计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):200−206 [doi: 10.15888/j.cnki.csa.006722]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

LBSN 中融合类别信息的混合推荐模型①

张岐山[1], 李　可[1], 林小榕[2]

[1](福州大学 经济与管理学院, 福州 350108)
[2](北京交通大学 下一代互联网互联设备国家工程实验室, 北京 100044)
通讯作者: 李　可, E-mail: 411700408@qq.com

摘　要: 针对基于位置的社交网络 (Location-Based Social Network, LBSN) 中用户签到数据的高稀疏性问题及用户隐私问题, 提出了一种混合推荐模型 (SoGeoCat). 首先, 通过用户潜在兴趣点数据模型, 学习用户的潜在兴趣点; 其次, 将用户的潜在兴趣点纳入融合类别信息的矩阵分解模型中并优化; 最后, 根据用户特征矩阵、兴趣点特征矩阵, 提出推荐策略. 基于 Foursquare 真实数据集, 实验结果表明: (1) 相比于其他几个推荐模型, 该算法将用户的潜在兴趣点填充至用户-兴趣点矩阵中, 可以有效地缓解数据稀疏性的影响; (2) 该算法可保护用户家庭信息; (3) 在推荐模型中纳入类别信息的影响能提高推荐效果.

关键词: 位置社交网络; 地理位置信息; 类别信息; 矩阵分解; 兴趣点推荐

引用格式: 张岐山,李可,林小榕.LBSN 中融合类别信息的混合推荐模型.计算机系统应用,2019,28(1):200–206. http://www.c-s-a.org.cn/1003-3254/6722.html

Hybrid Recommendation Model Integrating Category Information in LBSN

ZHANG Qi-Shan[1], LI Ke[1], LIN Xiao-Rong[2]

[1](School of Economics and Management, Fuzhou University, Fuzhou 350108, China)
[2](National Engineering Laboratory for NGI Interconnection Devices, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Aiming at the high sparsity problem of user's check-in data and user privacy in LBSN, a hybrid recommendation model (SoGeoCat) is proposed. Firstly, the user's potential point-of-interest is learnt from the user potential point of interest data model. Secondly, the user's potential point-of-interest is incorporated into a category based matrix factorization model and then optimized. Finally, the proposed recommended strategy is according to the user and feature matrix and the point-of-interest matrix. Based on the Foursquare real dataset, the experimental results show that: (1) compared with several other recommended models, the algorithm fills the user's potential point-of-interest into the matrix, which can effectively alleviate the impact of data sparsity; (2) the algorithm can protect the user's family information; (3) the influence of the category information in the recommendation model can improve the recommendation effect.

Key words: LBSN; geographical information; category information; matrix factorization; point-of-interest recommendation

近年来, 基于位置的社交网络服务 (Location-Based Social Network, LBSN) 得到迅速发展, 如 Loopt、Yelp、Foursquare、Whrrl 等[1]. 在这些 LBSNs 中, 用户访问线下的兴趣点 (Point-Of-Interest, POI), 如: 餐馆、电影院、博物馆等, 在线上进行“签到”活动, 并分享他们访问兴趣点时丰富的建议与经历[2]. 兴趣点推荐可以

① 基金项目: 国家自然科学基金 (61300104); 福建省自然科学基金 (2018J01791)
Foundation item: National Natural Science Foundation of China (61300104); Natural Science Foundation of Fujian Province (2018J01791)
收稿时间: 2018-06-30; 修改时间: 2018-07-27; 采用时间: 2018-08-08; csa 在线出版时间: 2018-12-26

减少用户的搜寻时间, 为商家提供精准营销策略. 所以如何利用这些信息, 为目标用户推荐正确的兴趣点集是一个很有前途、很有趣的研究问题. 目前, 有很多学者运用协同过滤、矩阵分解、LDA 模型等技术于兴趣点推荐之中, 但是普遍存在以下几个问题:

(1) 数据稀疏问题. 在 LBSNs 中兴趣点推荐研究遭遇到了严重的数据稀疏问题. 通常情况下, 一个用户访问的兴趣点的数量仅仅是兴趣点总数当中很小的一部分. 例如, Netflix 电影推荐的数据密度是 1.2%, 而兴趣点推荐研究实验中使用的数据密度通常在 0.1% 左右[3]. Ye 等人[4]提出了融合地理位置、用户偏好和社会影响的统一协同过滤模型, Lian 等人[5]提出了加权矩阵分解模型, 均容易受到数据稀疏性的影响. 协同过滤算法利用用户之间的相似性进行有效地推荐, 很容易受到数据稀疏性的影响. 而且该算法只考虑到了签到数据的显式反馈, 不能有效地融合异构数据源[6]. 矩阵分解算法可缓解数据稀疏性的影响, 但是其忽略了用户之间的相似性.

