基于稀疏轨迹数据的出租车载客区域推荐

廖祝华,张 健,刘毅志,肖 浩,赵肆江,刘建勋
(湖南科技大学计算机科学与工程学院,湖南湘潭 411201)

摘 要: 基于短期出租车轨迹数据的载客区域推荐能极大减少系统开销,提高推荐效率,但常伴随着数据稀疏性的问题.针对该问题,本文提出了一种融合地理信息的隐语义模型-GeoLFM.该模型通过将出租车司机所处的客观地理环境信息,融合到司机-载客区域矩阵分解的过程中,从而弥补数据稀疏性带来的不足.同时,根据出租车实时的空间位置信息,为身处不同地点的出租车推荐不同的载客区域.实验证明,本文提出的方法与常用方法相比,推荐结果与真实的出租车司机载客情况间的平均绝对误差和均方根误差都得到大幅降低,较好的提升了推荐效果.

关键词: 轨迹挖掘;载客推荐;数据稀疏性;隐语义模型;地理信息
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2020)11-2178-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.11.012

Taxi Pick-Up Area Recommendation Based on Sparse Trajectory Data

LIAO Zhu-hua, ZHANG Jian, LIU Yi-zhi, XIAO Hao, ZHAO Yi-jiang, LIU Jian-xun
(*College of Computer Science and Engineering, Hunan University of Science & Technology, Xiangtan, Hunan* 411201, *China*)

Abstract: Taxi pick-up areas recommendation based on the short-term taxi trajectory data can greatly reduce the system overhead and improve the efficiency of the recommendation, but it often has the problems of data sparseness. For this reason, a Latent Factor Model integrated with the geographic information, called GeoLFM, is put forward. This model makes up the faultiness of data sparseness by integrating the geographic information relating to drivers into the Matrix decomposition, which records the visiting relationship between drivers and pick-up areas. Meanwhile, different pick-up areas can be recommended for the taxis in various locations according to the real-time spatial context of the taxis. Experimental results show that, with the comparison between our proposed method and others, the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) between the recommended results and the actual value are significantly reduced, which indicates the recommendation effect is better improved.

Key words: trajectories mining; pick-up recommendation; data sparsity; latent factor model; geographic information

1 引言

随着物联网和大数据等技术的发展,滴滴出行、UBER 等网约车平台不仅为司机提供了在线预订服务,还记录了司机的使用记录和出租车的行驶轨迹等信息.系统通过挖掘这些数据中相关有价值的信息,可以为司机提供更好的推荐服务.而载客区域推荐[1]能够帮助司机有效提高寻找乘客的效率,减少空载巡游带来的能耗和尾气排放.

与传统的推荐系统相比,出租车载客区域推荐面临新的挑战.首先,不断增长的出租车轨迹数据会耗费大量存储资源并降低推荐效率[2].其次,很多城市不断新建或翻新道路,如果使用过多历史轨迹数据,反而引入噪声数据,导致推荐性能下降.因此,如果使用短时间内能反映司机的载客选择规律的轨迹数据,则能大幅度减少噪声数据.然而,出租车每天到达的载客区域非常有限,使得该方法将面临数据稀疏性的问题[3].此外,真实地理环境中客观存在的地理信息,如 POI(Point Of Interest)数量、类型等,也是司机关注的内容,但随着不同地理位置的变化,司机感兴趣的载客区域也会动

收稿日期:2019-12-04;修回日期:2020-05-26;责任编辑:孙瑶
基金项目:国家科学自然基金(No. 61370227, No. 41871320);湖南省自然科学基金(No. 2017JJ2081, No. 2018JJ4052);湖南省教育厅重点项目(No. 17A070);湖南省教育厅一般项目(No. 19C0755)

态变化[4].

为应对上述问题,本文提出基于稀疏轨迹数据的出租车载客区域推荐方法,其主要贡献在于:

(1) 提出一种融合地理信息的隐语义模型. 隐语义模型具有抗稀疏能力,而地理信息隐含载客需求变化,所以该方法既能有效应对数据稀疏性,又能提升推荐性能.

(2) 改进了基于当前位置信息的出租车载客推荐方案. 定义了基于高斯分布的网格可达度,优化了载客区域推荐结果,能够为不同司机提供位置服务(LBS).

