基于用户兴趣和地理因素的兴趣点推荐方法

苏 畅 武鹏飞 谢显中 李 宁
(重庆邮电大学计算机科学与技术学院 重庆 400065)

摘 要 在基于位置的社交网络中,协同过滤作为目前应用最广泛的推荐技术,存在数据稀疏性和冷启动等问题。针对协同过滤算法的不足,提出了一种结合用户兴趣和地理因素的兴趣点推荐算法。该方法首先通过自适应带宽的核密度分布、朴素贝叶斯算法以及兴趣点的流行度挖掘用户的地理偏好,并根据地理偏好模型筛选出一部分候选推荐兴趣点;然后,为了克服协同过滤算法的数据稀疏性问题和用户冷启动问题,结合用户签到相似性、类别信息和用户信任度构建用户偏好模型进行兴趣点推荐;最后,使用 Yelp 数据集进行实验分析,结果表明所提出的基于用户兴趣和地理因素的兴趣点推荐模型取得了良好的推荐效果。

关键词 兴趣点推荐,地理偏好,类别信息,信任关系,协同过滤
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.036

Point of Interest Recommendation Based on User's Interest and Geographic Factors

SU Chang WU Peng-fei XIE Xian-zhong LI Ning
(College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract In the location-based social network, collaborative filtering is the most widely used recommended technology, but it has some drawbacks, such as data sparsity and cold start. In light of this, this paper presented a point of interest (POI) recommendation algorithm combining user's interest and geographic factors. In this method, firstly, the adaptive kernel density distribution, naive Bayesian algorithm and the popularity of POIs are combined to mine user's geographical preferences, and some candidate recommended POIs are screened out according to the geographical preference model. Then, in order to overcome the problems of data sparsity and cold start in collaborative filtering algorithm, a user preference model is constructed to carry out POI recommendation based on the similarities of user checked-in, category information and user trust. Finally, the Yelp data set was used to conduct the experimental analysis. The results show that the proposed POI recommendation model based on user's interest and geographical factor obtains good recommendation effect.

Keywords POI recommendation, Geographical preferences, Category information, Trust relationship, Collaborative filtering

1 引言

近年来,随着基于位置的社交网络(Location Based Social Network,LBSN)的迅速发展,其包含的数据量不断增大并呈指数级增长。面对浩瀚的信息资源,如何为用户找到感兴趣的地点已受到工业界和学术界的广泛关注。在 LBSN 中,用户可以通过 LBSN 平台(如 Foursquare, Facebook places, Yelp, Google+)分享自己的签到地点以及生活经历,并可以与其他用户建立社交联系;同时,基于位置的社交网络也能帮助人们了解周围的信息,并探索周边的环境,从而辅助用户的决策。在此背景下,出现了 LBSN 的一种新应用——兴趣点(Point-of-Interest,POI)推荐[1-5]。

LBSN 中包含了大量的数据以及丰富的多源异构信息,如社交关系[6]、兴趣点的经纬度信息,以及用户对兴趣点的评分和点评文本[7]等数据,充分利用这些信息可以有效地提高兴趣点推荐的准确性。目前,大多数的 POI 推荐方法都是根

到稿日期:2018-03-10 返修日期:2018-06-29 本文受国家自然科学基金(61271259),重庆市基础科学与前沿技术研究项目(CSTC2016jcyA0398,CTSC2011jjA40006,CTSC2012jjA40038),重庆教育委员会研究项目(KJ120501C)资助。

苏 畅(1979-),女,博士,教授,CCF 会员,主要研究方向为基于位置的社交网络分析和空间数据挖掘,E-mail:changsu@cqupt.edu.cn;**武鹏飞**(1988-),男,硕士生,主要研究方向为基于位置的社交网络兴趣点预测算法;**谢显中**(1966-),男,博士,教授,主要研究方向为移动通信技术、通信信号处理,E-mail:xiexzh@cqupt.edu.cn(通信作者);**李 宁**(1991-),男,硕士生,主要研究方向为基于位置的社交网络的空间聚类和推荐算法。

据用户的签到数据及多源异构信息来挖掘用户对尚未签到的 POIs 偏好。如 Ye 等[8]用基于内存的协同过滤算法分别构建了用户偏好和社交关系模型,同时结合用户的地理位置特征,提出经典的 USG 推荐模型。文献[9]则主要采用核密度估计的方法考虑地理因素对位置推荐的影响,并认为用户的签到行为受到的地理影响是个性化的,且根据该理论提出了一个名为 iGeoRec 的个性化地理位置推荐框架。文献[10]认为用户的签到空间由社交朋友空间和用户的兴趣空间两部分组成,并分别构建了社交朋友概率分解模型(Social Friend Probabilistic Matrix Factorization,SFPMF)和用户兴趣概率分解模型(User Interest Probabilistic Matrix Factorization,UIPMF),然后将两种分解模型进行线性融合来进行兴趣点推荐。Cheng 等[11]主要考虑了类别信息、社交影响以及地理因素,认为这 3 种特征都服从幂律分布,并结合 3 种特征构建了统一的框架,实验证明该方法显著提高了基于线性的推荐方法的准确率。