(2) 隐私问题. 很多保护隐私意识比较强的用户, 他们在 LBSNs 中不会透露家庭住址、公司地址等有效信息. Li 等人[7]考虑了用户“家”的地理位置, 认为单考虑地理位置影响则家与兴趣点之间的距离同其访问该兴趣点的概率呈幂律分布. 但在这些信息不完全甚至是没有这些信息的情况下, 如何进行有效的兴趣点推荐是我们要研究的问题之一.

(3) 类别信息. 每个兴趣点都会有其类别信息, 如: 饭店、电影院、博物馆等. 从历史签到记录来看, 每个用户都会偏向于访问类别相同或者相似的兴趣点[7]. 因此, 如何利用兴趣点的类别信息提高兴趣点推荐的准确率是我们研究内容的重点.

本文针对上述的问题, 提出了 SoGeoCat(Social-Geography-Category, SoGeoCat) 模型, 主要贡献如下:

(1) SoGeoCat 模型结合了协同过滤算法和矩阵分解算法的优点, 首先根据用户行为相似性发现目标用户的潜在兴趣点, 然后将潜在兴趣点纳入矩阵分解模型当中, 克服了单纯协同过滤和矩阵分解算法的不足, 即考虑了用户相似度又很大程度上缓解了数据稀疏性的问题.

(2) 本文利用贝叶斯规则, 根据目标用户的历史签到轨迹来判断拟推荐兴趣点在地理位置因素上对目标用户的影响.

(3) 本文将兴趣点的类别标签纳入矩阵分解模型中, 提高 SoGeoCat 模型的推荐效率.

1 相关工作

协同过滤和矩阵分解是兴趣点推荐研究中主流的两种算法.

(1) 基于协同过滤算法的推荐. 协同过滤的主要思想是: 分析用户之间的关系和项目之间的相互依赖关系, 以识别新的用户—项目关联[8–10].

Ye 等人[4]提出了融合地理位置、用户偏好和社会影响的统一协同过滤方法. 采用幂律概率模型捕捉兴趣点之间的地理位置影响, 通过朴素贝叶斯方法实现基于地理影响的兴趣点协同推荐. Yuan 等人[11]在统一的协同过滤框架上纳入了时间信息的影响, 利用时间感知进行兴趣点推荐. 但该算法很容易受到数据稀疏性的影响, 也不能很好地实现对隐式反馈数据集的挖掘.

(2) 基于矩阵分解算法的推荐. 矩阵分解法的核心是训练出用户和兴趣点的特征向量, 并以此来预测用户对于某一特定兴趣点的偏好. 其不仅可以缓解数据稀疏性的影响还可以融合异构数据源, 考虑隐式反馈数据集[5, 12–14].

Lian 等人[5]将地理位置影响纳入加权矩阵分解框架当中, 根据签到记录的空间聚集现象, 提出了 GeoMF 模型, 模拟用户活动区域与地理位置之间的影响关系. 高榕等人[14]在经典的矩阵分解模型的基础上, 融合异构数据, 提出了 GeoSoRev 模型, 采用基于矩阵分解的主题模型来发现评论中的隐藏“主题”. 矩阵分解算法虽然缓解了数据的稀疏性, 也融合不同的异构数据, 但它没有考虑到用户之间的相似性.

(3) 混合算法推荐. 为了克服两种算法的不足之处, 有一些学者提出了混合算法. Li 等人[7]提出了“两步走”的框架. 第一步设计基于线性聚集和基于随机游走两种方法, 为每个用户学习一组他们可能感兴趣的潜在兴趣点. 在第二步骤中, 用基于平方误差的损失函数和基于排名误差的损失函数来模拟这三种签到.

文献[5]中认为用户的签到概率和从家到相应位置的距离遵循幂律分布. 一方面, 家的位置信息较难获得, 很多用户隐私保护意识越来越强, 不愿意透露家庭位置信息; 另一方面, 用户签到过的兴趣点可能会聚集在某两个距离比较远的区域, 如家和公司附近. 因此, 本文针对上述问题, 在文献[5]的基础上继续研究, 提出

了 SoGeoCat (Social-Geography-Category) 模型, 用朴素贝叶斯方法计算地理位置因素对于用户决策的影响, 保护用户家庭位置信息, 并将签到信息、朋友信息、地理位置信息和类别信息纳入混合模型中, 即考虑了用户相似性又缓解了数据稀疏问题, 提高了模型的推荐效果.

