的方法,即与 TGSC-PMF 比较的其它先进的兴趣点推荐技术如下:

(1) LCARS.此方法基于话题模型构建位置-内容感知的推荐系统推断用户个性化兴趣和位置偏好[12-13].

(2) TL-PMF.此方法通过采用文本信息和流行度提出话题-位置感知的兴趣点推荐系统[23].

(3) USG.此方法是一个统一的位置推荐框架结合用户偏好、社会和地理信息[2].

(4) NCPD.此方法应用矩阵分解结合邻域的影响、分类、流行度和 POIs 的地理距离[24].

我们设计 4 种基线方法进一步验证分别利用兴趣话题、地理相关性、社会相关性和分类相关性所带来的好处.GSC-PMF 是第 1 个 TGSC-PMF 模型的

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

简化版即没有考虑兴趣话题因素. TSC-PMF 是第 2 个 TGSC-PMF 模型的简化版即移除了地理相关性. TGC-PMF 是第 3 个 TGSC-PMF 模型的简化版即没有考虑社会相关性. TGS-PMF 是第 4 个 TGSC-PMF 模型的简化版即移除了分类相关性.

4.2 实验结果

4.2.1 性能对比

对于 *RMSE* 和 *MAE* 这两种预测指标,我们设置不同的维数数量与话题数量,对比 TGSC-PMF、TL-PMF 和 PMF. 首先,我们分别设置话题数量 $T=30$ 和 $T=50$ 与维数数量 $D=10$ 和 $D=30$. 从用户历史签到数据中学习,获取兴趣相关分数、地理相关分数、社会相关分数和分类相关分数,并对这些相关分数进行匹配获得偏好分数 $TGSC_{ij}$. 然而我们不能直接使用 $E(r_{u_i,l_j}\mid u_i,l_j)=g(TGSC_{ij}\cdot U_i^{\mathrm{T}}L_j)$ 来进行预测,但是我们可以通过逻辑函数 $g(x)=1/(1+e^{-x})$ 预测结果,限制预测分值在[0,1]范围内. 进一步,在每个话题数量 $T=30$ 和 $T=50$ 的情况下,我们设置不同的用户因素和 POI 因素的维数 $D=10$ 和 $D=30$,对 TGSC-PMF、TL-PMF 和 PMF 进行对比实验.

如表 3 和表 4 所示,TGSC-PMF 和 TL-PMF 的性能都超过 PMF 的性能. 因为 PMF 是最简单的模型没有融合任何因素;TL-PMF 的性能居中是因为其在 PMF 模型的基础上融合兴趣话题和流行度信息;然而我们所提的 TGSC-PMF 的性能最好是因为其在 PMF 模型的基础上融合兴趣话题、地理相关性、社会相关性、分类相关性和流行度信息. 例如,当话题数量 $T=30$ 和因素维数 $D=10$ 时,TGSC-PMF 相比于 PMF 的 *RMSE* 值和 *MAE* 值分别提高 13.5%和 11.2%;TGSC-PMF 相比于 TL-PMF 的 *RMSE* 值和 *MAE* 值分别提高 7.9%和 5.7%. 通过结合兴趣话题、地理相关性、社会相关性、分类相关性与流行度信息,我们可以看到 TGSC-PMF 很大程度地改善了推荐性能.

表 3 在两个不同的因素维数与两个不同的话题数量的设置下对比 TGSC-PMF、TL-PMF、PMF 的 *RMSE* 值

RMSE	$D=10$		$D=30$	
模型	$T=30$	$T=50$	$T=30$	$T=50$
PMF	0.6679		0.6668	
TL-PMF	0.6272	0.6065	0.6388	0.6298
TGSC-PMF	0.5776	0.5658	0.5972	0.5923

表 4 在两个不同的因素维数与两个不同的话题数量的设置下对比 TGSC-PMF、TL-PMF、PMF 的 *MAE* 值

MAE	$D=10$		$D=30$	
模型	$T=30$	$T=50$	$T=30$	$T=50$
PMF	0.4949		0.4908	
TL-PMF	0.4662	0.4617	0.4708	0.4689
TGSC-PMF	0.4395	0.4296	0.4555	0.4505