虽然近年来研究人员已提出了一些兴趣点推荐算法,但很多推荐模型均存在以下问题:在构建用户偏好模型时,只考虑了用户之间的相似性或用户签到类别的相似性而没有结合两者进行兴趣点推荐;在构建社交模型时,只使用直接的信任关系而造成那些没有朋友或朋友很少的用户很难进行社交建模。为此,本文提出了一种基于用户兴趣和地理因素的兴趣点推荐算法。该方法首先根据用户的地理偏好筛选出一部分候选 POIs,然后综合考虑用户的个人偏好和社交偏好进行兴趣点推荐。本文主要做了以下 4 方面的工作:

1)结合自适应带宽的核密度分布和朴素贝叶斯算法构建地理偏好模型,并根据该模型筛选出一部分候选 POIs;

2)在构建用户个人偏好模型时,融合了用户签到 POIs 的相似性以及用户签到类别相似性,使推荐更加准确;

3)在构建社交关系模型时,根据用户的直接信任关系进行一步信任推理,同时与用户签到行为相似性相结合,以解决数据稀疏性问题;

4)在进行兴趣点推荐时,通过将个人兴趣和由社交因素产生的兴趣进行有机结合,在一定程度上缓解了数据稀疏性和用户冷启动的问题;在 Yelp 数据集上进行实验分析,验证了所提方法的有效性。

本文第 2 节给出了一些符号的相关定义;第 3 节介绍了自适应带宽的核密度和朴素贝叶斯算法的推荐模型;第 4 节介绍了基于用户兴趣和地理因素的兴趣点推荐算法;第 5 节介绍了数据集的预处理过程,并在该数据集上验证了本文所提推荐模型的效果;最后对本文工作进行总结,并对未来的相关研究工作进行展望。

2　问题定义

定义 1(用户签到矩阵)　根据用户在 POIs 的历史签到记录,构建一个用户签到矩阵 $R_{|U|\times|L|}$,其中 $|U|$ 和 $|L|$ 分别代表用户总数和 POIs 的个数,r_{u_i,l_j} 则表示用户 u_i 在 l_j 的签到次数。

定义 2(分类偏好矩阵)　根据用户历史签到数据以及 POIs 所属的分类信息构建分类偏好矩阵 $B_{|U|\times|C|}$,其中,$|C|$ 代表类别总数。在矩阵中的元素 b_{u_i,c_a} 代表用户 u_i 的签到 POIs 中属于类别 c_a 的签到频率。其中,一个 POI 可以有多个类别信息。

定义 3(社交关系矩阵)　在基于位置的社交网络中构建一个社会关系矩阵 $S_{|U|\times|U|}$,若用户 u_i 和 u_i' 之间存在朋友关系,则 $s_{u_i,u_i'}=1$;否则,$s_{u_i,u_i'}=0$。

表 1 列出了本文所使用的关键符号及其含义。

表 1　相关符号及其含义

Table 1　Relevant symbols and their definitions

符号	含义
U	所有用户组成的集合
u_i	每个用户,并且有 $u_i\in U$
L	所有 POIs 组成的集合
l_j	某个 POI,$l_j\in L$ 且纬度和经度分别用 x_j 和 y_j 进行表示
C	所有 POIs 分类信息集合
c_a	某分类,$c_a\in C$
p_j	兴趣点 l_j 的流行度,表示所有用户在 l_j 的签到总次数

3　地理因素的构建

在基于位置社交网络的兴趣点推荐中,用户签到行为往往受距离因素的影响。一般而言,用户对一个地点签到的概率与其物理距离成反比,即距离临近的 POIs 的关联性比距离较远的 POIs 的地理关联性强。在对地理因素建模时,我们通过评估用户未签到的 POI 与已签到的 POIs 之间的关联性来计算用户访问未签到的 POI 概率。在分析地理因素对推荐概率的影响时,一般采用自适应带宽的核密度分布估计或朴素贝叶斯算法来计算用户访问每个新地点的可能性。

3.1　自适应带宽的核密度方法的构建

自适应带宽的核密度分布的构建由 3 个步骤构成:试点估计,确定本地带宽,地理相关性分数的自适应核估计。

步骤 1　试点估计。

首先,我们通过固定带宽的核密度估计方法来计算一个试点估计。当用户 u_i 在 n 个 POIs 进行签到时,其组成的 POIs 集合表示为 $L_{u_i}=\{l_1,l_2,\cdots,l_n\}$,每个 POI l_j 的经纬度坐标用(x_j,y_j)进行表示。一般而言,当用户在一个 POI 的签到次数越多时,暗示着用户对该 POI 越感兴趣,我们使用签到频率 r_{u_i,l_i} 表示用户 u_i 在 POI l_i 的签到权重。对于一个用户 u_i 未签到的 POI l_j,其试点估计的定义如下:

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

其中,l_i 表示用户 u_i 已经签到过的 POI。

$$N_{u_i}=\sum_{i=1}^{n}r_{u_i,l_i} \tag{2}$$

N_{u_i} 代表当前用户的总签到次数。

$$K_H(l_j-l_i)=\frac{1}{2\pi H_1H_2}\exp\left(-\frac{(x_j-x_i)^2}{2H_1^2}-\frac{(y_j-y_i)^2}{2H_2^2}\right) \tag{3}$$