(3) 在短期且真实的数据上,通过多组以平均绝对误差与均方根误差作为评价指标的对比实验,证明我们的方法具有较好效果.

2　相关工作

近年,基于轨迹数据的出租车载客区域推荐系统已成为国内外的研究热点.

其中,有一些研究是定量分析出租车载客点的空间分布情况,可以为司机推荐载客区域. Kong 等人[1]提出了一种结合时空信息的出租车服务推荐模型. 该模型通过网格划分的方式将研究区域分割为较小的网格,然后使用高斯过程回归模型和相应的统计方法预测各网格中的乘客分布. Liu 等人[5]使用改进的密度聚类算法,从历史载客点中提取出租车的候选载客点,然后根据候选点的位置、附近载客点数量等信息,并使用概率优化模型为用户提供推荐结果. Wang 等人[6]提出使用线密度探测模型对出租车上下客事件进行探测和分析,以获取出租车上下客的时空分布规律. Yang 等人[7]基于大规模的出租车轨迹数据,分析城市不同功能区的潜在乘客分布,以为出租车司机推荐载客效益高的区域. Yuan 等人[8]分别从出租车司机与乘客的角度计算利润较高的区域,然后根据出租车司机当前的时空上下文进行推荐.

另外一些研究是定量分析出租车司机对载客点的偏好情况,以为其提供载客区域的个性化推荐. Ren 等人[9]分别从地理、兴趣、社会等方面量化用户对候选区域的偏好程度,然后结合概率矩阵分解模型(PMF)为用户进行区域推荐. Liu 等人[10]使用网格划分城市区域,然后分析网格的载客率及载客热度等,以发现司机的寻客偏好. Jiang 等人[11]通过出租车司机的偏好信息,如:寻客策略、载客策略等构造司机的服务策略,然后通过服务策略预测各自的收益情况. Zhang 等人[12]通过比较不同司机的寻客策略、载客策略以及各自的服务范围,挖掘司机的偏好.

上述方法对大规模出租车轨迹数据具有良好的推荐效果. 但是,目前仍缺乏短期内,较少数据量情况下可用数据高度稀疏的解决方案. 本文结合前期研究成果,兼顾出租车历史载客点分布与司机对载客点的偏好,在应对数据稀疏的情况下,使用融合地理信息的隐语义模型,并基于司机当前的位置,为不同司机提供优良的个性化推荐服务.

3　出租车载客区域推荐框架

3.1　相关定义

定义 1　轨迹　出租车在运营过程中,由 GPS 记录仪按照固定频率记录其行驶的路径信息. 每条 GPS 信息 log 所包含的字段有:司机的标识(id)、时间(t)、经纬度(lot,lat)、海拔(alt)和状态(sta)等.

图 1 描述了出租车轨迹中的状态变化过程[13]. 图中 P 表示载客点,D 表示下客点. 出租车在空载时,其轨迹点中的状态为“0”;在载客情况下,其轨迹点中的状态为“1”. 本文中我们提取所有下客点“D”与载客点“P”进行分析.

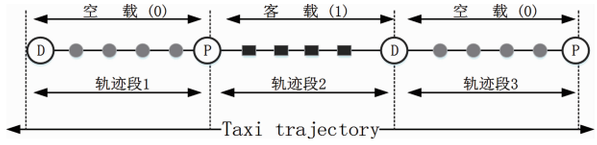


图1　轨迹状态示意图

定义 2　网格　为了增强系统的并行性,减少计算量,我们将出租车的载客区域划分为同等大小的正方形区域[1]. 整个区域由多个小正方形网格所覆盖,即为网格集合 $\text{GridSet} = \{g_1, g_2, g_3, \cdots, g_n\}$.

定义 3　司机-网格访问矩阵($\boldsymbol{C}$)　若司机 u 在网格 i 范围内载到一名乘客,则记为司机 u 对网格 i 的一次访问. 矩阵 $\boldsymbol{C}$ 的行表示司机的集合,列表示网格的集合. 矩阵中的元素 $\boldsymbol{C}_{ui}$ 表示在历史轨迹中,司机 u 对网格 i 的访问次数与该司机所有访问网格总次数的比值. 该比值在一定程度上体现了该司机对处在不同位置的网格的偏好程度. 参考文献[13],我们将一天的轨迹按时间等分为 48 个段. 网格属性值每 0.5h 更新 1 次,以适应司机在不同时段的偏好的变化.