4.2.2 方法对比

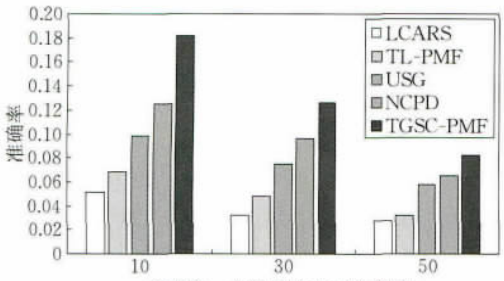
在训练集中,关于推荐给用户的 k 个 POIs 和被用户访问的 n 个 POIs,如图 4 和图 5 所示描述了我们所提的 TGSC-PMF 方法对比其他先进的兴趣点推荐技术的推荐精确度. 在所有评估的方法中,准确率和召回率的趋势是直观的. 例如,k 值增加,则准确率变低,召回率变高. 因为推荐给用户越多的 POIs 时,用户就会发现越多他们可能愿意签到的 POIs,但是一些被推荐的 POIs 被用户访问的机会就会随之减少. 随着 n 值的增加,准确率和召回率都逐渐上升. 因为使用越多的被用户访问的 POIs 的签到数据,TGSC-PMF 推荐模型就会更好地进行学习.

因为用户-POI 签到矩阵的密度很低,所以兴趣点推荐技术的绝对精度通常不高,然而随着签到数据收集得越多,则兴趣点推荐技术将表现得越好. 在前人的工作中[30,32],这种现象已经被反复观察. 相反,我们专注于对比被评估的兴趣点推荐技术的相对准确性.

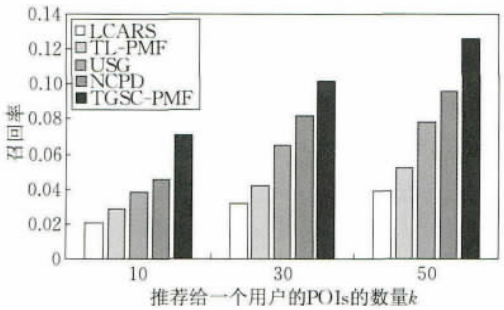
LCARS. 此方法利用话题模型(LDA)推断用户的个性化兴趣和区域的当地偏好[12-13]. 当地偏好或者个性化兴趣表现为话题的混合物,每个话题是 POIs 上的分布,并从 POIs 的签到数据和分类信息中学习到. 但是,LCARS 与 TGSC-PMF 相比忽略了 POIs 上的用户签到行为的地理特征、社会特征与流行度信息. 因此,如图 4 和图 5 所示,LCARS 的推荐精确度最低.

TL-PMF. 此方法利用文本信息和流行度提出话题和位置感知的推荐系统[23]. TL-PMF 利用一个 LDA 模型,通过挖掘 POIs 相关的文本信息,学习用户的兴趣话题并推断用户感兴趣的 POIs. 然后,TL-PMF 提出一个话题-位置感知的概率矩阵分解方法. 然而,TL-PMF 与我们所提的 TGSC-PMF 相比忽略了 POIs 上用户签到行为的地理、社会和分类相关性特征. (1) 我们所提的 TGSC-PMF 融合了 5 个上下文因素信息即 POIs 的文本信息、POIs 的地理信息、用户的社会信息、POIs 的分类信息与 POIs 的流行度信息;而 TL-PMF 只考虑了 POIs 的文本信息与流行度信息,并没有考虑地理、社会、分类信息;(2) 我们所提的 TGSC-PMF 模型,先利用 LDA 算法提出一种聚合 LDA 模型提取用户兴趣生成兴趣相关分数,然后我们提出一种自适应带宽核

