

基于卷积神经网络的评论文本兴趣点推荐算法^①



申晋祥, 鲍美英

(山西大同大学 计算机与网络工程学院, 大同 037009)

通信作者: 申晋祥, E-mail: shenjinx@sxdtdx.edu.cn

摘要: 针对基于位置社交网络中的兴趣点推荐存在用户签到数据稀疏、评论文本信息利用不充分、推荐准确度不高等问题, 提出一种基于卷积神经网络的评论文本兴趣点推荐模型 (RT-CNN). 首先采用高斯函数利用邻近地理位置加权方法填补矩阵分解模型中缺少的位置信息, 预测用户对未签到位置的潜在兴趣. 然后通过卷积神经网络处理评论文本信息挖掘潜在特征, 深度提取用户情感倾向, 使用 Softmax 逻辑回归函数获得评论文本与用户和位置兴趣点潜在特征相关的概率, 通过对目标函数的求解提取用户和位置潜在特征向量. 最后融合签到行为、地理位置影响、用户情感倾向、用户潜在特征和位置兴趣点潜在特征进行兴趣点推荐. 在公开的 Foursquare 网站纽约 (NYC) 和洛杉矶 (LA) 两个真实签到数据集进行实验, 结果表明 RT-CNN 模型比其他先进的兴趣点推荐模型提高了精确率和召回率, 具有更好的推荐性能.

关键词: 卷积神经网络 (CNN); 评论文本; 兴趣点; 推荐算法; 社交网络

引用格式: 申晋祥, 鲍美英. 基于卷积神经网络的评论文本兴趣点推荐算法. 计算机系统应用, 2022, 31(8): 314-318. <http://www.c-s-a.org.cn/1003-3254/8614.html>

Point of Interest Recommendation Algorithm of Review Text Based on CNN

SHEN Jin-Xiang, BAO Mei-Ying

(College of Computer and Network Engineering, Shanxi Datong University, Datong 037009, China)

Abstract: In view of the sparse user check-in data, underutilization of review text information, and low accuracy of point-of-interest recommendation in location-based social networks, this study proposes a point-of-interest recommendation model based on review texts and the convolutional neural network (CNN), or an RT-CNN point-of-interest recommendation model for short. To start with, the Gaussian function and adjacent geographical location weighting are used to fill in the missing location information in the matrix decomposition model and thereby predict the user's potential interest in unchecked locations. Then, review text information is processed by the convolutional neural network to mine potential features and ultimately to extract the user's emotional tendencies in depth. The Softmax logic regression function is utilized to obtain the probabilities of the review text related to the potential features of a user and a point-of-interest location, and potential feature vectors of the user and the location are extracted by solving the objective function. Finally, the check-in behavior, geographical location influence, user emotion tendencies, user potential features, and potential features of point-of-interest locations are integrated to recommend points-of-interest. Experiments are carried out on two real check-in datasets, namely NYC and LA, on the public website Foursquare. The results show that compared with other state-of-the-art point-of-interest recommendation models, the RT-CNN model improves the accuracy rate and the recall rate and has better recommendation performance.

Key words: convolutional neural network (CNN); review text; point-of-interest (POI); recommendation algorithm; social network

① 基金项目: 国家自然科学基金 (11971277); 山西省教育科学“十三五”规划项目 (GH-18045); 山西大同大学校级科研专项项目 (2020YGZX016); 山西大同大学校级科研项目 (2020K10)

收稿时间: 2021-11-01; 修改时间: 2021-12-02; 采用时间: 2021-12-13; csa 在线出版时间: 2022-05-30

1 引言

近年来,随着社交网络、智能手机以及移动设备自动定位技术的发展,使得基于位置的社交网络(location-based social networks, LBSN)得以快速发展^[1],典型的LBSN应用有Foursquare、Yelp、Gowalla、Facebook、大众点评和街旁等,在LBSN中兴趣点(point-of-interest, POI)推荐是目前研究的热点。

POI推荐注重研究用户和POI的相关信息对用户行为的影响,更好地了解用户可能感兴趣的地理位置及出行活动,探索新的POI以便在正确的时间为用户进行合适的推荐,有效解决了大数据环境下“信息过载”的问题^[2-5],丰富用户的生活体验,同时也对促进旅游发展、商家精准营销等,提高经济收入具有重要意义。

