计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100120

话题-位置-类别感知的兴趣点推荐

马理博　秦小麟
南京航空航天大学计算机科学与技术学院　南京 210016
(mlbcs@nuaa.edu.cn)

摘　要　随着基于位置的社交网络(Location-Based Social Networks,LBSN)的不断发展,有助于用户探索新地点和商家发现潜在客户的兴趣点(Point-of-Interest,POI)推荐受到了广泛关注。然而,用户签到数据的高稀疏性,为兴趣点推荐带来了严峻挑战。针对这一挑战,文中探索兴趣点的文本、地理和类别信息,有效融合兴趣话题、地理影响及类别偏好因素,提出了一种话题-位置-类别感知的协同过滤兴趣点推荐算法,称之为 TGC-CF。该算法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分布,并计算用户间兴趣话题分布的相似度,通过结合地理距离和用户的区域偏好来建模地理影响;使用 TF-IDF 统计方法评估目标用户对类别的偏好程度,并考虑其他用户的类别偏好在推荐过程中的作用和影响,最后将这些影响因素整合到一个协同过滤推荐模型中,从而生成包含用户感兴趣的兴趣点的推荐列表。在两个真实数据集上的实验结果表明,TGC-CF 算法比其他推荐算法表现更好。

关键词:基于位置的社交网络;兴趣点推荐;话题模型;地理影响;协同过滤
中图法分类号　TP311

Topic-Location-Category Aware Point-of-interest Recommendation

MA Li-bo and QIN Xiao-lin
College of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 210016,China

Abstract　With the continuous development of Location-Based Social Networks(LBSN),Point-of-Interest(POI) recommendations that help users explore new locations and merchants discover potential customers has received widespread attention. However,due to the high sparsity of the users' check-in data,POI recommendation faces serious challenges. To cope with this challenge,this paper explores the textual information,geographic information,and category information,incorporating interest topics,geographical influence,and category preference factors effectively,and proposes a topic-location-category aware collaborative filtering algorithm called TGC-CF for POI recommendation. The proposed algorithm uses the Latent Dirichlet Allocation(LDA) model to learn the interest topics distribution of users and calculate the similarity of interest topics distribution among users by mining textual information associated with POIs,models geographical influence by combining geographic distance and user's regional pre-ference,uses the TF-IDF statistical method to assess the target user's preference for the category and consider the impact of other users' category preference in the recommendation process,and finally integrate these influencing factors into a collaborative filtering recommendation model to generate a list of recommendations containing POIs that users are interested in. Experimental results on two real data sets show that TGC-CF algorithm performs better than other recommendation algorithms.

Keywords　Location-based social networks,POI recommendation,Topic model,Geographical influence,Collaborative filtering

1　引言

随着智能手机等移动设备的广泛使用以及全球定位系统和无线网络技术的快速发展,一些基于位置的社交网络服务如 Foursquare,Yelp,Gowalla 等应运而生且日益流行。通过它们,用户可以对当前访问的兴趣点进行签到,并与好友分享自己的签到信息。然而,用户在面对大量的候选项时难免会眼花缭乱,难以选出自己感兴趣的兴趣点[1]。

由于城市化的快速发展和城市的不断扩张,越来越多的兴趣点(如商场、餐厅、旅游景点等)涌现出来,使得人们有了更多的出行选择和探索城市的机会。很明显,人们可能对城市里大部分的兴趣点知之甚少,尤其是当他们处于一个不熟悉的区域时,位置推荐系统便愈发显示出重要性和价值[2]。

与传统的推荐任务不同,兴趣点推荐和上下文信息(如兴

到稿日期:2019-11-15　返修日期:2020-03-27　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61373015,61728204)
This work was supported by the National Natural Science Foundation of China(61373015,61728204).
通信作者:秦小麟(qinxcs@nuaa.edu.cn)

趣点的文本信息、位置信息和类别信息)密切相关。用户访问过的兴趣点的相关文本信息(如兴趣点的类别、名称、地址等)中隐含用户偏爱的兴趣点的特征。Tobler 曾提出地理学第一定律:"任何事物都是与其他事物相关的,只不过相近的事物关联更紧密"[3]。人们自然地倾向于访问其居住地或工作地点周围的兴趣点,并且常去的地方主要集中在某几个区域。不同的用户往往有不同的类别偏好,比如美食爱好者可能更有兴趣去探索各种高品质餐厅,而健身达人则更倾向于去健身房锻炼或参加户外运动。现实生活中存在不少场景可以说明上述上下文信息在兴趣点推荐中的重要性。例如:用户 A 常到附近的区域 $S1$ 而不是离他很远的区域 $S2$ 就餐并在 foursquare 上签到,他对类别集合 D(包括中式餐馆、法式餐馆等)感兴趣,则根据他的签到信息以及从相关文本信息中提取的兴趣话题,同时考虑地理、类别偏好和兴趣话题因素,向该用户推荐较近的中式餐馆、法式餐馆,或者和他有相似兴趣话题的用户所喜爱的兴趣点。