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

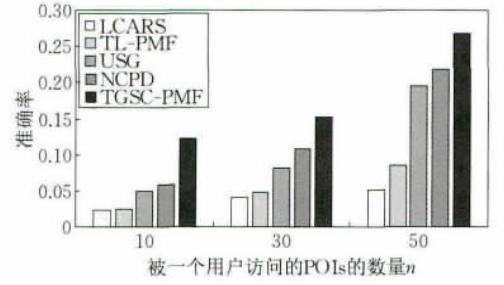


(a) 关于top-k值的推荐准确率

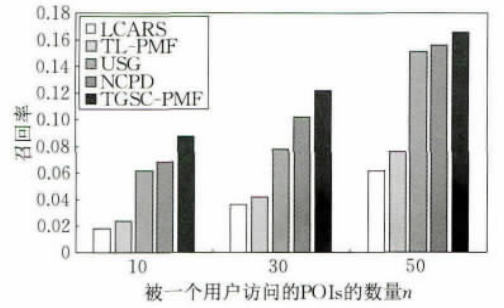


(b) 关于top-k值的推荐召回率

图 4　关于 top-k 值的推荐精确度



(a)关于given-n值的推荐准确率



(b)关于given-n值的推荐召回率

图 5　关于 given-n 值的推荐精确度

评估方法评估用户对 POIs 的地理相关分数,由于基于位置社交网络的兴趣点推荐很大程度依赖地理因素,而 TL-PMF 只考虑兴趣话题并没有考虑最重要的地理因素,TGSC-PMF 与 TL-PMF 相比较很大程度提高了推荐精确度;(3) 我们所提的 TGSC-PMF 增加了用户的社会相关性因素,我们不是简单地利用用户的社会关系,而是通过用户朋友关系的幂律分布评估用户对 POIs 的社会相关性生成社会相关分数,由于基于位置社交网络的兴趣点推荐一定程度依赖社会因素,而 TL-PMF 并没有考虑社会因素,TGSC-PMF 与 TL-PMF 相比较增加了社会因素,这一定程度提高了推荐精确度;(4) 我们所提的 TGSC-PMF 不仅同时考虑分类与流行度信息,并且对这两个因素进行融合,从而对 POIs 的流行度进行加权,基于 POIs 的分类的幂律分布评估用户对 POIs 的分类相关性生成分类相关分数,由于基于位置社交网络的兴趣点推荐一定程度依赖分类与流行度因素,而 TL-PMF 并没有考虑分类因素又只简单地考虑流行度因素,TGSC-PMF 与 TL-PMF 相比较增加了分类因素并考虑分类与流行度之间的有效加权融合,这一定程度提高了推荐精确度;(5) TL-PMF 只是简单地利用话题模型与 POIs 的流行度融合到概率矩阵分解模型中,TL-PMF 利用 POIs 的流行度信息时并没有构建流行度分布;而我们所提的 TGSC-PMF 利用话题模型、地理相关性、社会相关性、分类相关性推导出兴趣相关分数、地理相关分数、社会相关分数,然后对这些分数进行有效匹配生成偏好分数融合到概率矩阵分解模型中. 我们所提的上下文感知的概率矩阵分解模型 TGSC-PMF 相比 TL-PMF 推荐精确度有了显著地提高. 因此,如图 4 和图 5 所示,TL-PMF 的推荐精确度仅略高于 LCARS,远远不及 TGSC-PMF.

USG. 此方法基于用户的协同过滤与基于社会的协同过滤线性整合用户偏好、社会影响和所有用户距离分布的地理影响[2],但 USG 与 TGSC-PMF 相比没有考虑 POIs 的分类和流行度信息. 另外,USG 对于用户偏好、社会影响和地理影响,采用通用线性加权是不明智的,因为有些用户可能受社会的朋友影响更多,有些用户可能受地理影响更多. 然而 TGSC-PMF 采用乘法法则整合兴趣、地理、社会与分类相关分数能够更有效地进行融合. 因此,如图 4 和图 5 所示,USG 的推荐精度排第 3.

NCPD. 此方法利用矩阵分解方法推导每个用户、POI 和分类的潜在因素向量,推断每个用户的地理偏好和每个 POI 的流行度偏好[24]. 然后,基于用户的潜在因素向量、POI 的分类、POI 的邻域、用户的地理偏好和 POI 的流行度偏好,计算某用户对某 POI 的偏好分数. NCPD 相比 USG 增加了分类与流行度信息所以精确度有所提高. 但是 NCPD 与 TGSC-PMF 相比只是简单地把地理和流行度影响作为用户的偏好,没有把它们建模成地理或者流行度分布,并且忽略了社会因素. 因此,如图 4 和图 5