定义 4　网格-属性矩阵($\boldsymbol{X}$)　该矩阵的行表示网格集合,列表示网格属性的集合. 网格的属性包括:历史载客点数量、POI 数量、网格类型、网格几何中心位置和网格历史载客距离,分别由 g. pickups、g. poinum、g. type、g. lot、g. lat 和 g. dist 表示. 其中 g. dist 表示所有从网格 g 出发的出租车在载客状态下行驶路程的平均值;g. pickups 表示网格 g 内所有载客点数量;g. type 表示网格 g 的类型分布;g. poinum 与(g. lot,g. lat)分别指网格 g 中所有的 POI 数量之和与网格几何中心的经纬度. 同司机-网格访问矩阵一样,访问矩阵内容每 0.5h 更新 1 次.

定义5 寻客距离 指出租车在寻客过程中所行驶的距离. 具体指出租车从上一位乘客下车地点到下一位乘客上车地点的实际行驶路程.

定义6 网格可达度矩阵(A) 可达度表示司机从一个网格到另一个网格的便捷性,A_{ij}表示从网格g_i到网格g_j的便捷性. 其表达式如式(1)所示.

$$A_{ij}=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(dist(g_i,g_j)/\beta-\mu)^2}{2\sigma^2}\right) \tag{1}$$

其中$dist(g_i,g_j)$表示网格间的距离,由于出租车的载客区域一般与其当前的位置不远,因此本文用两网格几何中心点间的欧式距离表示. β是距离缩放因子,用于将距离缩放到一个较小的范围中,以减小计算结果的波动. μ和σ^2是均值与方差. 我们从轨迹数据集中随机抽取部分出租车的寻客距离,通过核密度估计[14]并生成概率密度曲线后发现,A与网格间的距离近似服从正态分布,即距离越近,便捷程度越高,可达度越大. 司机寻客时,在其他影响因素一定的情况下,若当前所在网格g_i与目标网格g_j的可达度越大,则其去往网格g_i的可能性会越大. 此外,我们假设从g_i到g_j的可达度与从g_j到g_i的可达度相等.

3.2 系统框架

图2是系统架构图,其主要由三大部分组成,而它们涉及到的源数据主要是POI数据与出租车的轨迹数据. 首先,我们以固定大小的网格划分区域,并将POI与轨迹中的载客点映射到相应网格中. 再以网格为单位分别构建属性集:历史载客点数量、POI数量、网格类型、网格几何中心的位置和网格的历史载客距离,形成网格-属性矩阵X. 然后,从出租车轨迹信息中提取载客点[1],并根据载客点的分布情况,构造司机-网格访问矩阵C. 接着进行矩阵分解,并融入矩阵X,以最小化该分解的损失值为目标,得到两个低秩矩阵. 再通过这两个低秩矩阵的乘积得到司机对网格的偏好情况. 最后结合司机当前的位置与网格可达度矩阵,推荐符合相应偏好且距离较近的网格列表给司机.