近年来,利用上下文信息来进行兴趣点推荐的研究工作屡见不鲜,但是潜藏在用户与兴趣点交互过程中的地理影响、类别偏好影响以及文本信息中隐含的用户偏好尚未被深度挖掘出来。虽然已有不少研究工作探索了地理影响在兴趣点推荐中的作用,但这些研究工作只考虑了地理距离因素,而没有考虑地理区域因素,无法反映出用户的区域偏好,即用户的活动范围往往集中在某些区域(如市中心的商圈)。少量利用类别信息进行兴趣点推荐的研究工作主要根据用户自己的类别偏好来推算兴趣点的预测分数,而未考虑其他用户的类别偏好在推荐过程中的作用和影响。部分研究工作通过提取文本信息中的潜在特征来进行兴趣点推荐,但是由于文本信息通常是模糊且不完整的,因此特征信息的提取往往很困难。

基于上述问题和挑战,本文整合地理信息、文本信息和位置类别信息,提出了一种话题-位置-类别感知的协同过滤兴趣点推荐算法。该算法在建模地理影响时不仅考虑距离因素,也考虑了用户的区域偏好,在评估目标用户对类别的偏好程度时,不仅考虑其自身的类别偏好,也参考了其他用户的类别偏好,并使用 LDA 主题模型探索文本信息,推算出用户间的兴趣话题分布相似度,最后将 3 部分的相关分数整合到一个协同过滤推荐模型中,进而有效提升推荐效果。本文主要有以下贡献。

1)本文利用主题模型挖掘文本信息,使用 JS 散度计算用户间兴趣话题分布的相似度;利用 K-Means 聚类算法建模地理影响;使用 TF-IDF 统计方法学习用户的类别偏好,并根据其他用户对目标用户访问过的类别的偏好程度构建用户间类别偏好影响权值。

2)本文将探索地理、文本、类别信息后得到的相关分数融合到协同过滤推荐模型中,提出了一种话题-位置-类别感知的兴趣点推荐算法 TGC-CF。

3)本文使用从 Foursquare 收集的真实数据集对 TGC-CF 进行大量的实验以评估其推荐效果,实验结果表明该算法优于其他推荐算法。

2 相关工作

2.1 兴趣点推荐

随着移动互联网的快速发展,兴趣点推荐既可以为用户提供更好的定位服务和出行体验,也可以为商家提供广告投放和潜在客户挖掘等服务,受到了学术界和工业界的广泛关注。Stepan 等[4]在基于用户的协同过滤框架上整合社会信息、时间信息、地理距离信息进行兴趣点推荐。Wu 等[5]考虑地理约束和时间相似性因素,并将其融入基于项目的协同过滤模型中。Wang 等[2]提出了一个潜在概率生成模型来模拟用户在本地和外地进行签到时的决策过程。Lian 等[6]使用二维核密度估计方法将地理信息整合到加权矩阵分解算法中,以向用户推荐兴趣点。

2.2 基于文本信息的兴趣点推荐

为了实现更好的推荐效果,提供更加令人满意的位置服务,近年来越来越多的研究开始探索文本信息。一些研究通过挖掘文本评论来获取用户偏好。Xing 等[7]使用卷积神经网络提取用户评论信息中隐含的语义信息特征,并对这些特征进行分析得出用户的偏好信息。Ren 等[8]利用与兴趣点相关的文本评论配置用户与 POI 之间的话题模型,然后通过话题提取与参数学习获取用户对兴趣点的兴趣偏好。Zhu 等[9]提出了一个位置感知的 LDA 模型来挖掘用户的潜在偏好主题分布和区域在旅游景点上的潜在主题分布,并进一步推断用户可能想去的旅游景点。然而,他们探索的都是评论文本信息。本文探索由位置类别名称和地址组成的文本信息,并采用话题模型解决文本信息模糊且不完整的问题,进而结合基于用户的协同过滤技术进行兴趣点推荐。

2.3 基于地理信息的兴趣点推荐

用户在选择兴趣点进行签到时,难免会受到地理邻近性的影响,已有大量研究工作使用地理信息进行兴趣点推荐。Wang 等[10]考虑用户的当前位置,过滤离用户较远的兴趣点。Yuan 等[11]认为人们倾向于访问距离他们以前签到过的兴趣点较近的位置,因此利用幂律分布建模地理信息对用户签到行为的影响。Si 等[12]提出了一个自适应推荐算法,针对活跃用户和非活跃用户分别使用二维高斯核密度估计方法和一维幂律函数来建模兴趣点推荐过程中的地理影响。

2.4 基于类别信息的兴趣点推荐

用户访问过的兴趣点类别信息隐式反映了其在该兴趣点进行的活动,用户活动可从一定程度上体现用户访问兴趣点的偏好。然而,只有少数研究利用类别信息进行兴趣点推荐。Bao 等[13]使用树状加权类别层次结构(Weighted Category Hierarchy,WCH)建模用户偏好,并使用迭代推理模型为城市中的每个位置类别提取本地专家,然后使用考虑本地专家意见的协同过滤模型推荐兴趣点。Xian 等[14]使用度量嵌入的方法将兴趣点映射到 Category 空间,使用不同兴趣点在 Category 空间中的欧氏距离来衡量其转移概率,进而得出当前兴趣点转移到候选兴趣点的概率。