基于LBSN的POI推荐已有大量研究^[6-9],其中大多数的研究都是考虑地理位置、用户签到行为、社交关系、兴趣点类别、流行度、评论文本以及时间因素等,融合多个因素进行POI推荐,实验证明推荐模型能够提高POI推荐性能。但仍然存在签到数据稀疏、评论文本信息利用不够充分等问题。高榕等人^[10]提出GeoSoRev模型,融合兴趣点的评论文本信息、用户社交关联以及地理位置信息3个因素进行POI推荐,采用基于矩阵分解的主题模型挖掘评论文本中的隐藏“主题”,模型在准确率和召回率等多个指标都有明显提高。王啸岩等人^[11]提出SoGeoCom模型,融合兴趣点评论文本信息、用户社交网络和地理位置信息进行POI推荐,采用隐狄利克雷分布(latent Dirichlet allocation, LDA)从评论文本中获取主题以及表征主题词,有效提高推荐准确率和召回率。但所述模型都是基于词袋或文档主题模型处理评论文本信息^[12],不能深度提取评论文本的潜在特征表示。目前基于卷积神经网络(convolutional neural network, CNN)技术已广泛应用于文本处理。冯浩等人^[13]提出MFM-HNN模型,融合评论文本信息和用户签到信息以提高兴趣点推荐性能,利用卷积神经网络处理评论文本信息提取特征表示,捕获更精确的上下文特征,具有更好的推荐性能,但模型没有能够基于深度学习技术融合多种上下文信息。

基于上述原因,提出RT-CNN模型,通过CNN处理评论文本内容深度提取上下文语义和情感信息,深度挖掘用户情感倾向、用户兴趣偏好以及位置兴趣点属性信息,再融合签到行为和地理影响因素进行POI推荐。实验证明模型能够有效提高推荐性能。

2 基于CNN的评论文本兴趣点推荐模型

在LBSN中有大量的历史签到数据,其中所包含的多源异构信息为深入分析用户兴趣偏好进而为用户推荐偏好的POI提供丰富的内容。主要有签到的POI地理位置、时间、POI的类别、用户的社交以及评论等信息,推荐模型可以融合多个因素,采用不同方法从签到数据中提取相关信息以便准确为用户生成POI推荐。为便于后叙内容理解,表1列出了所用符号解释。

表1 符号解释

符号	解释
U	LBSN中一组用户集合, $U=\{u_1, u_2, \dots, u_n\}$
u_i	用户集合 U 中的一个用户, $u_i \in U$
L	LBSN中一组POI位置集合, $L=\{l_1, l_2, \dots, l_m\}$
l_j	POI位置集合 L 中的一个POI, $l_j \in L$
C	LBSN中评论文本集合, $C=\{c_1, c_2, \dots, c_r\}$
$R \in \mathbb{R}^{n \times m}$	用户-兴趣点签到矩阵
R_{ij}	用户 u_i 在位置 l_j 签到频率
C_{ui}	用户 u_i 发布的评论文本集合
C_{lj}	POI位置 l_j 的所有评论文本集合

2.1 地理位置建模

通常情况用户更偏好距离自己活动空间较近的POI,正如Tobler地理学第一定律所讲的任何事物都相关,距离近的事物之间的相关性更大。在实际生活中,用户在远离其生活空间的POI签到概率较小,因此地理位置的远近对用户的签到行为有很大影响。为预测用户 u_i 对未签到位置 l_j 的签到兴趣,可由以下优化过程得到。

$$\min \frac{1}{2} (H \odot (R - UL^T))^2 \quad (1)$$

其中, $H \in \mathbb{R}^{M \times N}$ 为签到权重矩阵, H_{ij} 为1表示用户 u_i 在位置 l_j 有签到,为0表示没有签到。把矩阵参数 U 、 L 的两个正则化项加入式(1)防止过拟合,如式(2):

$$\min \frac{1}{2} \|H \odot (R - UL^T)\|_F^2 + \frac{\lambda_u}{2} \|U\|_F^2 + \frac{\lambda_l}{2} \|L\|_F^2 \quad (2)$$