2 用户潜在兴趣点数据模型

2.1 问题描述

本文主要研究的问题与传统的基于协同过滤的推荐模型或基于矩阵分解的推荐模型不同, 而是采用了“两步走”的框架模型 SoGeoCat: 首先, 建立用户潜在兴趣点数据模型, 利用用户的签到信息、朋友信息、地理位置信息对用户的签到信息进行有效地扩充; 然后, 建立一个融合类别标签的矩阵分解模型, 训练出用户特征矩阵和兴趣点特征矩阵; 最后考虑用户特征、兴趣点特征的影响, 估算出目标用户对于某一特定的兴趣点的访问概率, 进而推荐有效的兴趣点集.

假设 u_i 为目标用户, l_j 为拟推荐兴趣点. U 为用户集, 即 $U=\{u_1,u_2,\cdots,u_n\}$, L 为兴趣点集, 即 $L=\{l_1,l_2,\cdots,l_m\}$. 运用 SoGeoCat 模型计算出 u_i 访问每一个未访问过的 POI 的概率, 选取 Top S 作为 u_i 的拟推荐兴趣点集.

2.2 基于签到行为相似度建模

用户在 LBSNs 中有大量的签到信息, 签到信息包括用户 ID, 兴趣点 ID 和访问次数. 访问次数越多, 则说明用户对该兴趣点的偏好越强. 用户 i 与用户 u 已签到过的共同的兴趣点越多, 则他们的签到行为越相似, 即签到行为相似度 $Sim(u_i, u_u)$ 越高, 本文采用余弦相似度来度量两用户之间的签到行为相似度, 建模如下:

$$Sim(u_i,u_u)=\left(\sum_{z\in M_i^o} r_{i,z}^2 \sum_{z\in M_u^o} r_{u,z}^2\right)^{-\frac{1}{2}} \sum_{z\in M_i^o\cap M_u^o} r_{i,z}r_{u,z} \quad (1)$$

其中, $r_{i,z}$表示 u_i 在兴趣点 l_z 的签到次数, $r_{u,z}$表示 u_u 在兴趣点 l_z 的签到次数, M_i^o表示 u_i 访问过的兴趣点的集合, M_u^o表示 u_u 访问过的兴趣点的集合.

注意: 这里的 u_u 曾经在兴趣点 l_j 处有签到行为.

2.3 基于朋友相似度建模

用户在 LBSNs 上有一些相互关注的好友, 这些好友关系也反映了该用户在现实生活中的朋友圈. 现实中, 你朋友的推荐会激发你对某些兴趣点的兴趣, 在 LBSNs 中亦是如此. 所以, u_f(u_i 的朋友) 的签到记录很有可能是 u_i 想要访问的潜在兴趣点. 但是 u_i 有很多好友, 不一定每一个好友签到过的兴趣点, u_i 都会感兴趣. 对此, 提出了朋友相似度 $Sim(u_i, u_f)$, 朋友相似度越高, 其历史签到记录越有参考价值, 建模如下:

$$Sim(u_i,u_f)=\left(\sum_{z\in M_i^o} r_{i,z}^2 \sum_{z\in M_f^o} r_{f,z}^2\right)^{-\frac{1}{2}} \sum_{z\in M_i^o\cap M_f^o} r_{i,z}r_{f,z} \quad (2)$$

其中, $r_{i,z}$表示 u_i 在兴趣点 l_z 的签到次数, $r_{f,z}$表示 u_f 在兴趣点 l_z 的签到次数, M_i^o表示 u_i 访问过的兴趣点的集合, M_f^o表示 u_f 访问过的兴趣点的集合.

注意: 这里的 u_f 曾经在兴趣点 l_j 处有签到行为.

2.4 基于地理位置相似度建模

人们往往喜欢访问地理位置离自己近的兴趣点, 单考虑地理位置影响因素, 用户访问兴趣点的概率同其距离遵循幂率分布, 模型[5]如下:

$$Pr(d)=a\cdot d^b \quad (3)$$

其中, d 表示用户同兴趣点之间的距离, a 和 b 均为幂律分布的参数.