3 话题-位置-类别感知的兴趣点推荐

3.1 问题陈述

为了便于说明,令 $U=\{u_1,u_2,\cdots,u_M\}$ 为用户的集合,M 代表用户的数量;$L=\{l_1,l_2,\cdots,l_N\}$ 为兴趣点的集合,N 代表兴趣点的数量;$C=\{c_1,c_2,\cdots,c_X\}$ 为类别的集合,X 代表类别的数量;$W=\{w_1,w_2,\cdots,w_V\}$ 为与文本信息相关的所有词汇的集合,V 是唯一词的数量。根据用户在基于位置的社交网络上的历史签到数据,构建一个签到矩阵 $\boldsymbol{R}_{M*N}$,矩阵中的每个元素 $r_{u,l}\in\{0,1\}$ 代表用户 u 是否曾在位置 l 签到过,即 0 和 1 分别代表未签到和已签到。表 1 列出了本文用到的关键符号及其说明。

表 1 本文的关键符号

Table 1 Key symbols in this paper

符号	说明
M,N,X,V	用户、兴趣点、类别、词汇数目
U	用户集合
u	某用户:$u\in U$
L	兴趣点集合
l	某 POI:$l\in L$,与一对经纬度坐标对应
C	兴趣点类别集合
c	某类别:$c\in C$
W	词汇的集合
w	某词汇:$w\in W$
$\boldsymbol{R}_{M*N}$	签到矩阵

兴趣点推荐问题的定义:给定用户的签到数据(包括签到用户 ID、兴趣点 ID、位置类别、位置经纬度及签到时间),计算用户 u 未来会访问其未访问过的兴趣点 l(即 $r_{u,l}=0$)的预测分数 $S_{u,l}$,然后为用户 u 返回 top-k 个具有最高预测分数的兴趣点。

3.2 构建用户的话题模型

3.2.1 用户的兴趣话题学习

如上所述,数据稀疏性是兴趣点推荐中的常见问题。考虑到话题模型在自然语言处理领域对稀疏文本的语义解析具有良好的性能,本文建立话题模型来挖掘相关的文本信息。基于由用户签到的兴趣点的类别名称(如商场)和地址(如江苏省南京市江宁区秣陵街道太平社区)组成的文本信息,本文使用 LDA 生成模型学习用户的兴趣话题分布。文中将由同一用户签到过的兴趣点的类别名称和地址组成的文本信息聚集到一个用户文档 d_u。LDA 生成模型如图 1 所示,K 为话题数目,M 为用户数目,灰色圆圈表示可观察变量,白色圆圈表示潜在变量或超参数。此模型有两个潜在变量:1)文档-话题分布;2)话题-词语分布。

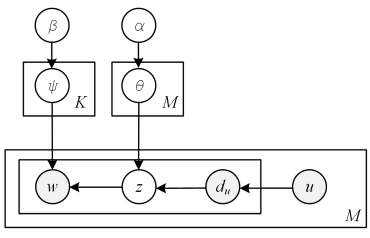


图 1 LDA 生成模型

Fig. 1 LDA generation model

在统一的话题生成模型中,每一个用户与话题的多项分布相关,每一个话题与词语的多项分布相关。用户 u 的兴趣话题多项分布表示为 θ,每个兴趣话题的词语多项分布表示为 ψ,LDA 模型的生成过程如下:

1)按照概率 $P(d_u)$ 选中一篇文档 d_u;

2)抽样生成文档 d_u 的话题分布 $\theta_{d_u}\in Dir(\alpha)$;

3)从 θ_{d_u} 中提取文档 d_u 第 n 个词的话题 $z_{d_u,n}$;

4)抽样生成话题 $z_{d_u,n}$ 的词分布 $\psi_{z_{d_u,n}}\in Dir(\beta)$;

5)从 $\psi_{z_{d_u,n}}$ 中抽样生成词语 $w_{d_u,n}$。

通过该 LDA 生成模型可推断用户 u 的兴趣话题分布 θ_u,进一步地,我们可以计算不同用户的兴趣话题分布的相似性。

3.2.2 模型参数学习

本文使用吉布斯采样方法来获得隐藏变量赋值的样本并估计未知参数,以进一步评估用户-话题分布与话题-词语分布。在采样过程中,潜在变量 z 的条件分布如下:

$$P(z_i=k\mid z_{\neg i},\vec{w})\propto\frac{n_{k,\neg i}^{(w)}+\beta_w}{\sum_{w=1}^{V}(n_{k,\neg i}^{(w)}+\beta_w)}\cdot\frac{n_{d_u,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{d_u,\neg i}^{(k)}+\alpha_k)}\tag{1}$$

其中,计数 $n_{k,\neg i}^{(w)}$ 代表除去话题 z_i 后,文档 d_u 中词汇 w 属于话题 k 出现的次数;计数 $n_{d_u,\neg i}^{(k)}$ 代表除去话题 z_i 后,文档 d_u 中话题 k 出现的次数。通过吉布斯采样算法后,最后的模型参数为:

$$\theta_{d_u k}=\frac{n_{d_u}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{d_u}^{(k)}+\alpha_k)}\tag{2}$$

$$\psi_{kw}=\frac{n_k^{(w)}+\beta_w}{\sum_{w=1}^{V}(n_k^{(w)}+\beta_w)}\tag{3}$$

其中,$n_k^{(w)}$ 是关于话题 k 的词配置频率,$n_{d_u}^{(k)}$ 是关于用户 u 的文档 d_u 的话题采样计数,V 是词汇数目,K 是话题数目。