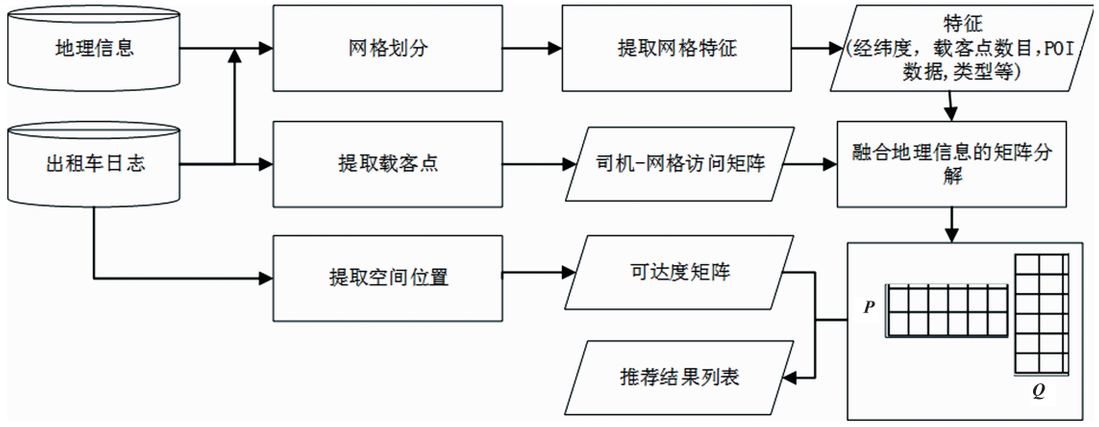


图2 系统架构图

4 出租车载客区域推荐算法

4.1 地理信息隐语义模型

矩阵分解具有一定的抗稀疏性的特质[15],其中,以隐语义模型(Latent Factor Model, LFM)[16]为代表的矩阵分解方法的抗稀疏能力较强. 本文通过分解司机-网格访问矩阵$\mathbf{C}$,得到司机-隐因子关系矩阵P与网格-隐因子关系矩阵Q. 其中,P的列与Q的行均表示隐因子的集合,而P的行与Q的列分别表示司机的集合与网格的集合,以此将司机与网格分别映射到具有相同的隐含语义空间中,这样用户对网格的访问概率就是P和Q对应项的内积表达. 求解过程以最小化分解后的矩阵与原始矩阵的差值为目标,分解过程的实现可表示成解决式(2)的优化问题:

$$loss=\min_{P,Q}[\ \|C-PQ\|^2+\lambda(\|P\|^2+\|Q\|^2)\ ] \tag{2}$$

这里的$C\in\mathbf{R}^{M\times N}$,$P\in\mathbf{R}^{M\times K}$,$Q\in\mathbf{R}^{N\times K}$. 式(2)的后半部分是正则化项,$\lambda$用于控制$P$和$Q$平方项的贡献度. 通过随机梯度下降的方法[13]来最小化差值$loss$,以得到矩阵P、Q,它们的乘积PQ近似等于原矩阵,但其中零元素的个数远少于原始矩阵,从而补充原矩阵的缺失数据.

司机对各网格的偏好概率与网格的属性具有相关性. 因此,基于客观的地理数据往往能够更加全面地描述司机载客的内在规律. 我们参考已有工作[5]并通过相关分析,发现网格的历史载客点数量、POI数量、网格类型、网格几何中心位置和网格历史载客距离对于出租车载客区域推荐的准确性有着重要影响.

网格类型主要根据网格内POI在不同类型上的分布所决定. 为完成网格类型值到数值的转换,我们采用数据预处理中常用的独热(one-hot)编码的方式,将网格类型属性扩展为若干个子类型属性. 本文以网格为

单位,用各个子类型数量在总 POI 数量中的占比,作为该网格相应类型属性的取值. 每种子类型的概率定义如式(3),其中 $|\mathrm{POI}_{\mathrm{type}_j}|$ 表示第 j 种类型的 POI 数量.

$$\mathrm{g.type}_j = \frac{|\mathrm{POI}_{\mathrm{type}_j}|}{\sum_{i=1}^{n}|\mathrm{POI}_{\mathrm{type}_i}|} \tag{3}$$

网格 g 的历史载客距离主要指从网格 g 出发的出租车在载客状态下行驶路程的平均值,其定义如式(4)所示.

$$g.dis = \left(\sum_{i=1}^{n} dist(pseg_i)\cdot c_i\right) \Big/ \sum_{i=1}^{n} c_i \tag{4}$$

其中,$pseg$ 表示历史轨迹中所有出租车在载客状态下的行驶轨迹的集合,n 为这些轨迹数量之和. $dist(\,)$ 表示计算一段轨迹的路程. c_i 取值为 0 或 1,当载客行程的载客点位于网格 g 内部时,c_i 取值为 1;否则,c_i 取值为 0.