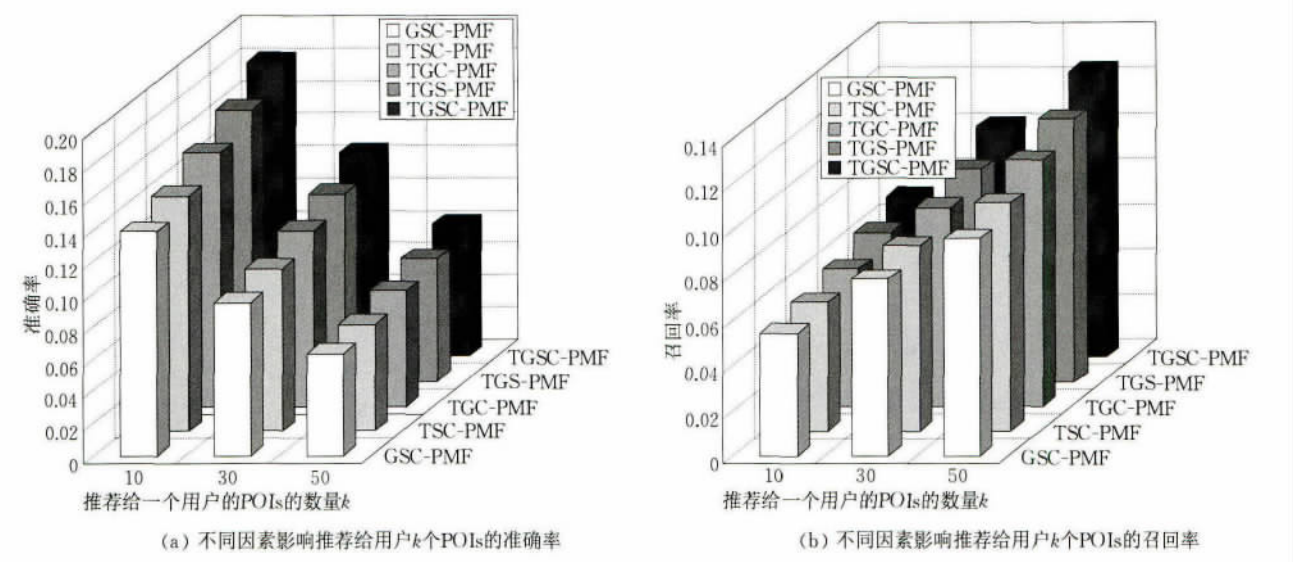
?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

所示，NCPD 的推荐精确度排第 2.

TGSC-PMF. 如图 4 和图 5 所示，关于准确率和召回率，我们所提的 TGSC-PMF 模型表现出最好的推荐精确度. 特别是，我们所提的方法相比精确度第 2 高的 NCPD 推荐技术有了明显的改善. 主要有以下几个原因：(1) TGSC-PMF 利用一个聚合 LDA 模型学习用户的兴趣话题并挖掘 POIs 相关的文本信息推断用户感兴趣的 POIs；(2) 从用户历史签到数据中学习，构建 POIs 之间的地理相关性，在地理坐标上评估个性化签到分布，采用自适应带宽核评估计算某用户对某 POI 的地理相关分数；(3) 从用户历史签到数据中学习，基于社会关系的幂律分布，TGSC-PMF 利用用户朋友的社会签到频率或者评价，有效地将社会签到频率或者评价转换为合理的社会相关分数. 这种方法优于传统的基于社会的协同过滤技术；(4) 从用户历史签到数据中学习，基于分类流行度的幂律分布，利用分类与流行度信息，无缝整合用户的分类偏好与 POIs 的流行度，有效地将其转换为合理的分类相关分数；(5) TGSC-PMF 整合兴趣相关分数、地理相关分数、社会相关分数与分类相关分数到概率矩阵分解模型中，有效地融合了文本、地理、社会、分类与流行度信息. 总结以上这些原因，所以我们所提的 TGSC-PMF 的准确率和召回率最高.