3.2.3 用户间的话题分布相似度计算

用户的话题分布可从一定程度上反映其偏好,由于用户的偏好不会完全相同,因此不同的用户具有不同的话题分布。本文使用 JS(Jensen-Shannon)散度测量两个话题分布之间的相似性。用户 u 和 v 的话题分布的 JS 散度定义如下:

$$D_{\mathrm{JS}}(d_u,d_v)=\frac{1}{2}D(\theta_{d_u}\parallel M)+\frac{1}{2}D(\theta_{d_v}\parallel M)\tag{4}$$

其中,$M=\frac{1}{2}(\theta_{d_u}+\theta_{d_v})$,$D(\cdot\parallel\cdot)$ 是 Kullback-Leibler 距离。JS 散度的值域范围是[0,1],相同为 0,相反则为 1,因此本文将用户 u 和 v 之间的兴趣话题分布相似度 $t_{u,v}$ 定义为:

$$t_{u,v}=1-D_{\mathrm{JS}}(d_u,d_v)\tag{5}$$

3.3 建模地理影响

在移动环境中向用户推荐位置时,地理距离是一个非常重要的因素,地理临近性对用户的签到行为具有显著影响。使用聚类算法分析用户访问过的兴趣点的坐标分布时,我们观察到这些兴趣点集中在某几个区域,并且这些区域相距不远,表现出一种地理聚类现象。这种现象可能归因于地理影响,可以通过以下用户心理来直观地解释:1)人们倾向于访问距离他们的家或办公室较近的兴趣点;2)人们有兴趣探索其兴趣点周围的兴趣点。因此,下文将研究并建模地理位置对用户签到行为的影响,旨在将其用于兴趣点推荐中。

本节建模的地理影响包括距离和用户的区域偏好。已有不少研究探索地理距离因素在兴趣点推荐中的作用,一些研

究人员认为签到概率和地理距离之间的关系遵循幂律分布[15]，还有一些研究人员认为签到概率和地理距离之间成反比[16]。这些研究看似不同，但基本思想是近乎一致的：人们倾向于访问附近的兴趣点，并且随着距离的增加，访问的概率会降低。与上述研究不同，本文探索的距离因素指的是兴趣点和聚类中心的距离而不是两兴趣点间的距离。本文使用K-Means算法聚类用户访问过的位置，采用肘部法则确定超参数 A，即聚类的数目。本文建模地理距离影响的主要思想是找出用户常去的区域中离候选兴趣点最近的区域，并计算它们间的距离，距离越短，用户访问此候选兴趣点的意愿越强，反之则反。为了方便说明，令 $G_u=\{g_1,g_2,\cdots,g_A\}$，$g_A$ 表示用户 u 的第 A 个聚类中心；令 $D_{u,l}=\{d_1,d_2,\cdots,d_A\}$，$d_A$ 表示候选兴趣点 l 离用户 u 的第 A 个聚类中心的距离；令 $d_{u,l}=\min(d_1,d_2,\cdots,d_A)$ 为候选兴趣点 l 离各聚类中心的最短距离。地理距离影响的定义如下：

$$GD_{u,l}=1+2^{-d_{u,l}} \tag{6}$$

本文探索的距离因素与用户访问过的兴趣点的聚集区域相关。由于某区域可能具有更多用户感兴趣的兴趣点，因此用户访问该区域的频率更高，导致各个聚集区域包含的兴趣点数量不尽相同。聚集区域包含的兴趣点数目能从一定程度上反映用户对该区域的偏好。令 $H_u=\{n_1,n_2,\cdots,n_A\}$，$n_A$ 表示第 A 个聚类中心包含的兴趣点数目；令 $e_{u,l}$ 表示离候选兴趣点 l 最近的聚类中心所包含的兴趣点数目。用户的区域偏好影响的定义如下：

$$GN_{u,l}=2^{e_{u,l}/|L_u|} \tag{7}$$

同时考虑地理距离和用户的区域偏好，则地理影响的定义如下：

$$g_{u,l}=GD_{u,l}\cdot GN_{u,l} \tag{8}$$

3.4 类别偏好学习

实际上，兴趣点的类别能在很大程度上反映用户访问该兴趣点时进行的活动，并且人们对类别的偏好可能明显不同。因此，我们可以根据用户访问过的兴趣点类别推测用户的类别偏好。直观地，如果用户喜欢某类别，则他将访问属于该类别的更多位置。此外，如果用户访问其他人很少访问的属于某类别的位置，则其可能更加偏爱该类别。例如，在签到历史记录中访问餐馆的次数通常比寺庙等其他类别多，但由于吃饭是人们日常必需的活动，因此这并不意味着食品是人们最感兴趣的；然而，如果我们发现某用户频繁访问寺庙，则该用户对宗教或神学感兴趣的概率较高。本文用TF-IDF统计方法来评估用户对类别的偏好程度，其中用户的签到历史记录被视为文档，类别被视为文档中的词语。用户 u 对类别 c 的偏好程度 $p_{u,c}$ 的定义如下：

$$p_{u,c}=1+TF_{u,R_u}\cdot IDF_{u,L} \tag{9}$$

其中，TF_{u,R_u} 是用户 u 的签到历史中的类别 c 的TF值，由式(10)计算得出；$IDF_{u,L}$ 为该类别的IDF值，由式(11)计算得出。

$$TF_{u,R_u}=\frac{|\{u.l_i:l_i.c'=c\}|}{|R_u|} \tag{10}$$

$$IDF_{u,L}=\log\frac{\sum_{v\in U}|R_v|}{\sum_{v\in U}|\{v.l_i:l_i.c'=c\}|} \tag{11}$$