其中,$K_H(l_j-l_i)$是包含两个全局固定带宽的标准核函数,H_1 和 H_2 的计算公式如下:

$$H_1=1.06n^{-\frac{1}{5}}\sqrt{\frac{1}{N_{u_i}}\sum_{i=1}^{n}(r_{u_i,l_i}\cdot x_i-\frac{1}{N_{u_i}}\sum_{k=1}^{n}r_{u_i,l_k}\cdot x_k)} \tag{4}$$

$$H_2=1.06n^{-\frac{1}{5}}\sqrt{\frac{1}{N_{u_i}}\sum_{i=1}^{n}(r_{u_i,l_i}\cdot y_i-\frac{1}{N_{u_i}}\sum_{k=1}^{n}r_{u_i,l_k}\cdot y_k)} \tag{5}$$

全局固定带宽 H_1 和 H_2 是根据当前用户 u_i 的历史签到数据分别计算纬度值和经度值的标准偏差而得出的。

步骤 2 确定本地带宽。

本文并不是直接使用式(1)的试点估计来计算用户对未签到 POIs 的签到概率,而是利用试点估计来计算用户 u_i 的每个签到 POI l_i 的本地自适应带宽 h_i,其计算方法如下:

$$h_i=(d^{-1}\cdot f_{KDE}(l_i|u_i))^{-\theta} \tag{6}$$

其中,θ 是灵敏度参数,其取值区间为[0,1],θ 的取值越大,本地带宽 h_i 对 $f_{KDE}(l_i|u_i)$的值越敏感。d 的定义如下所示:

$$d=\sqrt[n]{\prod_{i=1}^{n}f_{KDE}(l_i|u_i)} \tag{7}$$

步骤 3 地理相关性分数的自适应核估计。

最后,基于式(4)和式(5)计算出的全局固定带宽 H_1 和 H_2,以及式(6)计算出的本地自适应带宽 h_i,根据自适应带宽的核密度评估方法,用户 u_i 对一个未签到的 POI l_j 的推荐概率如下所示:

$$F_{KDE}(l_j|u_i)=\frac{1}{N_{u_i}}\sum_{i=1}^{n}(r_{u,l_i}\cdot K_{Hh_i}(l_j-l_i)) \tag{8}$$

$$K_{Hh_i}(l_j-l_i)=\frac{1}{2\pi H_1H_2h_i^2}\exp(-\frac{(x_j-x_i)^2}{2H_1^2h_i^2}-\frac{(y_j-y_i)^2}{2H_2^2h_i^2}) \tag{9}$$

4 基于用户兴趣和地理因素的兴趣点推荐模型

本节主要介绍所提出的基于用户兴趣和地理因素的兴趣点推荐算法,该模型利用 POIs 的分类信息、社交信息以及经纬度等信息,有效地融合了用户的个人偏好和社交影响等特征。

4.1 结合多种模型构建地理因素

在构建地理因素时,自适应带宽的核密度估计方法较适合用户的个性化地理推荐,而朴素贝叶斯算法在用户签到 POIs 个数比较多的情况下,具有低耗时并且推荐准确率较高的特点。因此,根据每个用户 u_i 访问 POIs 个数的不同,分别采用自适应带宽的核密度分布方法或朴素贝叶斯算法,即当前用户 u_i 访问 n 个 POIs 时,若 n 的值小于 v,则采用自适应带宽的核密度分布方法,否则就采用朴素贝叶斯算法。同时,结合推荐 POI l_j 的流行度 p_j 的相关信息,以便进一步提高根据地理相关性所计算出的推荐准确度。最终根据地理相关性,得到用户 u_i 对一个未签到的 POI l_j 的推荐得分,计算公式如下式所示:

$$F_{Geo}(l_j|u_i)=\begin{cases}F_{KDE}(l_j|u_i)\cdot p_j, & n<v\\ F_{NBM}(l_j|u_i)\cdot p_j, & n\geqslant v\end{cases} \tag{14}$$

4.2 用户社交影响模型的构建

在传统的社交网络以及 LBSN 服务中,用户之间的社交关系对用户的签到行为有很大程度的影响。好友对用户的强组织关系的影响力,可能使用户访问一些没有去过的商家或者地方。例如在现实生活中,一个用户通常在中式餐馆进行点菜,但在好友的邀请或影响下他很有可能也去西式餐馆吃顿牛排。在构建社交相似性时,本文同时考虑了用户之间的信任关系以及签到数据的相似性,在一定程度上克服了由于用户缺少社交关系以及签到数据稀疏性所带来的问题。其相似度的定义如下所示:

$$sim_s(u_i,u_i')=(1-\alpha)\cdot\frac{|L_{u_i}\cap L_{u_i'}|}{|L_{u_i}\cup L_{u_i'}|}+\alpha\cdot tru(u_i,u_i') \tag{15}$$

其中,α 是一个调谐参数,用于控制信任关系和签到数据的相似性所占的权重,$\alpha\in[0,1]$。$|L_{u_i}\cap L_{u_i'}|$表示用户 u_i 和 u_i'都进行过签到 POIs 的个数,而$|L_{u_i}\cup L_{u_i'}|$则表示两个用户一共访问 POIs 的数量,$tru(u_i,u_i')$表示用户 u_i 和 u_i'之间的信任关系。