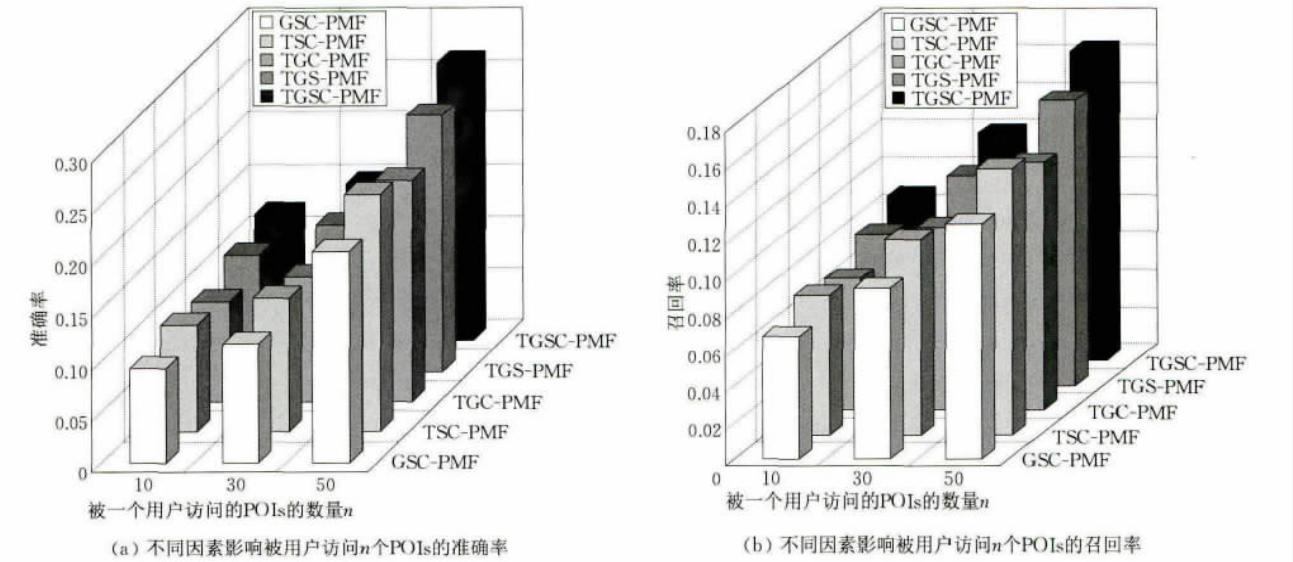
4.2.3 不同因素影响

为了分别研究 TGSC-PMF 模型融合话题分布、地理相关性、社会相关性、分类相关性所带来的好处，我们对比了我们所提的 TGSC-PMF 模型的 4 个基线方法，GSC-PMF、TSC-PMF、TGC-PMF 和 TGS-PMF. 对比的结果如图 6 和图 7 所示. 从结果



(a) 不同因素影响推荐给用户k个POIs的准确率　(b) 不同因素影响推荐给用户k个POIs的召回率

图 6 不同因素影响推荐给用户 k 个 POIs 的精确度



(a) 不同因素影响被用户访问n个POIs的准确率　(b) 不同因素影响被用户访问n个POIs的召回率

图 7 不同因素影响被用户访问 n 个 POIs 的精确度

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

中,我们首先观察到在推荐给用户 k 个 POIs 和被用户访问 n 个 POIs 的两种评估方法中 TGSC-PMF 始终优于 4 个基线方法,并表明在两种评估方法中 TGSC-PMF 受益于 4 个因素来提高推荐精确度. 此外,另一个观察是同一个因素在两种不同的评估方法中的影响程度是不同的.

特别是,在推荐给用户 k 个 POIs 的评估方法中,根据 4 个因素的重要性,它们的影响程度排列如下:兴趣话题>地理相关性>社会相关性>分类相关性;然而在被用户访问 n 个 POIs 的评估方法中,它们的影响程度排列如下:兴趣话题>社会相关性>地理相关性>分类相关性. 观察结果如下:(1)兴趣话题在两种评估方法中都起着最重要的作用;(2)关于兴趣点推荐,4 个因素在 TGSC-PMF 算法中都起着重要的作用而且又是彼此竞争的. 例如,在推荐给用户 k 个 POIs 的评估方法中,地理相关性相比社会相关性更重要一些,然而在被用户访问 n 个 POIs 的评估方法中,社会相关性比地理相关性更重要一些;(3)4 个因素的集成有助于提高推荐质量,因为 TGSC-PMF 的推荐精确度明显优于每个因素,原因是在实践中人们不同程度地受兴趣、地理、社会和分类相关性的影响,只考虑一种类型的相关性无法建模所有用户的签到行为.

4.2.4　参数影响

调整模型参数,即话题的数量 T 和维度的数量 D 对于 TGSC-PMF 模型的性能是重要的. 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 α 和 β,我们设置固定值 $\alpha=50/T$ 和 $\beta=0.1$. 我们尝试不同的设置,发现 GTSC-PMF 模型的性能对这些超参数是不敏感的,但是 TGSC-PMF 模型的性能对主题和维数的数量是敏感的. 因此我们通过设置主题和维数的数量来测试 TGSC-PMF 模型的性能,结果如图 8 和图 9 所示.