其中，$|\{u.l_i:l_i.c'=c\}|$ 表示用户 u 访问类别为 c 的兴趣点的次数，$|R_u|$ 表示用户 u 的总访问次数。

若某用户经常访问某类别，则他可能对属于该类别的场所更加了解，成为该类别的“专家”，相比于很少访问该类别的普通用户，更可能找到属于该类别的高质量场所，即他访问过的属于该类别的兴趣点对推荐更有参考价值。例如，美食家喜欢到餐馆品尝各种菜品，因而他对城中各餐馆的位置、口味、环境等可能非常了解，当为一个同样热爱美食的用户推荐兴趣点时，该美食家访问过的兴趣点将有很大的参考价值。给定一个用户 u，为其推荐兴趣点，本文方法考虑了其他用户对该用户访问过的位置类别的偏好程度，认为偏好程度越高的用户访问过的位置越具有参考性，越可能是用户 u 感兴趣的。用户 v 的类别偏好对用户 u 的参考价值 $h_{u,v}$ 由式(12)计算得出：

$$h_{u,v}=\sum_{c\in u.C}\frac{|\{v.l_i:l_i.c'=c\}|}{\sum_{v\in U}|\{v.l_i:l_i.c'=c\}|} \tag{12}$$

其中，$u.C$ 代表用户 u 访问过的类别集合，$|\{v.l_i:l_i.c'=c\}|$ 表示用户 v 访问类别为 c 的兴趣点的次数。

3.5 协同过滤POI推荐模型

上述提到的各影响因素，即用户间的兴趣话题分布相似度、地理影响和用户的类别偏好，皆可用于提升POI的推荐效果。自然地，本文提出一个协同过滤POI推荐模型来整合这些因素，模型的目标是计算用户 u 将来会访问兴趣点 l 的预测分数，进而生成推荐列表。关于兴趣点推荐，在以往的研究工作中[17-18]，乘法法则被广泛应用于融合不同的元素，并显示了高鲁棒性，因此基于乘法法则，本文用 $S_{u,l}$ 表示候选兴趣点的预测分数，定义如下：

$$S_{u,l}=\frac{\sum_{v\in T}w_{u,v}r_{v,l}g_{u,l}p_{u,l.c}}{\sum_{v\in T}w_{u,v}} \tag{13}$$

其中，$g_{u,l}$ 表示地理影响权值，由式(8)计算得出；$p_{u,l.c}$ 表示用户 u 对类别 $l.c$ 的偏好程度，由式(9)计算得出；$w_{u,v}$ 表示用户 u 和用户 v 间的相互影响权值，由式(14)计算得出；T 表示计算完所有其他用户和用户 u 间相互影响权值并排序后取排名前二十对应的用户所组成的集合。具有最高分数 $S_{u,l}$ 的top-k 个兴趣点将被推荐给用户 u。

$$w_{u,v}=sim_{u,v}\cdot t_{u,v}\cdot h_{u,v} \tag{14}$$

$$sim_{u,v}=\frac{\sum_{l\in L}r_{u,l}r_{v,l}}{\sqrt{\sum_{l\in L}r_{u,l}^2}\sqrt{\sum_{l\in L}r_{v,l}^2}} \tag{15}$$

其中，$sim_{u,v}$ 表示用户 u 和用户 v 间的余弦相似度，由式(15)计算得出；$t_{u,v}$ 表示用户 u 和用户 v 间的兴趣话题分布相似度，由式(5)计算得出；$h_{u,v}$ 表示用户 v 的类别偏好对用户 u 的参考价值，由式(12)计算得出。

值得指出的是，本文使用集合 T 而不是用户集 U 在兴趣点 l 上的签到记录的加权组合来计算预测分数的原因是：1)可以有效提高系统的运行效率；2)可以筛去和用户 u 相互影响力很低的用户，提升推荐的效果。

根据该模型，我们提出了一种话题-位置-类别感知的兴趣点推荐算法TGC-CF，如算法1所示。

算法1 TGC-CF算法

输入：签到矩阵 $\mathbf{R}_{M*N}$，类别集合C，词汇集合W，兴趣点集合L，用户集合U，目标用户u，u未访问过的兴趣点l

输出：预测分数 $S_{u,l}$

1. weightDict←[]
2. for each v in U do
3. $sim_{u,v}$←getCosSim($\mathbf{R}_{M*N}$,u,v)//计算用户间的余弦相似度
4. $t_{u,v}$←getTopicSim($\mathbf{R}_{M*N}$,W,u,v) //计算用户间的兴趣话题分布相似度
5. $h_{u,v}$←getCateInflu($\mathbf{R}_{M*N}$,C,u,v) //计算用户间的类别偏好影响权值
6. $w_{u,v}$←$sim_{u,v}$ · $t_{u,v}$ · $h_{u,v}$
7. Add {key:v,value:$w_{u,v}$} to weightDict
8. end for
9. T←getSortUser(weightDict) //对 weightDict 排序，得到排名前 20 的用户
10. $g_{u,l}$←getGeoInflu($\mathbf{R}_{M*N}$,L,u) //计算地理影响权值
11. $p_{u,l,c}$←getCatePrefer($\mathbf{R}_{M*N}$,C,u,l) //计算用户 u 对兴趣点 l 的类别偏好值
12. $S_{u,l}$←getRateScore(weightDict,$g_{u,l}$,$p_{u,l,c}$,u,l) //计算用户 u 将来会访问兴趣点 l 的预测分数
13. return $S_{u,l}$