楚琴用户之间的信任关系具有传递性,即两个陌生人可以通过共同好友建立起信任关系。假如用户 A 信任用户 B,而用户 B 信任用户 C,则用户 A 和 C 之间也存在信任关系。但随着传播路径越长,两个用户之间通过共同好友建立起的信任值就越小。当用户 u_i 和 u_i'为直接的朋友关系时,$tru(u_i,u_i')=1$,否则进行一步信任推理来考虑用户之间的间接社交关系,其信任推理公式采用文献[13]中的定义:

$$tru(u_i,u_i')=\frac{1}{2}\times\frac{1}{1+e^{\frac{-d}{2}}} \tag{16}$$

其中,d 表示从用户 u_i 到 u_i'的一步信任推理路径数。图 1 展示了一步用户信任推理的过程:用户 u_1 和 u_2,u_3,u_4 是直接的好友关系,则 $tru(u_1,u_2)=tru(u_1,u_3)=tru(u_1,u_4)=1$;而 u_1 与

3.2 朴素贝叶斯算法的构建

定义 $Pg[L_{u_i}]$为用户 u_i 在 POIs L_{u_i} 的签到概率,其计算公式如下所示:

$$Pg[L_{u_i}]=\prod_{l_i,l_{i'}\in L_{u_i}\wedge i\neq i'}Pg[d(l_i,l_{i'})] \tag{10}$$

其中,$d(l_i,l_{i'})$表示 POI l_i 和 $l_{i'}$之间的距离。一般而言,若两个 POIs 的距离越小,则其地理关联性越强,因此 $d(l_i,l_{i'})$与 $Pg[d(l_i,l_{i'})]$之间呈反比例关系。对于用户 u_i 尚未访问的兴趣点 l_j,其访问的概率定义如式(11)所示:

$$\begin{aligned}Pg[l_j|L_{u_i}]&=\frac{Pg[l_j\cup L_{u_i}]}{Pg[L_{u_i}]}\\&=\frac{Pg[L_{u_i}]\times\prod_{l_i\in L_{u_i}}Pg[d(l_j,l_i)]}{Pg[L_{u_i}]}\\&=\prod_{l_i\in L_{u_i}}Pg[d(l_j,l_i)]=\prod_{l_i\in L_{u_i}}\frac{1}{d(l_j,l_i)}\end{aligned} \tag{11}$$

由于待推荐的兴趣点很多,为了防止 $Pg[l_j|L_{u_i}]$的值出现向下溢出的情况,本文采用文献[12]中对访问概率取对数的操作,其定义如式(12)所示:

$$Pg'[l_j|L_{u_i}]=\sum_{l_i\in L_{u_i}}\log\frac{1}{d(l_j,l_i)} \tag{12}$$

当 $d(l_j,l_i)>1$ km 时,$\log\frac{1}{d(l_j,l_i)}$的值为负数。通过对所计算出的推荐概率进行标准化处理,最终根据朴素贝叶斯算法得到用户 u_i 对一个未签到的 POI l_j 的推荐概率定义,如式(13)所示:

$$F_{NBM}(l_j|u_i)=\frac{Pg'[l_j|L_{u_i}]-\min\limits_{l_j\in(L-L_{u_i})}\{Pg'[l_j|L_{u_i}]\}}{\max\limits_{l_j\in(L-L_{u_i})}\{Pg'[l_j|L_{u_i}]\}-\min\limits_{l_j\in(L-L_{u_i})}\{Pg'[l_j|L_{u_i}]\}} \tag{13}$$

u_6,u_7,u_8 虽不是直接的朋友关系,但是根据一步信任推理可以与其建立起联系。根据间接信任关系,我们计算用户 u_1 与用户 u_6,u_7,u_8 之间的信任权重。根据有向图的传递性,用户 u_1 到 u_6 有 1 条一步信任路径,根据式(16)的一步信任推理过程进行计算,则 $tru(u_1,u_6)=0.311$;从用户 u_1 到 u_7 有 2 条路径,则 $tru(u_1,u_7)=0.366$;而从用户 u_1 到 u_8 有 3 条路径,则 $tru(u_1,u_8)=0.409$。由此可以看出,信任推理路径数越多,其所计算出的信任权重就越大,但间接信任权重的最大值也不超过 0.5。当用户没有到达相似用户的一步信任路径时,其信任权重为 0,如从 u_1 到 u_5 没有路径,其信任度 $tru(u_1,u_5)=0$。至此,得到用户 u_1 与直接朋友和间接朋友的信任度权重。

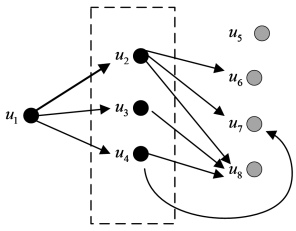

图 1　一步用户信任推理

Fig. 1　Trust inference for one step user

根据用户之间的社交相似性,选取 N' 个与用户 u_i 相似度最高的用户组成用户近邻用户集合 $U'(U'\in U)$,然后对于用户未签到的 POI l_j,根据协同过滤算法计算用户 u_i 对 POI l_j 感兴趣的程度,其公式定义如下:

$$f_{Soc}(l_j\mid u_i)=\frac{\sum\limits_{u_{i'}} sim_s(u_i,u_i')\cdot cif_{u_{i'},l_j}}{\sum\limits_{u_{i'}} sim_s(u_i,u_i')}\times\frac{1}{1+e^{-\frac{n_{soc}}{2}}} \tag{17}$$