从图 8 的结果中,首先我们观察到 TGSC-PMF 的推荐均方根误差 RMSE 值随着维度的数量 D 增加而增加,由于均方根误差值越小精确度越高,因此随着维度的数量增加,精确度有所下降;然后当维度的数量大于 50 时,推荐均方根误差值的变化就不明显了. 类似的观察,随着话题的数量 T 增加,TGSC-PMF 的推荐均方根误差值减小,随着话题的数量增加,精确度有所上升,然后当话题的数量大于 50 时推荐均方根误差值的变化也不明显. 原因是 D 和 T 代表模型的复杂度. 因此,一方面,当 D 和 T 的值太小时,模型描述数据的能力受限;另一方面,当 D 和 T 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 D 和 T 的值对提高模型性能的帮助就不明显了.

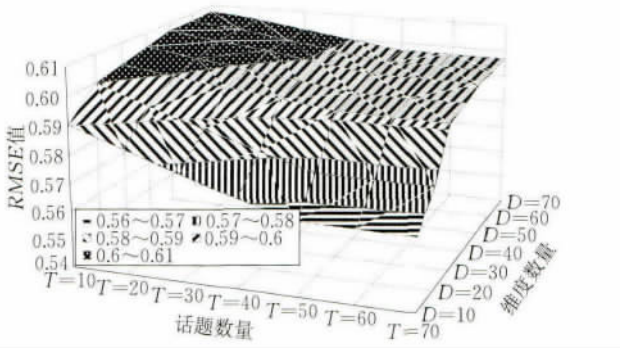


图 8　参数 T 和 D 的不同值时 $RMSE$ 值

从图 9 的结果中,首先我们观察到 TGSC-PMF 的推荐平均绝对误差 MAE 值相比图 8 中 TGSC-PMF 的推荐均方根误差 $RMSE$ 值都有所减小. 其次我们观察到 TGSC-PMF 的推荐平均绝对误差 MAE 值随着维度数量 D 的增加而增加,由于平均绝对误差值越小精确度越高,因此随着维度的数量增加,精确度有所下降;然后当维度的数量大于 50 时,推荐平均绝对误差值的变化就不明显了. 类似的观察,随着话题数量 T 的增加,TGSC-PMF 的推荐平均绝对误差值减小,随着话题数量的增加,精确度有所上升,然后当话题的数量大于 50 时推荐平均绝对误差值的变化也不明显. 原因是 D 和 T 代表模型的复杂度. 因此,一方面,当 D 和 T 的值太小时,模型描述数据的能力受限;另一方面,当 D 和 T 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 D 和 T 的值对提高模型性能的帮助就不明显了.

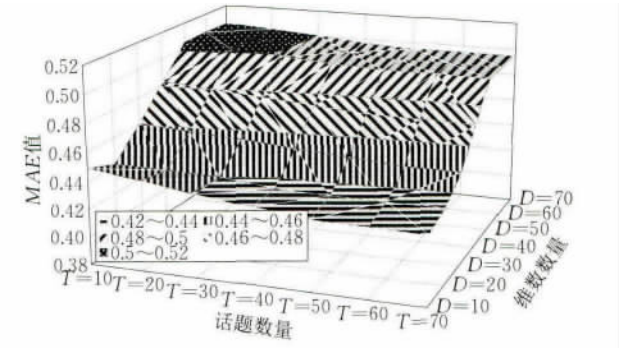
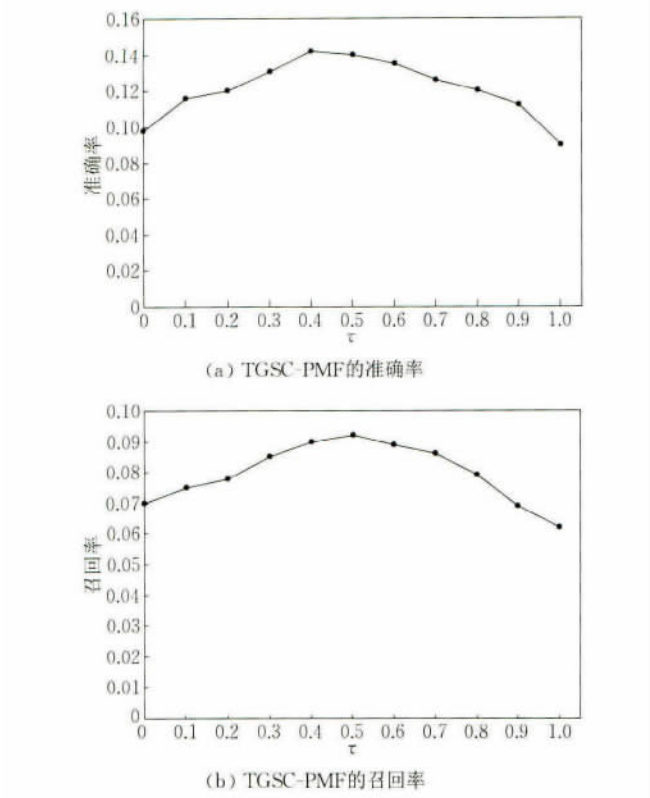