本文考虑融入上述因素到隐含因子分解的过程中,即融合地理信息隐语义模型 GeoLFM. 该模型的表达如式(5)所示. 该模型以最小化损失函数 $loss$ 为优化目标. 其中: $\boldsymbol{X}\in\mathbf{R}^{N\times l}$ 为网格属性矩阵,表示 N 个网格在 l 种属性下的取值矩阵,其矩阵元素都可以通过历史数据获得;$\boldsymbol{P}$、$\boldsymbol{Q}$ 和 $\boldsymbol{T}$ 是需要求解的三个子矩阵,其中 $\boldsymbol{T}$ 为网格属性-隐因子关系矩阵,且 $\boldsymbol{T}\in\mathbf{R}^{l\times k}$,它的行是网格属性的集合,列为隐因子的集合,主要表示 l 种属性在 K 种隐含因子空间下的取值矩阵,其在式(2)中的正则化部分加入了对 $\boldsymbol{T}$ 的约束. 为了消除隐因子在实际取值上的差异,本文将这三个子矩阵的元素值初始化为(0,1)之间的随机值. 在子矩阵求解时,首先按照算法 1 进行初始化,然后经过多轮迭代使得损失函数值之间的差值收敛于一定范围内. 这时的 $\boldsymbol{P}'$、$\boldsymbol{Q}'$ 和 $\boldsymbol{T}'$ 即为我们所求的子矩阵. 将 $\boldsymbol{P}'$ 和 $\boldsymbol{Q}'$ 返回,作为求解新 $\boldsymbol{C}'$ 的子矩阵. 在该模型中,属性的数量与内容可以根据不同的场景而调整. 对这些属性进行归一化后,再交由模型进行处理. 其中归一化操作是将各个属性中的数值等比例缩放到[0,1]之间. 然后,将网格属性矩阵 X 融入到矩阵分解中,在一定程度上补充原本稀疏矩阵的语义信息.

$$loss = \min_{\boldsymbol{P},\boldsymbol{Q}}\Big[\ \|\boldsymbol{C}-\boldsymbol{P}\boldsymbol{Q}\|^2 + \frac{l_2}{2}\|\boldsymbol{X}-\boldsymbol{Q}\boldsymbol{T}'\|^2 + \frac{l_1}{2}\left(\|\boldsymbol{P}\|^2+\|\boldsymbol{Q}\|^2+\|\boldsymbol{T}\|^2\right)\Big] \tag{5}$$

GeoLFM 详细训练过程如算法 1 所示. 其中 $(\,)^{\mathrm{T}}$ 代表着矩阵的转置,该算法主要是使用随机梯度下降对子矩阵 $\boldsymbol{P}$、$\boldsymbol{Q}$ 内容进行更新. 当两轮迭代的误差变化小于阈值 e 时,就停止更新,并返回分解结果. l_1 和 l_2 是用于防止结果过拟合的正则化因子,若取值过大,可能导致结果不能收敛,故在本文的对比实验中均取值为 0.01.

算法1　融合地理信息的隐语义模型GeoLFM

Input:driver-grid access matrix $\boldsymbol{C}$, grid-attribution matrix $\boldsymbol{X}$.
Output:submatrix $\boldsymbol{P},\boldsymbol{Q}$ after decomposition.
Initialize $\boldsymbol{P}\in\mathbf{R}^{M\times K},\boldsymbol{Q}\in\mathbf{R}^{N\times K},\boldsymbol{T}\in\mathbf{R}^{l\times K}$ with random values,$\lambda_1=\lambda_2=0.01$
Set a as step length in gradient descent.
While $L_t - L_{t+1} > a$
  For each $\mathrm{item_R}$ in $\boldsymbol{C}$
    $\boldsymbol{C}'_{ij} = \boldsymbol{P}_{i*}\boldsymbol{Q}_{j*}$
    $\boldsymbol{P}_{i*} = \boldsymbol{P}_{i*} - a((\boldsymbol{C}'-\boldsymbol{C})\boldsymbol{Q}_{j*} - \lambda_1\boldsymbol{P}_{i*})$
    $\boldsymbol{Q}_{j*} = \boldsymbol{Q}_{j*} - a((\boldsymbol{C}'-\boldsymbol{C})\boldsymbol{P}_{i*} + \lambda_2(\boldsymbol{Q}_{j*}\boldsymbol{T}_{k*}-\boldsymbol{X})\boldsymbol{T} + \lambda_1\boldsymbol{Q}_{j*})$
    $\boldsymbol{T} = \boldsymbol{T} - a(\lambda_2)\boldsymbol{Q}_{j*}\boldsymbol{T}-\boldsymbol{X})\boldsymbol{Q}_{j*} + \lambda_1\boldsymbol{T})$
  end
end
Return $\boldsymbol{P},\boldsymbol{Q}$

4.2　基于空间位置的载客区域推荐

使用 GeoLFM 对司机-网格访问矩阵 $\boldsymbol{C}$ 进行分解后,便可获得司机 i 对地点 j 的偏好,即矩阵 $\boldsymbol{P}$ 的第 i 行与矩阵 $\boldsymbol{Q}$ 的第 j 列的内积.