其中, λ_u 、 λ_l 是正则化参数, $\|\cdot\|_F^2$ 是Frobenius范数,其优化问题采用梯度下降的方法获得局部最优解。

由事物相关性规则,如果用户对位置 l_j 的邻近若干位置感兴趣则对位置 l_j 感兴趣的可能性也很大,因此可以对矩阵分解模型中缺少的地理位置通过邻近位置加权的方法加以补充,目标函数的最小化如式(3)所示:

$$\min_{U, L} \frac{1}{2} (H \odot (R - UBL^T))^2 \quad (3)$$

其中, $B = \gamma UL^T + (1 - \gamma)A^T$, $A \in R^{n \times n}$, γ 是邻近位置影响的权重参数, $A_{j,x} = \frac{sim(l_j, l_x)}{Z(l_j)}$, $Z(l_j)$ 是正则化项, $sim(l_j, l_x)$ 是位置 l_j 邻近位置 l_x 的地理权重, 采用高斯函数如式 (4):

$$sim(l_j, l_x) = e^{-\frac{\|l_j - l_x\|^2}{r^2}} \quad (4)$$

2.2 评论文本建模

充分利用评论文本信息能够有效缓解签到数据稀疏问题, 采用 CNN 能够深度挖掘评论文本潜在特征, 将其融入 POI 推荐有重要作用。

2.2.1 用户情感倾向建模

利用 CNN 处理评论文本信息获取用户情感倾向模型, 以词向量作为输入数据, 以情感倾向作为输出数据。模型分为 4 层: 嵌入层、卷积层、池化层和输出层。具体过程如下:

1) 嵌入层: 将用户 u_i 的所有评论文本的单词合并成一个文档, 利用词向量模型将每个单词按照出现的先后顺序映射为相应的词向量, 生成用户 u_i 词序不变的词向量矩阵。如式 (5) 所示, 其中, w_p 表示第 p 个词的词向量表示。

$$M_i = (w_0) : (w_1) : \dots : (w_{n-1}) \quad (5)$$

2) 卷积层: 把输入的用户词向量矩阵 M_i 进行卷积操作提取新的特征, 每个神经元 N_q 中对应 $d \times t$ 的过滤器 F_q , d 表示嵌入维度, q 表示卷积窗口大小, 每层有 y 个神经元, 每个卷积对应产生一个特征如式 (6):

$$T_y = f(M_i \otimes F_q + b_q) \quad (6)$$

其中, f 表示激活函数 ReLU, \otimes 表示卷积操作, b_q 是过滤器 F_q 对应的偏置项。

3) 池化层: 采用最大池化操作从上下文特征向量中提取最大的特征向量生成新的固定维度的特征, 最大池化操作能够有效处理评论文本长度不同的问题, 对特征进行压缩减小其规模, 并且只提取主要特征, 即降低网络计算的复杂度又避免过拟合的现象。池化特征表示如式 (7) 所示:

$$D_j = T_y(\text{Max}(\text{allfeature})) \quad (7)$$

4) 输出层: 将池化层中提取的评论文本情感特征向量输入到 Softmax 函数, 计算各情感的预测概率并和标准实验数据对比获得误差, 采用梯度下降和反向传播进行误差传递来更新参数。

模型最后输出用户情感倾向分为 3 种: 1、0、-1, 1 表示感兴趣, 0 表示一般, -1 表示不感兴趣。结合签到权重矩阵, 通过函数重构使得情感评分 S 值在 $(-1, 1)$

范围, 显然用户签到行为与用户情感分数相关。

2.2.2 用户兴趣与位置 POI 属性建模

对于位置 POI 评论文本内容通过 CNN 能够深度提取其潜在特征, 采用 Softmax 逻辑回归函数定义用户发布评论的概率函数如式 (8) 所示:

$$P(\varphi_{il} = 1 | u_i, c_l) = \frac{e^{(u_i^T \cdot C \cdot \text{CNN}(W, C_l))}}{\sum_{C_q \in C} e^{(u_i^T \cdot C \cdot \text{CNN}(W, C_q))}} \quad (8)$$

其中, φ_{il} 表示用户 u_i 是否发布了评论 c_l , c_l 是评论文本集合, $C \in R^{q \times d}$ 是一个交互矩阵, 用来分析用户 u_i 是否发布了评论 c_l , $\text{CNN}(W, C_l)$ 表示通过 CNN 提取的评论文本特征, W 是 CNN 的内部权重。Softmax 函数的输出值相互关联, 其概率总和为 1, 要获取用户潜在特征向量 u_i , 将概率函数式 (8) 转换为目标函数式 (9) 求解得到。

$$\sum_{i=1}^n \sum_{c_q \in C_{u_i}} \log P(\varphi_{iq} = 1 | u_i, c_q) \quad (9)$$