图 9　参数 T 和 D 的不同值时 MAE 值

图 10 描述了基于位置社交网络数据集中式(11)的敏感参数 τ 对 TGSC-PMF 的准确率与召回率的影响. 请注意参数 γ 和 δ 可以从签到数据中学习,不是自由参数. 观察结果如下:(1)τ 的最优值在 0.4~0.6 之间,此时可生成最高的推荐精确度;(2)τ 值在

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

0～0.4 之间变化时，当地带宽对式(6)的试点估计不那么敏感，即当地带宽与签到数据不太相关. 特别是，当 $\tau=0$ 时，自适应带宽退化成固定带宽，这样就与签到数据无关了. 结果，精确率和召回率降低；(3)相反，当 τ 值升高至 0.6～1.0 之间，当地带宽对式(6)的试点估计就更加敏感，关于签到数据当地带宽容易过度拟合. 因此，推荐质量也退化了.



(a) TGSC-PMF的准确率

(b) TGSC-PMF的召回率

图 10 敏感参数 τ 对 TGSC-PMF 推荐精确度的影响

5 总结与未来工作

针对兴趣点推荐本文提出一个上下文感知的概率矩阵分解模型，称为 TGSC-PMF，并利用文本信息、地理信息、社会信息、分类信息与流行度信息，有效地融合了兴趣话题、地理相关性、社会相关性与分类相关性. 首先，我们利用一个聚合 LDA 模型学习用户的兴趣话题，挖掘 POIs 相关的文本信息推断用户感兴趣的 POIs. 其次，构建地理相关性，提出一种自适应带宽核评估方法，评估用户对 POIs 的地理相关分数. 然后，构建社会相关性，通过用户的朋友到 POIs 的幂律分布，评估社会签到频率或者评价将社会相关性转换成社会相关分数. 再次，构建分类相关性，结合用户的分类偏好与 POIs 的流行度将其转换为用户对 POIs 的分类相关分数. 我们所提的 TGSC-PMF 兴趣点推荐能有效地匹配兴趣相关分数、地理相关分数、社会相关分数与分类相关分数生成偏好分数，将偏好分数整合到概率矩阵分解模型中，从而有效地融合兴趣、地理、社会与分类相关性. 最终，在真实的 LBSNs 的数据集中，实验结果有效验证该方法的推荐效果，并显示 TGSC-PMF 的精确度相比其他当前先进的兴趣点推荐技术有了明显提高.

在未来工作中我们将计划结合用户对 POIs 评价的文本信息中提取出的情感因素，或者结合时间因素，进一步提高兴趣点推荐的性能.

致 谢 本文工作是在北京邮电大学信息网络工程研究中心完成的. 该中心为教育部重点实验室. 本文审稿专家和编辑提出了宝贵意见和建议，在此致谢！

参 考 文 献

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

[2] 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

[3] Gao H J, Liu H. Mobile Social Networking: Data Analysis on Location-Based Social Networks. New York, USA: Springer, 2014

[4] 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

[5] 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

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

[7] 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

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[8] 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

[9] 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)

[10] Liu Y, Wei W, Sun A X, Miao C Y. Exploiting geographical neighborhood characteristics for location recommendation//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai, China, 2014: 739-748

[11] Hu B, Ester M. Spatial topic modeling in online social media for location recommendation//Proceedings of the 7th ACM Conference on Recommender Systems. Hong Kong, China, 2013: 25-32

[12] Yin H Z, Cui B, Sun Y Z, et al. LCARS: A spatial item recommender system. ACM Transactions on Information Systems, 2014, 32(3): 11. 1-11. 37

[13] 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

[14] Farrahi K, Gatica-Perez D. Discovering routines from large-scale human locations using probabilistic topic models. ACM Transactions on Intelligent Systems and Technology, 2011, 2(1): 3. 1-3. 27