但是, 在本文中只有用户历史签到记录信息, 没有用户的实时地理位置信息, 所以, 不能算出用户与某一兴趣点之间的准确距离. 为了解决此问题, 且又能保护用户的家庭住址或公司地址等常驻地址信息, 本文采用了基于朴素贝叶斯规则的模型,计算地理位置相似度 $P(l_j|L_{u_i})$, 已知 u_i 的全部历史签到记录 L_u, 我们计算 $P(l_j|L_{u_i})$作为每个候选兴趣点 l_j 的排名分数, 然后向用户推荐排名前 S 个的兴趣点, 建模如下:

$$\begin{aligned} P(l_j|L_{u_i}) &= \frac{P(l_j\cup L_{u_i})}{P(L_{u_i})} = \frac{P(L_{u_i})\times\prod_{l_k\in L_{u_i}} P(l_k|l_j)}{P(L_{u_i})} \\ &= \prod_{l_k\in L_{u_i}} P(l_k|l_j) \end{aligned} \quad (4)$$

注意: 这里假定L_{u_i}中的兴趣点的签到概率彼此独立.

2.5 相似度的线性聚集

综合考虑上述三个因素的影响, 对签到行为相似度、朋友相似度和地理位置相似度进行线性聚合. 但是, 它们是通过不同的方法来衡量的, 具有不同的价值范围. 因此, 我们采用最小-最大归一化进行处理, 然后

再进行聚集.

同时, 签到次数也能侧面地反映用户的偏好. 根据公式 (8) 计算出 u_i 对于拟推荐兴趣点 l_j 的分数, 选取分数高的前 S 个兴趣点作为 u_i 的潜在兴趣点.

$$Sim'(u_i,u_u)=\frac{Sim(u_i,u_u)-\min(Sim(u_i,u_u))}{\max(Sim(u_i,u_u))-\min(Sim(u_i,u_u))} \quad (5)$$

$$P'\left(l_j|L_{u_i}\right)=\frac{P\left(l_j|L_{u_i}\right)-\min\left(P\left(l_j|L_{u_i}\right)\right)}{\max\left(P\left(l_j|L_{u_i}\right)\right)-\min\left(P\left(l_j|L_{u_i}\right)\right)} \quad (6)$$

$$Sim\left(\mathrm{u_i},l_j\right)=\zeta\max(Sim'(u_i,u_u))+(1-\zeta)P'\left(\mathrm{l_j}|L_{u_i}\right) \quad (7)$$

$$\mathrm{score}=Sim\left(u_i,l_j\right)\times rating_{u,j} \quad (8)$$

其中, U 表示与 u_i 访问过相同兴趣点的用户及 u_i 的朋友的集合且$u_u\in U$, $Sim(u_i, l_j)$ 表示用户 u_i 对拟推荐兴趣点 l_j 的聚集相似度. ζ是调整参数.

3 SoGeoCat 模型

3.1 SoGeoCat 模型

用户 u_i 对于兴趣点 l_j 的偏好程度受用户潜在特征和兴趣点潜在特征影响. 令用户特征矩阵为 U, 兴趣点特征矩阵为 V, 偏好矩阵为 P, 则:

$$P\approx\hat{P}=U^{\mathrm{T}}V \quad (9)$$

用$\hat{P}$值来估计P值. 为了缓解数据稀疏问题, 我们从用户潜在兴趣点数据模型中挖掘到了用户的潜在兴趣点, 并用于扩充偏好矩阵. 但是用户对于潜在兴趣点和已签到过的兴趣点的偏好是有不同的, 对于这一现象, 本文将二元偏好变量 P_{ij} 扩充为三元值, 公式为:

$$P_{ij}=\begin{cases}1 & \text{if } j\in M_i^o\\ \alpha & \text{if } j\in M_i^p\\ 0 & \text{otherwise}\end{cases}\quad \text{where } \alpha\in[0,1] \quad (10)$$

其中, M_i^o表示目标用户 u_i 访问过的兴趣点集, M_i^p表示目标用户 u_i 的潜在兴趣点集.

在 LBSNs 的兴趣点推荐中, 其类别信息发挥着重要的作用. 从历史签到记录来看, 每个用户都会偏向于访问类别相同或相似的兴趣点, 如: u_i 之前经常访问饭店, 但几乎没去过电影院, 此时如果给他推荐电影院, 则其访问的可能性就会大大降低. 设q_{ic_j}表示 u_i 对于 l_j 对应的类别 c 的偏好程度, Q 表示类别特征矩阵. 将类别信息纳入矩阵分解模型中, 模型为:

$$\hat{P}_{ij}=\left(q_{ic_j}+\varepsilon\right)u_i^{\mathrm{T}}v_j \quad (11)$$

其中, ε为调整参数.