4 实验及分析

本节将通过实验来评估 TGC-CF 算法相比其他推荐方法在 POI 推荐质量方面的优势。

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签到。本实验使用的数据集由文献[19]提供，来源于 Foursquare 的用户真实签到记录，包括从 2012 年 4 月到 2013 年 2 月共 10 个月间的纽约和东京两个城市的用户签到记录。每一条签到记录包括用户 ID、位置 ID、位置类别名称、位置的经纬度及签到时间。纽约的数据集共有 1 083 个用户和 38 471 个兴趣点，总签到记录有 227 482 条。东京的数据集中有 2 293 个用户和 61 886 个兴趣点，总签到记录有 573 703 条。纽约和东京的数据集密度分别为 0.546% 和 0.404%。数据集的统计信息如表 2 所列。

表 2 Foursquare 数据集描述

Table 2 Foursquare datasets description

数据集	用户数目	POIs 数目	签到数目	密度/%
纽约	1083	38471	227482	0.546
东京	2293	61886	573703	0.404

为了便于评估，对于数据集中的每个用户，我们随机划定用户签到历史记录的 30%作为测试数据集，剩余 70%作为训练数据集。在实验中，待评估的兴趣点推荐算法用于恢复已被划定到测试集的签到记录。

4.1.2 评价指标

利用待评估的兴趣点推荐算法计算候选兴趣点的预测分数，并返回排名前 N 的兴趣点作为对目标用户的推荐。为了评估预测的准确性，我们须找出推荐列表中和测试集中兴趣点重合的数目，重合数目越多，说明预测越准确。本文的对比实验采用两种指标来评估兴趣点推荐的质量：*Precision@N* 和 *Recall@N*。*Precision@N* 表示推荐结果中用户将来真正访问的兴趣点数量占推荐总数的比例，反映了推荐的准确性。*Recall@N* 表示推荐结果中用户将来真正访问的兴趣点数量占用户将来访问的兴趣点总数的比例，反映了推荐的全面性。

$$Precision@N=\frac{1}{|U|}\sum_{u\in U}\frac{|R(u)\bigcap T(u)|}{N} \tag{16}$$

$$Recall@N=\frac{1}{|U|}\sum_{u\in U}\frac{|R(u)\bigcap T(u)|}{|T(u)|} \tag{17}$$

其中，$R(u)$表示用户 u 的兴趣点推荐列表，$T(u)$表示用户 u 实际访问过的兴趣点集合，$|U|$表示用户的总数，N 表示推荐给用户 u 的兴趣点数目。本文实验测试 $N=5,10,20$ 时推荐算法的表现。

4.1.3 对比方法

3 个因素，即用户的兴趣话题(T)、地理影响(G)、用户的类别偏好(C)，被统一纳入本文的协同过滤推荐算法中，在评估中由 TGC-CF 表示。除了 TGC-CF，下面给出其他待评估的兴趣点推荐方法。

1)UBCF：传统的基于用户的协同过滤推荐算法。

2)GCF：一个融合了地理因素的协同过滤模型[15]。

3)CD-MF：结合邻域的影响、类别和兴趣点间地理距离的矩阵分解算法[20]。

为了进一步验证分别利用兴趣话题、地理影响和类别偏好对推荐效果的提升，本文设计了 3 种基线方法：GC-CF，TC-CF 和 TG-CF。GC-CF 是未考虑兴趣话题因素的 TGC-CF 的简化版，TC-CF 是未考虑地理影响的 TGC-CF 的简化版，TG-CF 是未考虑类别偏好的 TGC-CF 的简化版。

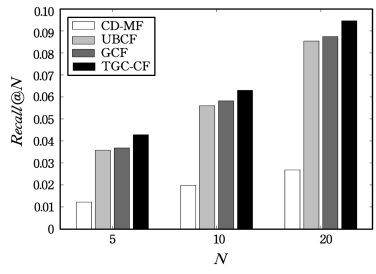
4.2 实验结果

4.2.1 方法对比结果分析

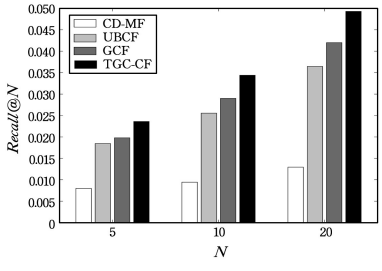
本节对比并分析各推荐方法的效果。由于数据集密度很低，如上所述，纽约和东京两城市的数据集密度分别仅为 0.546%和 0.404%，并且数据集中某些用户的签到记录很少，向他们推荐时很可能会遭遇冷启动问题，因此兴趣点推荐方法的绝对精度通常不高。本实验也不例外，其中各推荐算法的准确率和召回率大都不高于 0.1。因此，在实验中，我们专注于对比待评估的兴趣点推荐方法的相对表现。