[15] 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

[16] Yin Z J, Cao L L, Han J W, Huang T. Geographical topic discovery and comparison//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 247-256

[17] 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

[18] Agarwal D, Chen B C. Flda: Matrix factorization through latent dirichlet allocation//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2010: 91-100

[19] Pennacchiotti M, Gurumurthy S. Investigating topic models for social media user recommendation//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 101-102

[20] Bao J, Zheng Y, Mokbel M F. Location-based and preference-aware recommendation using sparse geo-social networking data//Proceedings of the 20th International Conference on Advances in Geographic Information Systems. Redondo Beach, California, USA, 2012: 199-208

[21] 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. San Francisco, USA, 2013: 721-726

[22] Wang H, Terrovitis M, Mamoulis N. Location recommendation in location-based social networks using user check-in data//Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Orlando, USA, 2013: 374-383

[23] 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

[24] 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

[25] 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

[26] Kurashima T, Iwata T, Hoshide T, et al. Geo topic model: Joint modeling of user's activity area and interests for location recommendation//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome, Italy, 2013: 375-384

[27] Liu X, Liu Y, Aberer K, Miao C Y. Personalized point-of-interest recommendation by mining users' preference transition //Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. Burlingame, USA, 2013: 733-738

[28] Yao Z, Liu B, Fu Y, et al. User preferencelearning with multiple information fusion for restaurantrecommendation//Proceedings of the 2014 SIAM International Conference on Data Mining. Philadelphia, USA, 2014: 470-478

[29] 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

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[30] Yuan Q, Cong G, Sun A X. Graph-based point-of-interest recommendation with geographical and temporal influences// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai, China, 2014: 659-668

[31] Zhang J D, Chow C Y. iGSLR: Personalized geo-social location recommendation — A kernel density estimation approach// Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Orlando, USA, 2013: 334-343

[32] Zhang J D, Chow C Y. CoRe: Exploiting the personalized influence of two-dimensional geographic coordinates for location recommendations. Journal of Information Sciences, 2015, 293(1): 163-181

[33] Zhang J D, Chow C Y, Li Y H. Lore: Exploiting sequential influence for location recommendations//Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Dallas, USA, 2014: 103-112

[34] 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

[35] Yang D Q, Zhang D Q, Yu Z Y, Wang Z. A sentiment enhanced personalized location recommendation system// Proceedings of the 24th ACM Conference on Hypertext and Social Media. Paris, France, 2013: 119-128

[36] Liu X, Wu W. Learning context-aware latent representations for context-aware collaborative filtering//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 887-890

[37] Rahimi S M, Wang X. Location recommendation based on periodicity of human activities and location categories// Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Gold Coast, Australia, 2013: 377-389

[38] Zhao Y L, Nie L Q, Wang X Y, Chua T S. Personalized recommendations of locally interesting venues to tourists via cross-region community matching. ACM Transactions on Intelligent Systems and Technology, 2014, 5(3): 50.1-50.26

[39] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022

[40] Griffiths T L, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Supplement 1): 5228-5235

[41] 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

[42] Salakhutdinov R, Mnih A. Probabilistic matrix factorization// Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2007: 1257-1264

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.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

POI recommendation is a personalized, location-aware, and context depended recommendation. Therefore, we propose a context-aware probabilistic matrix factorization method called TGSC-PMF for POI recommendation, exploiting geographical information, text information, social information, categorical information and popularity information, incorporating these factors effectively.

First, we exploit an aggregated Latent Dirichlet Allocation (LDA) model to learn the interest topics of users and infer the interest POIs by mining textual information associated with POIs and generate interest relevance score. Second, we propose a kernel estimation method with an adaptive bandwidth to model the geographical correlations and generate geographical relevance score. Third, we build social relevance through the power-law distribution of user social relations to generate social relevance score. Then, we model the categorical correlations which combine the category bias of users and the popularity of POIs into categorical relevance score. Further, our exploit probabilistic matrix factorization model (PMF) to integrate the interest, geographical, social and categorical relevance scores for POI recommendation. Finally, we implement experiments on a real LBSN check-in dataset. Experimental results show that TGSC-PMF achieves significantly superior recommendation quality compare to other state-of-the-art POI recommendation techniques.

The work described in this paper is supported by the National Key Project of Scientific and Technical Supporting Programs of China (Grant No. 2014BAK15B01); the Cosponsored Project of Beijing Committee of Education; Engineering Research Center of Information Networks, Ministry of Education.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net