损失函数为:

$$\min_{U,V,Q}\left\|W\odot\left(P-\hat{P}\right)\right\|_2^2+\frac{\lambda_u}{2}\|U\|_2^2+\frac{\lambda_v}{2}\|V\|_2^2+\frac{\lambda_q}{2}\|Q\|_2^2 \quad (12)$$

其中, λ_u、λ_v、λ_q为正则化常数, W 为权重矩阵, w_{ij} 表示 u_i 访问 l_j 的重要度量, 考虑用户的签到次数的影响, 本文采用平方根的方法计算 W, 如下:

$$w_{ij}=\begin{cases}1+\sqrt{1+\gamma\times r_{ij}} & \text{if } j\in M_i^o\\ 1 & \text{otherwise}\end{cases} \quad (13)$$

其中, γ为调整参数.

3.2 SoGeoCat 模型优化

本文采用变更最小二乘 (ALS) 优化损失函数, 训练出特征矩阵 U,V 和类别特征矩阵 Q. U,V,Q 的更新公式如下:

$$u_i=\left(\lambda_u I_k+\sum_j w_{ij}\left(q_{ic_j}+\varepsilon\right)^2v_jv_j^{\mathrm{T}}\right)^{-1}\sum_j w_{ij}\left(q_{ic_j}+\varepsilon\right)p_{ij}v_j \quad (14)$$

$$v_j=\left(\lambda_v I_k+\sum_i w_{ij}\left(q_{ic_j}+\varepsilon\right)^2u_iu_i^{\mathrm{T}}\right)^{-1}\sum_i w_{ij}\left(q_{ic_j}+\varepsilon\right)p_{ij}u_i \quad (15)$$

$$q_{ic}=\frac{\left(\sum_{j\in N_c}w_{ij}\left(p_{ij}-\varepsilon\right)\left(u_i^{\mathrm{T}}v_j\right)^2\right)}{\left(\lambda_q+\sum_{j\in N_c}w_{ij}\left(u_i^{\mathrm{T}}v_j\right)^2\right)} \quad (16)$$

其中, I_k 为 k 维单位矩阵, N_c 为类别为 c 的兴趣点的集合.

4 实验

4.1 实验数据集

本实验的数据来自 Foursquare 真实数据集[13], 采集的是 2009 年 12 月至 2013 年 6 月期间在加利福尼亚的签到数据, 包括用户 ID、朋友信息、兴趣点 ID、兴趣点经纬度及其类别信息. 数据集中一共含有 2551 名用户, 13 474 个兴趣点及 124 933 条签到记录. 用户-兴趣点矩阵密度为 0.002 91. 由于 LBSNs 中存在严重的数据稀疏性, 所以 LBSNs 背景下的推荐模型准确率和召回率普遍较低. 数据集的相关内容详见表 1.

为了验证 SoGeoCat 模型的准确性, 对 Foursquare 数据集做了如下的处理.

表 1 实验数据集

数据集	Foursquare
用户数	2551
兴趣点数	13 474
类别数	10
签到记录	124 933
测试集	100 033
训练集	24 900
矩阵密度	0.002 91

1) 剔除访问少于 10 个兴趣点的用户.

2) 剔除少于 10 个用户访问的兴趣点.

3) 采用数据集中的 80% 的数据作为训练集, 剩余的 20% 作为测试集.

4.2 评价指标

本文采用准确率 (Precision) 和召回率 (Recall) 来评估推荐算法的性能, 计算公式如下:

$$P@k = \frac{1}{N}\sum_{i=1}^{N}\frac{T(u_i)\cap H(u_i)}{k} \tag{17}$$

$$R@k = \frac{1}{N}\sum_{i=1}^{N}\frac{T(u_i)\cap H(u_i)}{T(u_i)} \tag{18}$$

其中, $P@k$表示当向目标用户推荐前 k 个兴趣点时的准确率, $R@k$表示当向目标用户推荐前 k 个兴趣点时的召回率, N 为用户数, $T(u_i)$表示在测试集中 u_i 访问过的兴趣点, $H(u_i)$表示在推荐的 k 个兴趣点中击中的兴趣点.

实验中, 我们将 k 设置为: 5, 8, 10, 12, 15, 20.