其中,$cif_{u_{i'},l_j}$ 表示一个二进制参数,若用户 u_i' 在 POI l_j 进行了签到,则 $cif_{u_{i'},l_j}=1$,否则 $cif_{u_{i'},l_j}=0$。n_{soc} 表示有多少当前用户 u_i 的社交相似用户访问过 POI l_j,如果访问过 POI l_j 的社交相似用户很多,则表明 POI l_j 在用户 u_i 的社交相似用户中的知名度越高,那么用户访问 POI l_j 的可能性也随之增大。本文采用 $\frac{1}{1+e^{-\frac{n_{soc}}{2}}}$ 作为 POI l_j 的社交知名度,其取值范围为(0,1)。

4.3　用户个人偏好特征的构建

传统的相似度计算主要基于用户共同签到 POIs 的信息,而忽略了用户签到 POIs 所属的类别信息。本文提出结合用户协同过滤以及用户签到 POIs 所属类别的相似性,来计算用户之间的相似度。该方法在一定程度上可以解决用户协同过滤所存在的数据稀疏性问题,使得所计算出的相似性更加符合实际情况,其计算公式如下所示:

$$sim_u(u_i,u_i')=(1-\beta)\cdot\frac{\sum\limits_{l_j\in L} cif_{u_i,l_j}\cdot cif_{u_{i'},l_j}}{\sqrt{\sum\limits_{l_j\in L} cif^2_{u_i,l_j}}\sqrt{\sum\limits_{l_j\in L} cif^2_{u_{i'},l_j}}}+$$
$$\beta\cdot\frac{\sum\limits_{a=1}^{|C|} I_{u_i,c_a}\cdot I_{u_{i'},c_a}}{\sqrt{\sum\limits_{a=1}^{|C|} I^2_{u_i,c_a}}\sqrt{\sum\limits_{a=1}^{|C|} I^2_{u_{i'},c_a}}} \tag{18}$$

其中,β 是一个调谐参数,用于控制用户协同过滤所计算出的相似度和根据类别计算出的相似度的比例,其取值范围为[0,1]。而 I_{u_i,c_a} 则表示用户 u_i 对项目 c_a 的兴趣度,采用文献[14]中的定义进行计算:

$$I_{u_i,c_a}=\frac{b_{u_i,c_a}}{N_{u_i}} \tag{19}$$

其中,N_{u_i} 表示用户 u_i 的总签到次数。I_{u_i,c_a} 的值越大,表明用户 u_i 对类别 c_a 越感兴趣。

得到用户之间的相似度后,选取 N' 个与用户 u_i 相似度最高的用户组成用户近邻集合 U',然后对于用户未进行签到的 POI l_j,根据协同过滤算法计算用户 u_i 对 POI l_j 感兴趣的程度,其公式定义如下:

$$f_{pre}(l_j\mid u_i)=\frac{\sum\limits_{u_{i'}} sim_u(u_i,u_i')\cdot cif_{u_{i'},l_j}}{\sum\limits_{u_{i'}} sim_u(u_i,u_i')}\times\frac{1}{1+e^{-\frac{n_{pre}}{2}}} \tag{20}$$

其中,n_{pre} 表示有多少当前用户 u_i 的相似用户访问过 POI l_j。如果访问过 POI l_j 的相似用户很多,则表明 POI l_j 在用户 u_i 的相似用户中的知名度较高,用户访问 POI l_j 的可能性也随之增大。本文采用 $\frac{1}{1+e^{-\frac{n_{pre}}{2}}}$ 作为 POI l_j 的偏好知名度,并将其取值规定在(0,1)范围内。

4.4　用户特征偏好的构建

本文将用户的兴趣偏好分为个人兴趣和社交因素而产生的兴趣。个人兴趣主要是从用户的签到 POIs 的所属类别信息以及相似签到用户中挖掘到的长期兴趣偏好。因社交因素产生的兴趣,主要是受好友的影响而访问某 POI,一般为用户的短期兴趣偏好。通过把个人兴趣和社交因素产生的兴趣进行有机结合,可在一定程度上缓解数据稀疏性和用户冷启动的问题。本文采用线性加权的方式把两种兴趣模型融合在一起,其公式定义如下:

$$F_{pre}(l_j\mid u_i)=(1-\gamma)\cdot f_{pre}(l_j\mid u_i)+\gamma\cdot f_{soc}(l_j\mid u_i) \tag{21}$$

其中,γ 用于控制个人兴趣和社交因素在兴趣点推荐时所占的比例,其取值范围为[0,1]。

4.5　推荐 POIs

本文认为用户访问某个 POI 主要是受两方面因素的影响:用户的兴趣偏好和地理偏好。地理偏好是指临近的 POIs 的关联性比距离较远的 POIs 的地理关联性更强,用户一般遵循就近访问的原则。本文首先根据地理偏好筛选出一部分POIs,然后再根据用户的兴趣偏好计算这一部分的 POIs 的访问概率。当向用户推荐 k 个 POIs 时,推荐过程如图 2 所示,共由 4 步组成。