本文统计了数据集中所有出租车司机在寻客过程中所行驶的路程,即寻客距离. 然后通过分析发现司机的寻客距离与概率密度之间服从 $\mu=0,\sigma^2=0.77$ 的正态分布. 因此,本文采用正态分布来刻画出租车在网格间移动的便捷性,即据定义 6 获得网格可达度矩阵 $\boldsymbol{A}_{N\times N}$,该矩阵的行、列均是网格集合. 通过结合矩阵 $\boldsymbol{C}$ 与司机当前位置,为司机推荐符合其偏好的网格列表.

整个工作的流程如过程 1 所示. 其中使用边长为 l 的网格对研究区域进行网格划分,得到网格集合 GridSet. 首先,依据网格间的距离,构造矩阵 $\boldsymbol{A}$. 然后结合轨迹数据与网格集合,构造矩阵 $\boldsymbol{C}$. 接着再依据 POI、网格

过程1　应对稀疏数据的出租车载客区域推荐

Input:grid side length l;Trajectories $T_1,T_2,...,T_n$;
  the current position grid i of driver u;
  POIset $p_1,p_2,p_3,...,p_n$;
Output:Gridset
GridSet=GridDivide(l)
$\boldsymbol{A}$ = AccessMatrix(GridSet)
for T_i in Trajectories
  for G_j in GridSet
    $\boldsymbol{C}_{ij}$ = GridVisit(T_i,GridSet$_j$)
  end
end
$\boldsymbol{X}$ = GridAttribute(GridSet,POIset,Trajectories)
$\boldsymbol{P},\boldsymbol{Q}$ = GeoLFM($\boldsymbol{C},\boldsymbol{X}$)
$\boldsymbol{R}^*_{ui} = (\boldsymbol{P}\boldsymbol{Q})_{u*}\odot\boldsymbol{A}_i$
return Top($\boldsymbol{R}^*_{ui}$)

与轨迹数据构建网格-属性矩阵 $\boldsymbol{X}$. 最后依据融合地理信息的隐语义模型对矩阵 $\boldsymbol{C}$ 进行缺失值填充,并根据矩阵 $\boldsymbol{A}$ 与填充后的矩阵,为司机推荐载客区域.

5 实验与评估

5.1 数据集说明与处理

5.1.1 轨迹数据描述与处理

本文实验所采用的数据集是成都市主城区 2014 年 8 月份共 13000 辆出租车在一个月内的轨迹数据. 在对数据集进行了包括轨迹平滑[5]、异常状态修正和去除重复轨迹点等预处理后,最终选择了成都市中心区域([103.96,30.58]~[104.18,30.73])作为研究范围. 我们使用边长为 100 米的正方形网格划分该研究区域,共得到了 47427 个网格. 实验中,我们先预提取轨迹中的所有载客点作为实验数据集,然后根据数据集的时间属性,将其分为工作日与周末两种情况. 同时将一天的数据划分得到三个时段的数据子集:7 时前的数据作为训练集,7~9 时的作为测试集;12 时前的数据作为训练集,12~14 时的数据作为测试集;21 时以前的数据作为训练集,21~23 时的数据作为测试集. 以天为单位依次验证本文提出的方法在早、中、晚各时段的表现.

5.1.2 POI 数据描述与处理

为构建矩阵 $\boldsymbol{X}$,本文从高德地图上共爬取到 343851 个研究范围内的 POI 数据点. 每个 POI 数据点包含经纬度信息和相应的类型信息. 该数据集包含 12 种 POI 类型,如餐饮、风景名胜、公共设施、交通设施、科教文化、金融保险、汽车服务、商务住宅、生活服务、体育休闲、医疗保险和住宿服务. 其中不同类型的 POI 数量排在前三的是餐饮、交通设施和商务住宅,恰好分别代表着人们的食、行和住.