4.3 推荐模型对比

为了评估 SoGeoCat 模型的性能, 本文选取三个经典模型同本模型进行对比:

IRenMF[15]采用了融合地理位置信息的矩阵分解模型, 根据地理特征将领域分为实例级别领域和区域级别领域这两个层次, 利用领域的特征进行个性化推荐;

USG[4]采用了统一的协同过滤框架, 综合考虑了用户偏好、朋友信息和地理位置信息对兴趣点推荐的影响;

ASMF-LA[13]采用了“两步走”框架, 融合用户偏好、朋友信息、地理位置信息和类别信息对兴趣点推荐的影响.

参考文献[13], 实验的相关参数设置如下:

特征矩阵维度 l 设置为 12; ζ用于调整用户偏好与地理位置影响之间的权重, 设置为 0.4; α为潜在兴趣点的偏好常量, 设置为 0.1; 调整参数ε为 0.1; 正则化常数 λ_u、λ_v分别用于调整用户特征矩阵和兴趣点特征矩阵的权重, 均设置为 0.01; λ_q用于调整类别特征矩阵的权重, 设置为 1; γ为 10.

4.4 实验结果分析

为了评估 SoGeoCat 模型的性能, 本节从推荐模型 (USG、IRenMF、ASMF-LA、SoGeoCat) 之间比较、SoGeoCat 模型中各要素影响和用户潜在兴趣点数据模型影响这三个方面进行分析, 具体内容如下.

4.4.1 推荐模型的比较与分析

在 k=5, 8, 10, 12, 15, 20 的条件下准确率和召回率分别用 $P@k$、$R@k$ 表示, 各模型的准确率和召回率见表 2.

表 2 各模型在 Foursquare 数据集中的性能

Precision	$P@5$	$P@8$	$P@10$	$P@12$	$P@15$	$P@20$
IRenMF	0.0603	0.0487	0.0452	0.0418	0.0369	0.0322
USG	0.0518	0.0424	0.0396	0.0361	0.0342	0.0297
ASMF-LA	0.0621	0.0508	0.0458	0.0427	0.0401	0.0342
SoGeoCat	0.0657	0.0529	0.0484	0.0442	0.0409	0.0363
Recall	$R@5$	$R@8$	$R@10$	$R@12$	$R@15$	$R@20$
IRenMF	0.0362	0.0448	0.0500	0.0545	0.0612	0.0711
USG	0.0308	0.0403	0.0475	0.0534	0.0601	0.0693
ASMF-LA	0.0389	0.0495	0.0533	0.0582	0.0683	0.0761
SoGeoCat	0.0391	0.0502	0.0548	0.0601	0.0698	0.0775

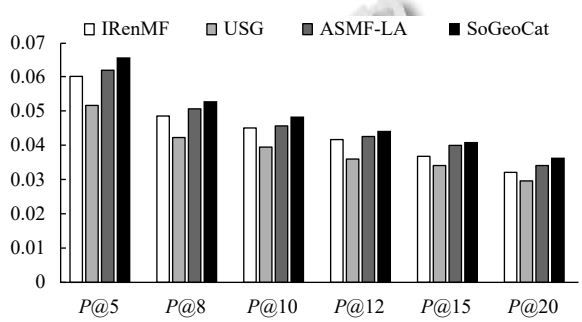

图 1 基于 Foursquare 数据集各模型的准确率对比

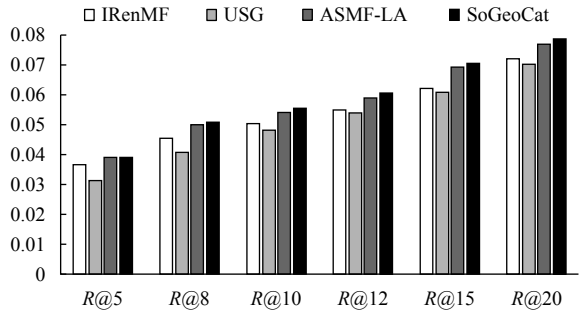

图 2 基于 Foursquare 数据集各模型的召回率对比

从表 2 中可以看出:

(1) IRenMF 采用了加权矩阵分解模型, 对于实例级别领域和区域级别领域分别采用兴趣点相似性和用户相似性进行个性化推荐, 但由于没考虑朋友信息和类别信息, 因此相对于 ASMF-LA 和 SoGeoCat 而言表现出了更差的推荐效果, 如表三所示, IRenMF 表现出了第 3 好的推荐效果;