(1)根据式(14)计算用户在未签到的 POIs 的地理偏好推荐得分;

(2)选出得分最高的 m 个 POIs,作为候选推荐兴趣点,并且有 $m>k$;

(3)对这 m 个 POIs,根据式(21)计算用户的偏好推荐概率;

(4)选出得分最高的 k 个 POIs 推荐给用户。

将本文所提出的这种方法称为基于用户兴趣和地理因素的兴趣点推荐(POI Recommendation based on User interest and Geographical factors,PRUG)。

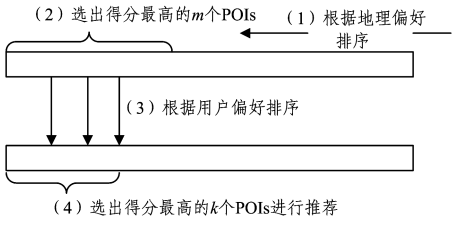

图 2 POIs 推荐流程

Fig. 2 POIs recommendation process

5 实验评估

本节主要是通过在真实数据集中进行实验来评估 PRUG 算法的有效性,并与其他类似算法进行对比实验。

5.1 数据集

采用 Yelp 数据集[15]来进行评估实验。Yelp 是美国著名商户点评网站,在该网站上用户可以在兴趣点进行签到、评分、撰写评论以及结交好友。在该数据集中,共有 2225213 条历史签到记录,签到数据的时间范围为 2004-10-12 至 2015-12-24,这些数据分布于 159 个城市。其中,在签到次数最多的 Las Vegas 城市中有 861536 次签到记录,因此本文先对数据集进行预处理。首先,选取签到 POIs 属于 Las Vegas 城市的签到记录,并在此基础上仅保存用户和兴趣点的签到记录不少于 20 条的数据作为标准数据集。经预处理所得到的标准数据集中,共有 3564 个用户和 2397 个 POIs,333 个类别信息,以及 158636 条历史签到记录。

5.2 评估指标

本文采用信息检索和统计学分类相关领域的准确率(Precision)和召回率(Recall)两个度量值作为模型方法的评估指标,其定义分别如下:

$$Precision@k=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{S_i(k)\bigcap T_i}{k} \tag{22}$$

$$Recall@k=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{S_i(k)\bigcap T_i}{|T_i|} \tag{23}$$

其中,$S_i(k)$表示给用户 u_i 推荐 k 个 POIs,T_i 表示在测试数据集中用户访问过的兴趣点集合,$|U|$表示用户的数量。本实验中分别选取 $k=\{5,10,15,20\}$来计算准确率和召回率的值,进而作为本文的评估指标。

5.3 对比模型方法

本节为了验证所提出的 PRUG 模型的有效性,选取了经典的 USG 模型以及结合类别信息进行兴趣点推荐的 GeoSoCa 模型进行对比实验。

(1)USG[8]:采用协同过滤的方法构建了用户偏好和社会影响模型,并采用朴素贝叶斯算法构建了地理因素模型,最后使用线性融合框架将 3 种特征融合在一起。

(2)GeoSoCa[16]:采用自适应带宽的核密度方法构建了 POI 推荐的地理相关性模型,利用用户类别信息以及加权流行度等构建了用户偏好模型,使用直接朋友关系构建了社会相关性模型,最后通过乘积形式将 3 种模型整合在一起。

(3)PRUG:本文所提出的推荐模型,首先结合自适应带宽的核密度分布和朴素贝叶斯算法构建 POI 推荐的地理偏好模型,并筛选出一部分候选 POIs;在此基础上,根据结合个人偏好和社交关系构建的用户偏好模型为用户推荐 k 个 POIs。

对于每个用户 u_i,本文聚合其所有签到记录,并按照用户访问时间的先后顺序进行排序,选择前 80%的记录作为本实验的训练数据集,剩余部分作为测试数据集。为了进行公平对比,参照对比模型的相应文献设置的不同参数,使各个算法都能取得最优性能。在 USG 中,$\alpha=\beta=0.1$;对于本文中所提出的 PRUG 算法,分别设 $\alpha=0.05$,$\beta=0.09$,$\gamma=0.11$,$\upsilon=18$,使其推荐性能达到最优。

5.4 实验模型对比与分析

本节讨论并总结所得出的实验结果。从图 3 和图 4 可以看出,本文所提出的 PRUG 模型在 Precision 和 Recall 这两个性能指标上相比 USG 和 GeoSoCa 模型均有所提高。当为用户推荐 5 个 POIs 时,在准确率上,PRUG 模型比 USG 模型提高了 25.21%,比 GeoSoCa 模型提高了 85.02%;在召回率上,PRUG 模型比 USG 模型提高了 22.31%,比 GeoSoCa 模型提高了 84.48%。由此可以看出,在准确率和召回率方面,PRUG 模型相对于 USG 模型和 GeoSoCa 模型均有很大幅度的提高,并且在推荐其他数量的 POIs 时准确率和召回率也有明显的提升。通过实验对比,证明了本文所提 PRUG 模型的有效性。