与此同理, 分析评论 c_k 与位置 l_j 的相关性概率函数如式 (10) 所示:

$$P(\varphi_{jk} = 1 | l_j, c_k) = \frac{e^{(l_j^T \cdot P \cdot \text{CNN}(W, C_k))}}{\sum_{C_q \in C} e^{(l_j^T \cdot P \cdot \text{CNN}(W, C_q))}} \quad (10)$$

要获取位置潜在特征向量 l_j , 将概率函数式 (10) 转换为目标函数式 (11) 求解得到。

$$\sum_{j=1}^m \sum_{c_q \in C_{l_j}} \log P(\varphi_{jq} = 1 | l_j, c_q) \quad (11)$$

3 RT-CNN 模型优化及算法实现

RT-CNN 模型融合签到行为 R 、地理位置影响 B 、用户情感倾向 S 、用户潜在特征 U 以及位置潜在特征 L , 针对签到数据 R 稀疏的问题, 通过卷积神经网络处理, 充分挖掘和利用评论文本信息能够有效缓解, 其中的 S 、 U 、 L 都是从卷积神经网络中学习得到的, W 是卷积神经网络权重值, P 是概率函数。RT-CNN 模型需要解决最大化优化问题, 优化问题利用梯度下降的方法学习目标函数局部最优解。在词向量模型中, 主要有 Skip-gram 和 CBOW 两种模型, 前者是给定文中的某个词来预测上下文内容, 而后者是给定上下文来预测上下文之间的结果。Skip-gram 模型虽然预测次数要多于 CBOW, 但通过多次的预测、调整会使得词向量更加准确。因此, 通过 Skip-gram 模型获取词向量矩阵。负抽样不仅能够

减少训练过程的计算负担, 而且还能提高其结果词向量的质量, 采用负抽样方法近似计算目标函数.

RT-CNN 算法的伪代码如下算法 1.

算法 1. RT-CNN 算法的伪代码

输入: $R, C_{u_i}, u_i \in U, C_{l_j}, l_j \in L$

输出: Top-N POI

- 1) 随机初始化 W, U, L, P, C
- 2) 采用 CNN 获取用户情感倾向 S 的值
- 3) 通过 U 和 L 计算 B
- 4) While (not convergent) do
- 5) 计算 $\frac{\partial \mathcal{L}}{\partial U}, \frac{\partial \mathcal{L}}{\partial L}, \frac{\partial \mathcal{L}}{\partial C}, \frac{\partial \mathcal{L}}{\partial P}, \frac{\partial \mathcal{L}}{\partial B}$
- 6) 进行更新 $U \leftarrow U + \eta \frac{\partial \mathcal{L}}{\partial U}$
- 7) L, C, P, B 分别类同 6) 依次进行更新
- 8) 通过反向传播方法调整 CNN 中的参数
- 9) End
- 10) 计算 $U^T L$ 推荐 Top-N POI 给用户

4 实验结果与分析

4.1 实验数据集

为了验证 RT-CNN 模型的性能, 需要进行一次实验. 选用 Foursquare 网站 (https://download.csdn.net/download/weixin_41665541/10219398) 数据集中纽约 (NYC) 和洛杉矶 (LA) 的签到及评论数据, 对两个数据集进行预处理, 过滤掉签到次数小于 10 的用户和访问次数小于 10 的 POI, 实验数据集统计如表 2 所示. 数据集分成 3 份, 80% 作为训练集, 10% 作为验证集, 10% 作为测试集.

4.2 评价指标的设定

使用精确率 (Precision) 和召回率 (Recall) 作为评价指标评估算法的性能, 简记为 $P@N, R@N$, 对于一个用户 u , 其计算公式如式 (12) 和式 (13).

$$P@N = \frac{1}{|U|} \sum_{u \in U} \frac{|T(u) \cap R(u)|}{N} \quad (12)$$

$$R@N = \frac{1}{|U|} \sum_{u \in U} \frac{|T(u) \cap R(u)|}{|T(u)|} \quad (13)$$

其中, $T(u)$ 表示测试集中用户 u 签到过的 POI 列表, $R(u)$ 表示为用户 u 推荐的 POI 列表.