图 2 和图 3 分别描述了在东京数据集和纽约数据集上，TGC-CF 方法对比其他先进的兴趣点推荐方法的召回率和准确率。可以看出，无论 N 取何值，TGC-CF 在两个数据集上的准确率和召回率都优于其他方法，表明了整合兴趣话题、地理影响和类别偏好 3 个因素的有效性。

GCF 使用幂率分布建模地理影响并将其整合到基于用户的协同过滤算法中；TGC-CF 通过结合地理距离和用户的区域偏好来建模地理影响，并将其整合到一个协同过滤模型中。我们发现，GCF 和 TGC-CF 的推荐效果均优于基线方法 UBCF，这证明了地理影响因素在兴趣点推荐中起着重要的作用。CD-MF 通过将用户的位置偏好和 POI 的类别信息整合到矩阵分解模型中来进行兴趣点推荐，不过相比于 TGC-CF，其只是简单地使用候选位置离用户家庭位置的距离来建模地理影响，并且忽略了文本信息，因此如图 2 和图 3 所示，CD-MF 的推荐精确度不如 TGC-CF。



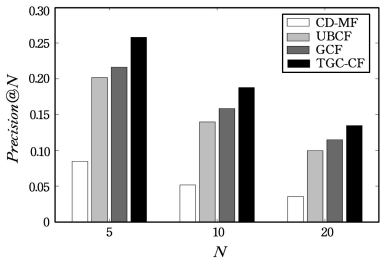
(a)东京数据集推荐结果的召回率



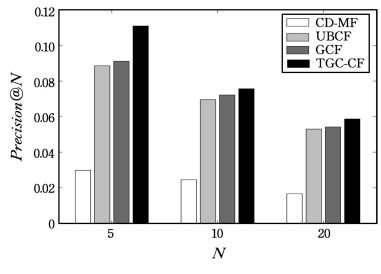
(b)纽约数据集推荐结果的召回率

图 2 两个城市数据集的召回率随 N 变化的情况

Fig. 2 Recall of two city datasets varies with N



(a)东京数据集推荐结果的准确率



(b)纽约数据集推荐结果的准确率

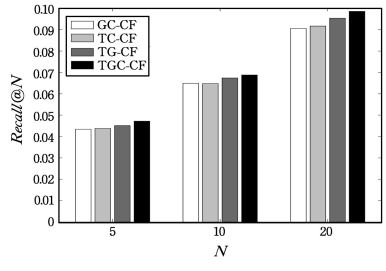
图 3 两个城市数据集的准确率随 N 变化的情况

Fig. 3 Precision of two city datasets varies with N

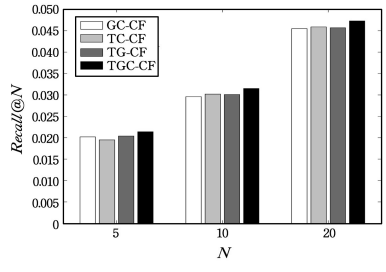
我们注意到 TGC-CF 的推荐表现相比 GCF 和 CD-MF 有了明显的提升，这主要有以下几个原因：1)TGC-CF 使用 LDA 生成模型挖掘兴趣点相关的文本信息，丰富了有效信息，从一定程度上缓解了数据稀疏性问题；2)建模地理影响时，不仅考虑了距离因素，还考虑了用户的区域偏好；3)考虑了用户的类别偏好在推荐过程中的作用；4)TGC-CF 将用户的兴趣话题、地理影响和类别偏好因素整合到一个协同过滤模型中，有效利用了文本、地理和类别信息。另外，实验中的矩阵分解推荐模型相比于基于用户的协同过滤推荐模型明显表现不佳，这可能是由于前者受冷启动问题影响较大。

4.2.2 不同因素影响分析

TGC-CF 与 3 种基线方法 GC-CF，TC-CF 和 TG-CF 的对比结果如图 4 和图 5 所示。图 4 和图 5 分别描述了在东京和纽约城市数据集上 TGC-CF 与基线方法的召回率和准确率。



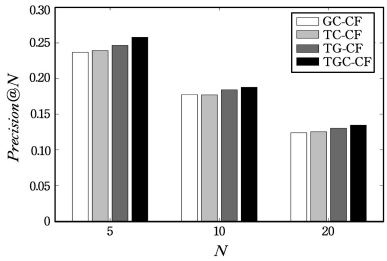
(a) 东京数据集上受不同因素影响的召回率



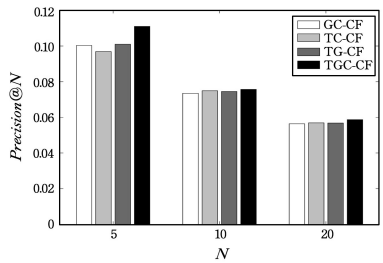
(b) 纽约数据集上受不同因素影响的召回率

图 4 两个城市数据集上受不同因素影响的召回率随 N 变化的结果

Fig. 4 Recall affected by different factors on two city datasets varies with N



(a) 东京数据集上受不同因素影响的准确率



(b) 纽约数据集上受不同因素影响的准确率

图 5 两个城市数据集上受不同因素影响的准确率随 N 变化的结果

Fig. 5 Precision affected by different factors on two city datasets varies with N