(2) USG 是采用了融合用户偏好、朋友信息和地理位置信息的统一协同过滤模型, 但由于其没有考虑类别信息, 且各要素的影响只是进行简单的线性加权组合, 忽略了要素之间的相互作用, 再者, 协同过滤算法很容易受到数据稀疏性的影响. 所以, USG 模型表现出最差的推荐效果;

(3) ASMF-LA 采用了“两步走”框架, 考虑了直接朋友、邻居朋友、位置朋友和类别信息对兴趣点推荐的影响, 表现出了不错的推荐效果. 但是在获取邻居朋友和计算地理位置因素对兴趣点推荐的影响时, 都需要用到用户“家”的信息. 实际上, 越来越多的用户不愿意公开自己“家”的位置等隐私信息, 而且, 并非用户只愿意访问离家近的兴趣点, 如: 白领小 A, 他家和公司相离 10 公里, 他经常访问的兴趣点就容易集中在以家和公司为圆心的两个领域当中. 所以, ASMF-LA 表现出了第 2 好的推荐效果;

(4) SoGeoCat 同样采用了“两步走”框架, 既考虑到了用户之间的相似性, 又缓解了数据稀疏性, 融合了签到信息、朋友信息、地理位置信息和类别信息对兴趣点推荐的影响. 而且, 本模型中, 改进了地理位置对兴趣点推荐的影响, 根据用户的历史签到足迹来估计地理位置因素对目标用户的影响, 保护了用户的隐私信息, 表现出了最好的推荐效果.

4.4.2 要素影响分析

从图 3、图 4 中我们可以看出: (1) 三个要素对于兴趣点推荐都发挥着重要作用, 且融合三个要素时推荐效果最好; (2) 朋友信息、地理位置信息对兴趣点推荐的影响大于类别信息对于推荐的影响. 分析其原因, 主要在于用户在选择兴趣点时受到了多个方面的影响, 如朋友的介绍、距离的远近和自己的爱好等等, 所以我们不能片面地根据某一影响因素进行建模. 在 SoGeoCat 模型的第二步中运用了矩阵分解算法, 在矩阵分解算法中训练出的用户特征向量和矩阵特征向量中也有考虑到社会关系、地理位置等因素的影响, 但是在特征矩阵中没有具体地说明.

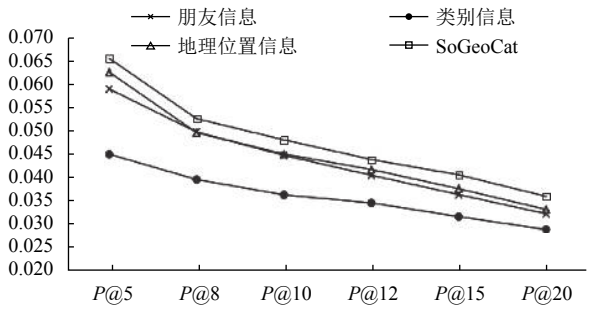

图 3 基于 Foursquare 数据集各要素间的准确率对比

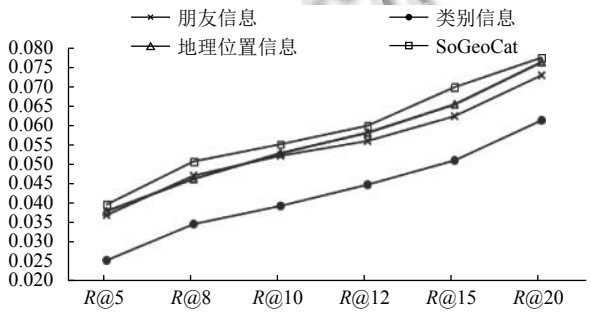

图 4 基于 Foursquare 数据集各要素间的召回率对比

4.4.3 用户潜在兴趣点数据模型影响分析

在这个部分中, 我们比较纳入用户潜在兴趣点数据模型的推荐模型和未纳入用户潜在兴趣点数据模型的推荐模型的推荐效果, 图 5、图 6 结果表明, 纳入用户潜在兴趣点数据模型的推荐效果优于未纳入用户潜在兴趣点数据模型的推荐效果. 分析其原因, 主要有两点: (1) 虽然矩阵分解算法中已经将朋友信息、类别信息和地理位置信息考虑在特征矩阵之中, 但是不能确切地说明. 我们通过用户潜在兴趣点数据模型, 单独考虑了朋友信息和地理位置信息的影响, 利于发挥其对推荐效果的影响; (2) 用户潜在兴趣点数据模型不仅考虑了这三个要素, 它还为偏好矩阵填充了大量的潜在兴趣点的签到信息, 缓解了数据稀疏性.