表 2 实验数据集统计

数据集	用户	位置 POI	签到 (评论)
NYC	45013	107237	298690
LA	30216	86925	175264

4.3 不同模型比较

选择较新的典型先进模型: SoGeoCom 模型^[9]、

TGSC-PMF 模型^[12] 和 MFM-HNN 模型^[13] 与 RT-CNN 模型比较, 设定 $N=1, 5, 10$. 基于 NYC 数据集的实验结果如图 1 和图 2 所示.

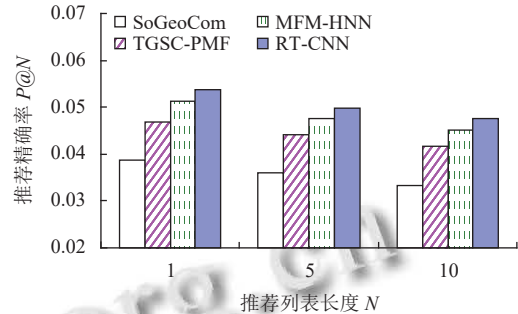


图 1 不同模型基于 NYC 数据集的精确率对比

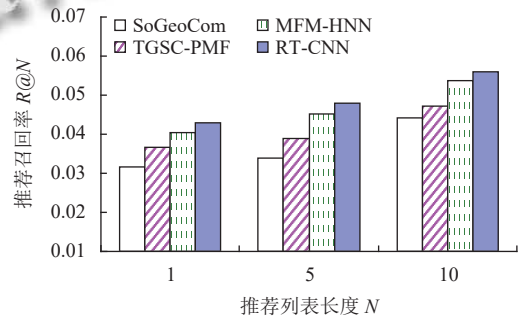


图 2 不同模型基于 NYC 数据集的召回率对比

基于 LA 数据集的实验结果如图 3 和图 4. 由实验结果分析得出, RT-CNN 模型在两个数据集上精确率和召回率都高于其他模型. 其中, SoGeoCom 模型融合用户社交、地理位置及 POI 评论文本进行推荐, 认为用户是否真正访问某 POI 是由 3 方面因素决定, 分别是用户对 POI 的兴趣、用户与 POI 之间的心理距离和物理距离. 兴趣方面可通过对用户的评论文本进行挖掘得到, 心理距离和物理距离则通过社交网络关系建模和地理位置信息建模来获得. 兴趣越大、距离越近, 访问概率就越高, 对缓解签到数据稀疏以及冷启动方面有较好的效果. 但是, 模型使用 LDA 处理评论文本信息, 忽略了评论文本语义上下文信息, 因此表现最差. TGSC-PMF 模型通过挖掘兴趣点评论文本信息, 了解用户的兴趣意向, 采用一种自适应带宽核评估方法构建兴趣点之间的地理相关性, 通过分析用户社会关系构建用户之间的社会相关性, 针对用户的分类喜好和兴趣点的流行度构建分类相关性. 最后将各相关分数进行匹配并融合到概率矩阵分解模型中. 也是使用 LDA 处理评论文本信息, 但模型融合了 POI 评论文本、地理、社交、分类与流行度信息, 并利用概率矩

阵分解模型进行有效处理,推荐性能高于 SoGeoCom 模型,表明融合多种因素能提高推荐性能。MFM-HNN 模型融合评论信息与用户签到信息进行推荐,通过 CNN 充分考虑词序及上下文信息获取评论文本潜在特征表示,克服了 LDA 处理评论文本信息所存在的问题,并利用深度堆栈降噪自动编码器研究了特征矩阵的初始化问题,推荐性能高于 TGSC-PMF 模型。所提模型 RT-CNN 通过 CNN 深度挖掘评论文本信息,获取用户情感倾向、用户兴趣偏好以及位置 POI 属性信息,融合签到行为以及地理位置影响进行推荐,在签到数据稀疏的情况下充分利用评论文本隐含的潜在语义和情感信息,有效地提高了位置 POI 推荐性能,实验结果表明,RT-CNN 模型具有最好的推荐效果。