我们观察到：1)不同因素在推荐过程中的影响程度是不同的，影响程度排列如下：地理影响＞兴趣话题＞类别偏好。地理影响在使用 TGC-CF 算法进行兴趣点推荐的过程中起着关键作用，当然兴趣话题和类别偏好因素同样起着重要作用，不过由于在融合过程中因素间相互竞争，各因素产生的作用可能会被削弱，表现为图 4 和图 5 中 TGC-CF 相比 3 种基线方法的提升低于预期。2)同时整合上述 3 个因素相比于只整合其中 1 个或 2 个因素更有助于提升推荐性能，表现为

TGC-CF 的推荐精确度优于 3 种基线方法,实际上在现实生活中人们不同程度地受兴趣、地理位置和类别偏好的影响。

结束语 针对基于位置的社交网络中的兴趣点推荐,本文提出了一种话题-位置-类别感知的协同过滤兴趣点推荐算法 TGC-CF。该算法利用文本信息、地理信息和类别信息,有效地融合了兴趣话题、地理影响和类别偏好因素,从而在一定程度上缓解了数据稀疏性问题。最终,在 Foursquare 的两个真实数据集上验证了 TGC-CF 算法的推荐效果,实验结果表明该算法比其他算法表现更好。为了进一步提升兴趣点推荐的质量,未来我们准备通过提取并分析文本信息中隐含的 POI 属性或者整合时间信息等其他相关的上下文信息,来拓展本文提出的推荐模型。

参 考 文 献

[1] SASSI I B, MELLOULI S, YAHIA S B. Context-aware recommender systems in mobile environment: On the road of future research[J]. Information Systems, 2017, 72: 27-61.

[2] WANG H, FU Y, WANG Q, et al. A location-sentiment-aware recommender system for both home-town and out-of-town users [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 1135-1143.

[3] TOBLER W R. A Computer Movie Simulating Urban Growth in the Detroit Region[J]. Economic Geography, 1970, 46(Supp 1): 234-240.

[4] STEPAN T, MORAWSKI J M, DICK S, et al. Incorporating spatial, temporal, and social context in recommendations for location-based social networks[J]. IEEE Trans. Comput Soc. Syst., 2016, 3(4): 164-175.

[5] WU H, SHAO J, YIN H, et al. Geographical Constraint and Temporal Similarity Modeling for Point-of-Interest Recommendation[C]// International Conference on Web Information Systems Engineering. 2015: 426-441.

[6] LIAN D, ZHAO C, XIE X, et al. Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 831-840.

[7] XING S, LIU F, WANG Q, et al. Content-aware point-of-interest recommendation based on convolutional neural network[J]. Appl. Intell., 2019, 49: 858.

[8] REN X Y, SONG M N, SONG J D. Context-aware Point-of-interest Recommendation in Location-Based Social Networks[J]. Chinese Journal of Computers, 2017, 40(4): 824-841.

[9] ZHU Z Q, CAO J X, WENG C H. Location-time-sociality aware personalized tourist attraction recommendation in LBSN[C]// IEEE 22nd International Conference on Computer Supported Cooperative Work in Design. 2018: 636-641.

[10] WANG H, TERROVITIS M, MAMOULIS N. Location recommendationin location-based social networks using user check-in data[C]// Proceedings of the 21st ACM SIGS PATIAL International Conference on Advances in Geographic Information Systems. 2013: 374-383.

[11] YUAN Q, CONG G, MA Z, et al. Time aware point-of-interest recommendation[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2013: 363-372.

[12] SI Y L, ZHANG F Z, LIU W Y. An adaptive point-of-interest recommendation method for location-based social networks based on user activity and spatial features[J]. Knowledge-Based Systems, 2019, 163: 267-282.

[13] BAO J, ZHENG Y, MOKBEL M F. Location-based and preference-aware recommendation using sparse geo-social networking data[C]// Proceedings of the 20th International Conference on Advances in Geographic Information Systems. 2012: 199-208.

[14] XIAN X F, CHEN X J, ZHAO P P, et al. Next point of interest recommendation based on context awareness and personalized metrics[J]. Computer Engineering and Science, 2018, 40(4): 616-625.

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

[16] CHENG C, YANG H, KING I, et al. Fused matrix factorization with geographical and social influence in location-based social networks[C]// Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012: 17-23.

[17] LIU B, FU Y, YAO Z, et al. Learning geographical preferences for point-of-interest recommendation[C]// 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 1043-1051.

[18] ZHANG J D, CHOW C Y. GeoSoCa: exploiting geographical, social and categorical correlations for point-of-interest recommendations[C]// 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015: 443-452.

[19] YANG D Q, ZHANG D Q, ZHENG V W, et al. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2015, 45(1): 129-142.

[20] HU L K, SUN A X, LIU Y. Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction [C]// Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. 2014: 345-354.

MA Li-bo, born in 1997, postgraduate. His main research interests include data management and recommendation system.

QIN Xiao-lin, born in 1953, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include spatio-temporal database, distributed data management and security, etc.