还有一个有趣的发现, 表 2 中只考虑类别信息的模型的推荐效果低于未纳入用户潜在兴趣点数据模型的的推荐模型的推荐效果. 因为前者在计算用户潜在兴趣点数据模型时, 没有考虑朋友信息和地理位置信息, 使得计算出来的潜在兴趣点与实际用户偏好有较大的出入, 于是将其带入矩阵分解算法中的时候产生了噪声, 影响推荐效果.

5 结论与展望

SoGeoCat 模型采用了混合算法, 融合了两种算法

的优点, 既考虑了用户之间的相似性又缓解了数据稀疏问题. SoGeoCat 模型还融合了类别标签, 保护了用户的常驻位置信息. 通过对真实的 Foursquare 数据集进行实验, 实验结果表明, SoGeoCat 模型相对于其他三个对比模型而言在 Precision 和 Recall 上都表现出较好的推荐效果.

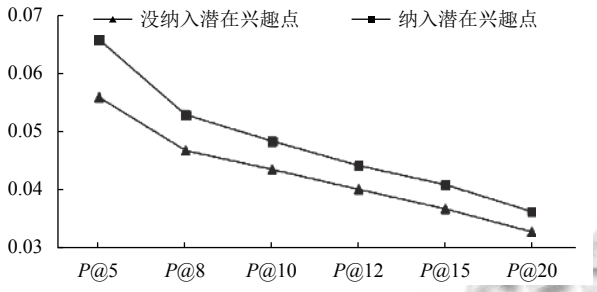

图 5 基于 Foursquare 数据集是否纳入潜在兴趣点模型的准确率对比

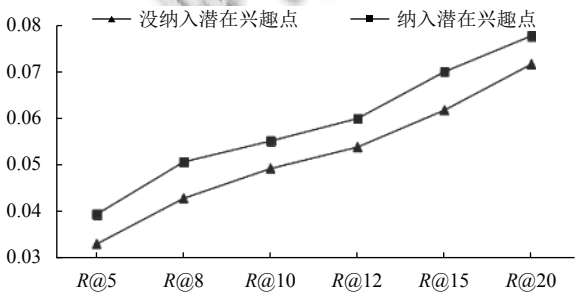

图 6 基于 Foursquare 数据集是否纳入潜在兴趣点模型的召回率对比

未来, 希望在此模型的基础上, 纳入“时间信息”和“评论信息”等上下文信息, 进一步地提高推荐算法的精确度和召回率.

参考文献

1 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.

2 余永红. 融合多源信息的推荐算法研究[博士学位论文]. 南京: 南京大学, 2017.

3 Bell RM, Koren Y. Lessons from the Netflix prize challenge. ACM SIGKDD Explorations, 2007, 9(2): 75–79. [doi: 10.1145/1345448]

4 Ye M, Yin PF, Lee WC, *et al*. 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.

5 Lian DF, 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, NY, USA. 2014. 831–840.

6 Liu YD, Pham TAN, Cong G, *et al*. An experimental evaluation of point-of-interest recommendation in location-based social networks. Proceedings of the VLDB Endowment, 2017, 10(10): 1010–1021. [doi: 10.14778/3115404]

7 Li HY, Ge Y, Hong RC, *et al*. Point-of-interest recommendations: Learning potential check-ins from friends. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA. 2016. 975–984.

8 杨志文, 刘波. 基于 Hadoop 平台协同过滤推荐算法. 计算机系统应用, 2013, 22(7): 108–112. [doi: 10.3969/j.issn.1003-3254.2013.07.024]

9 冯晓敏. 基于项目综合相似度和因子分析的协同过滤算法研究[硕士学位论文]. 青岛: 中国石油大学 (华东), 2013.

10 范波, 程久军. 用户间多相似度协同过滤推荐算法. 计算机科学, 2012, 39(1): 23–26. [doi: 10.3969/j.issn.1002-137X.2012.01.005]

11 Yuan Q, Cong G, Ma ZY, *et al*. Time-aware point-of-interest recommendation. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland. 2013. 363–372.

12 Zheng N, Jin XM, Li LH. Cross-region collaborative filtering for new point-of-interest recommendation. Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil. 2013. 45–46.

13 Li XT, Cong G, Li XL, *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.

14 高榕, 李晶, 杜博, 等. 一种融合情景和评论信息的位置社交网络兴趣点推荐模型. 计算机研究与发展, 2016, 53(4): 752–763.

15 Liu Y, Wei W, Sun AX, *et al*. 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.