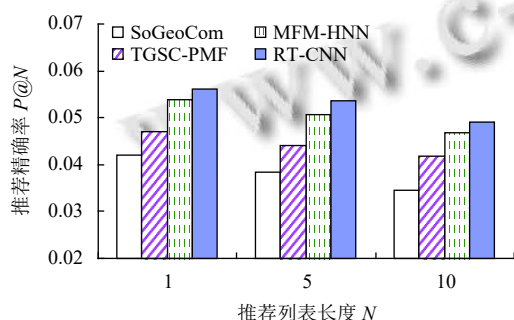


图3 不同模型基于 LA 数据集的精确率对比

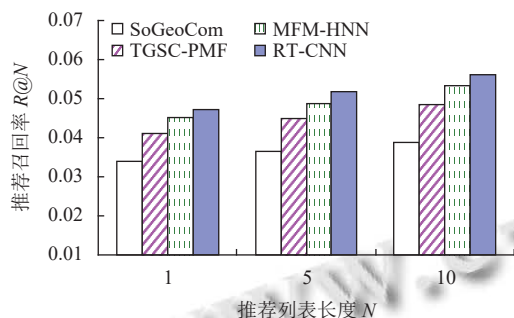


图4 不同模型基于 LA 数据集的召回率对比

通过不同模型在两个数据集上精确率和召回率的实验结果可以看出,随着推荐列表长度 N 值的增大,推荐精确率 $P@N$ 逐渐减小,推荐召回率 $R@N$ 逐渐增大。这是因为推荐列表中 POI 的数目越多,用户真正访问的 POI 在推荐列表中的比例显然会降低,也就是说在推荐列表中有许多列出的 POI 并不是用户真正访问的 POI,从而导致推荐精确率下降。然而随着推荐列表中 POI 数目的增加,则会有更多的用户真正访问的 POI 出现在推荐列表中,使得推荐召回率增加。

5 结束语

为缓解位置 POI 推荐中签到数据稀疏问题,提出一种 RT-CNN 模型,基于卷积神经网络深度提取评论文本内容的隐含信息,对位置 POI 和用户进行建模,同时融合签到行为以及地理位置信息进行 POI 推荐。通过实验与其他模型对比,结果表明模型提高了精确率和召回率,具有更优的推荐效果。

参考文献

- 刘辉,曾斌,刘子恺.融合邻居选择策略和信任关系的兴趣点推荐.计算机工程与科学,2020,42(2):365-372.[doi:10.3969/j.issn.1007-130X.2020.02.023]
- 李志刚,郝振霆.基于灰色理论的图书推荐模型的仿真分析.计算机仿真,2017,34(3):391-394.[doi:10.3969/j.issn.1006-9348.2017.03.088]
- 左怀见,徐聪,卢胜强,等.基于深度学习的芯片位置检测系统.制造业自动化,2021,43(5):18-21,48.[doi:10.3969/j.issn.1009-0134.2021.05.005]
- 宋思聪,王兴芬,杜惠英.基于改进深度学习算法的文本极性智能判断方法研究.现代电子技术,2020,43(1):76-79,85.
- 郭原东,雷帮军,聂豪,等.基于深度学习的智能高精度图像识别算法.现代电子技术,2021,44(4):173-176.
- 夏永生,王晓蕊,白鹏,等.基于时序和距离的门控循环单元兴趣点推荐算法.计算机工程,2020,46(1):52-59.
- 张岐山,李可,林小榕.基于时间效应的兴趣点推荐混合模型.计算机工程,2019,45(8):203-209.
- 许朝,孟凡荣,袁冠,等.融合地点影响力的兴趣点推荐算法.计算机应用,2019,39(11):3178-3183.
- 张进,孙福振,王绍卿,等.融合社交关系与地理信息的兴趣点推荐模型.计算机工程与应用,2020,56(5):173-178.[doi:10.3778/j.issn.1002-8331.1811-0290]
- 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.[doi:10.7544/issn1000-1239.2016.20151005]
- 王啸岩,袁景凌,秦凤.位置社交网络中基于评论文本的兴趣点推荐.计算机科学,2017,44(12):245-248,278.[doi:10.11896/j.issn.1002-137X.2017.12.044]
- 任星怡,宋美娜,宋俊德.基于位置社交网络的上下文感知的兴趣点推荐.计算机学报,2017,40(4):824-841.
- 冯浩,黄坤,李晶,等.基于深度学习的混合兴趣点推荐算法.电子与信息学报,2019,41(4):880-887.[doi:10.11999/JEIT180458]

(校对责编:孙君艳)