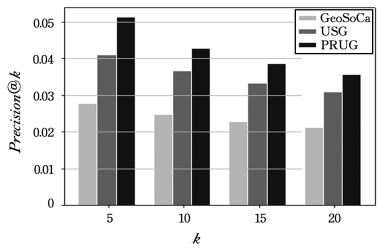

图 3 推荐准确率

Fig. 3 Recommendation precision

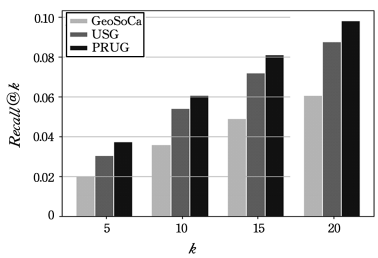

图 4 推荐的召回率

Fig. 4 Recommendation recall

接着,将本文的 PRUG 模型拆分成个人兴趣偏好 U_{PRUG}、用户社交偏好 S_{PRUG} 以及地理因素偏好 G_{PRUG};把 USG 模型拆分成用户偏好 U_{USG}、基于社交的协同过滤模型 S_{USG} 以及朴素贝叶斯构建的地理模型 G_{USG};把 GeoSoCa 模型拆分成类别偏好模型 Ca、社交偏好模型 So 以及由自适应带宽的核密度分布构建的 Geo 地理模型。通过实验比较不同模型下的拆分方法的推荐准确率和召回率。图 5 和图 6 表示当 $k=5$ 时,3 种模型的拆分方法的推荐准确率和召回率。从图中可以看出,相对其他模型,本文所提模型不仅整体推荐效果最好,而且拆

分方法也是最优的。例如在构建个人用户偏好模型时,单独使用类别信息构建 Ca 模型的推荐准确率只有0.0288,U_{USG} 模型的推荐准确率为 0.0403,但结合用户的协同过滤以及类别信息的推荐准确率提高至 0.0510,这说明结合用户的协同过滤和类别信息的有效性。同理,在构建社交影响模型时,所提 S_{PRUG} 模型相对于 S_{USG} 和 So 模型在性能方面也有明显的提升。在地理因素建模时,通过自适应带宽的核密度分布构建的 Geo 模型的推荐准确率仅为 0.0109;通过朴素贝叶斯算法构建的 G_{USG} 模型的推荐准确率为 0.0147;G_{USG} 模型与流行度相结合时其推荐准确率为 0.0352;而本文的 G_{PRUG} 模型的推荐准确率为 0.0360。由此可以看出,流行度对地理偏好模型的推荐准确率有很重要的作用,通过结合朴素贝叶斯算法、自适应带宽的核密度估计方法,以及 POI 的流行度,所构建的地理偏好模型的推荐准确率和召回率都有很大幅度的提高。

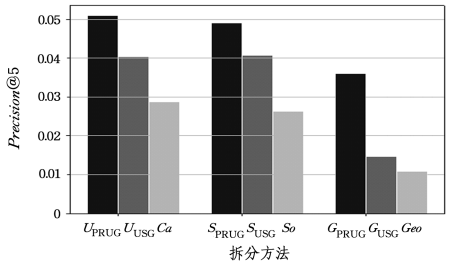

图 5　当 $k=5$ 时拆分方法的准确率

Fig. 5　Precision of split method when k is 5

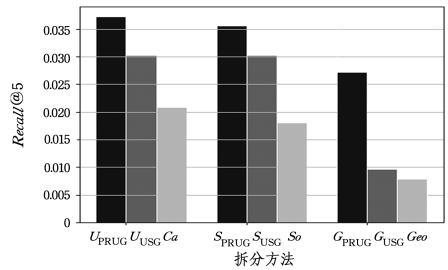

图 6　当 $k=5$ 时拆分方法的召回率

Fig. 6　Recall of split method when k is 5

5.5　参数设置

在地理偏好模型中,v 参数是在推荐时采用自适应带宽的核密度分布方法或朴素贝叶斯算法的一个重要参数。本文将 v 的值分别设置为{12,15,18,…,39,42},然后给用户推荐 5 个 POIs,观察其在不同参数下的推荐准确率和召回率,其结果如图 7 所示。

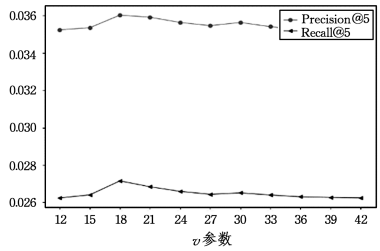

图 7　v 对准确率和召回率的影响

Fig. 7　Effect of parameters v on precision and recall

从图中可以看出,当 $v=18$ 时其推荐的准确率和召回率最高。

在进行用户社交影响模型的构建时,α 参数是控制信任关系和签到数据的相似性所占权重的调谐参数。在为用户推荐 5 个 POIs 时,其推荐的准确率和召回率如图 8 所示。从图中可以看出,当 $\alpha=0.05$ 时其推荐的准确率和召回率最高。

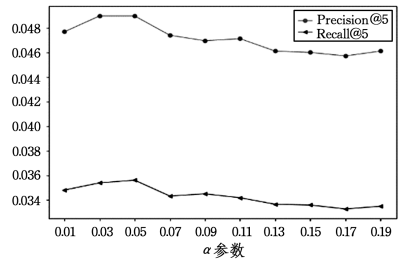

图 8　α 对准确率和召回率的影响

Fig. 8　Effect of parameter α on precision and recall

同理,在进行个人偏好模型和用户偏好模型的构建时,通过同样的方法可以验证,当 $\beta=0.09$ 和 $\gamma=0.11$ 时,其推荐效果最优。

结束语　为了解决兴趣点推荐的问题,本文提出了一种基于用户兴趣和地理因素的兴趣点推荐算法,称为 PRUG,并利用了 POIs 的分类信息、社交信息以及经纬度等信息,有效地融合了用户的个人偏好和社交影响等特征。首先,结合自适应带宽的核密度分布和朴素贝叶斯算法挖掘用户的地理偏好;其次,根据地理偏好推荐得分筛选出一部分候选推荐 POIs;然后构建用户偏好模型,通过基于用户的协同过滤和类别信息构建用户的个人偏好,结合用户之间的信任关系构建社交影响模型,并将两者进行线性融合;最后,根据用户偏好模型对候选 POIs 计算推荐概率,把得分最高的前 k 个 POIs 推荐给用户。在真实的 Yelp 数据集上进行实验,结果显示本文所提出的 PRUG 模型在准确率和召回率上相比其他的推荐方法有了明显的提高。

未来将结合用户对 POIs 的评论文本信息进行分析,使所得到的用户偏好更准确,进一步提高推荐的性能。

参 考 文 献

[1]　LI H,GE Y,ZHU H,et al. Point-of-Interest Recommendations: Learning Potential Checkins from Friends[C]// In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM Press, 2016:975-984.

[2]　YU Y,WANG H,SUN S,et al. Exploiting location significance and user authority for point-of-interest recommendation[C]// Advances in Knowledge Discovery and Data Mining- 21st Pacific-Asia Conference. South Korea: PAKDD Press, 2017: 119-130.

[3]　YANG S,HUANG G,XIANG Y,et al. Modeling User Preferences on Spatiotemporal Topics for Point-of-Interest Recom-

mendation[C]// IEEE International Conference on Services Computing. Honolulu:IEEE Press,2017:204-211.

[4] LIM K H,CHAN J,LECKIE C,et al. Personalized trip recommendation for tourists based on user interests,points of interest visit durations and visit recency[J]. Knowledge & Information Systems,2017(2):1-32.

[5] XIE M,WANG S,WANG H,et al. Learning Graph-based POI Embedding for Location based Recommendation[C]// ACM International on Conference on Information and Knowledge Management. Indianapolis:ACM,2016:15-24.

[6] FANG M Y,DAI B R. Power of Bosom Friends,POI Recommendation by Learning Preference of Close Friends and Similar Users[C]// Big Data Analytics and Knowledge Discovery. Porto:Springer Press,2016:179-192.

[7] ZHANG J D,CHOW C Y,ZHENG Y. ORec:An Opinion-Based Point-of-Interest Recommendation Framework[C]// ACM International on Conference on Information and Knowledge Management. Melbourne:ACM,2015:1641-1650.

[8] YE M,YIN P,LEE W C,et al. Exploiting geographical influence for collaborative point-of-interest recommendation[C]// Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing:ACM Press,2011:325-334.

[9] ZHANG J D,CHOW C Y,LI Y. iGeoRec:A Personalized and Efficient Geographical Location Recommendation Framework [J]. IEEE Transactions on Services Computing,2015,8(5):701-714.

[10] LI H,HONG R,ZHU S,et al. Point-of-Interest Recommender Systems:A Separate-Space Perspective[C]// IEEE International Conference on Data Mining. Barcelona:IEEE Press,2016:231-240.

[11] CHENG C,HUANG J,ZHONG N. Point-of-Interest Recommendations by Unifying Multiple Correlations[C]// Web-Age Information Management. Nanchang:Springer Press,2016:178-190.

[12] LIN K,WANG J,ZHANG Z,et al. Adaptive location recommendation algorithm based on location-based social networks [C]// International Conference on Computer Science & Education. Cambridge:IEEE Press,2015:137-142.

[13] WANG R Q,PAN J,LI Y X,Research on collaborative recommendation method based on multiple data sources of social network. Telecommunications Science [J]. Telecommunications Science,2015,31(6):78-84. (in Chinese)
王瑞琴,潘俊,李一啸. 基于多社交数据源的协同推荐方法研究[J]. 电信科学,2015,31(6):78-84.

[14] HE M,XIAO R,LIU W S,et al. Collaborative Filtering Recommendation Algorithm Combing Category Information and User Interests [J]. Telecommunications Science,2017,44(8):230-235. (in Chinese)
何明,肖润,刘伟世,等. 融合类别信息和用户兴趣度的协同过滤推荐算法[J]. 计算机科学,2017,44(8):230-235.

[15] Yelp. Challenge Data Set [OL]. (2015-12-24). http://www.yelp.com/dataset_change.

[16] ZHANG J D,CHOW C Y. GeoSoCa:Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations[C]// Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago:ACM Press,2015:443-452.