5.2 实验性能与对比分析

为了合理选取网格属性,我们随机抽样 30 位出租车司机与 250 个网格区域,从数据集中获取相应的日志信息后,构造网格属性信息和相关的司机偏好信息. 本文通过分析 POI 数据和出租车轨迹数据,得到 18 种网格属性.

本文的评价指标参考推荐系统[17]中常见的平均绝对误差(MAE)与均方根误差(RMSE). 这两个指标描述了推荐载客结果与司机实际载客情况的差异大小,误差越小,准确度就越高. 二者定义如式(6)所示.

$$MAE = \frac{\sum_{u,i} |r_{ui} - \hat{r}_{ui}|}{N} \qquad RMSE = \sqrt{\frac{\sum_{u,i} (r_{ui} - \hat{r}_{ui})^2}{N}} \tag{6}$$

这里 r_{ui} 表示司机 P_u 对网格 Q_i 的真实访问情况,如果访问了则 $r_{ui}=1$,反之,$r_{ui}=0$. $\hat{r}_{ui}$ 表示相应的司机对网格访问的概率大小. N 表示测试集样本个数. 从上述定义可知,越低的 MAE 与 RMSE 的值意味着越好的推荐效果.

5.2.1 对比方法

为了更好地测验 GeoLFM 的有效性,本实验采用常见的推荐算法包括协同过滤和相关的矩阵分解方法进行比较.

(1)基于时空关系推荐(TLR). 即使用网格划分目标区域[1],并使用高斯过程回归模型与相应的统计方法分别获得乘客在周末与工作日的需求量、起终点的平均行车路程与时间,以为不同司机推荐乘客数量多、距离近、耗时短的网格区域.

(2)吸引力模型(AM)[18]. 即先通过聚类方式将研究区域内的载客点集规约成较少的类簇;然后参照传统的引力模型 Huff Model,分析司机的选择行为. 这样目标区域的载客点数量与行车开销的乘积共同组成了载客区域选择时的依据.

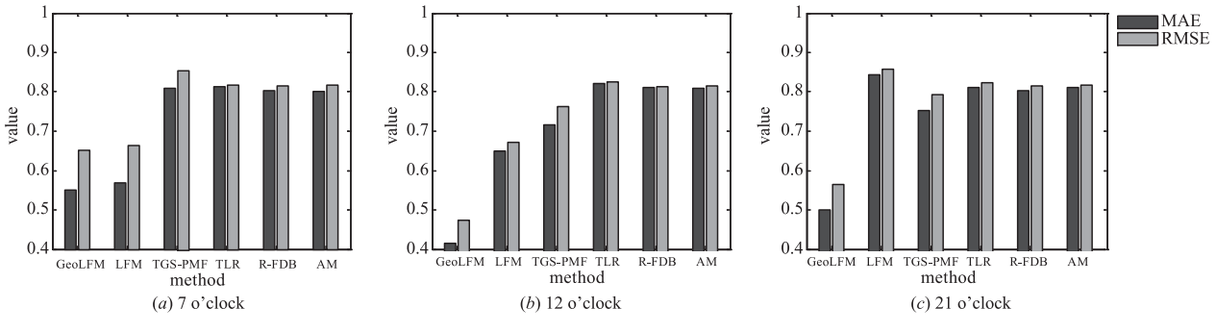
(3)带参数 R 的密度聚类(R-FDB). 即在 Fast-DBSCAN 的基础上,加入聚类簇的半径限制参数 Radius[19],以有效防止类内数量过大和信息丢失严重的问题,并发现热点载客区域.

(4)概率矩阵分解(PMF). 是在原始矩阵 $\boldsymbol{C}$ 与分解后的矩阵 $\boldsymbol{P}$、$\boldsymbol{Q}$ 乘积间的误差服从正态分布的假设下,进行矩阵分解. Ren 等人[9]在该模型的基础上,结合用户兴趣偏好和地理相关性因素,提出了 TGS-PMF,在推荐兴趣区域时,可以提供更多客观因素的参考.

(5)隐语义模型(LFM). 旨在参照地显示出融合地理信息隐语义模型的效果提升程度. 使用 LFM 直接分解司机-网格访问矩阵 $\boldsymbol{C}$,生成两个子矩阵后,将二者相应项的内积作为司机对网格的访问概率.

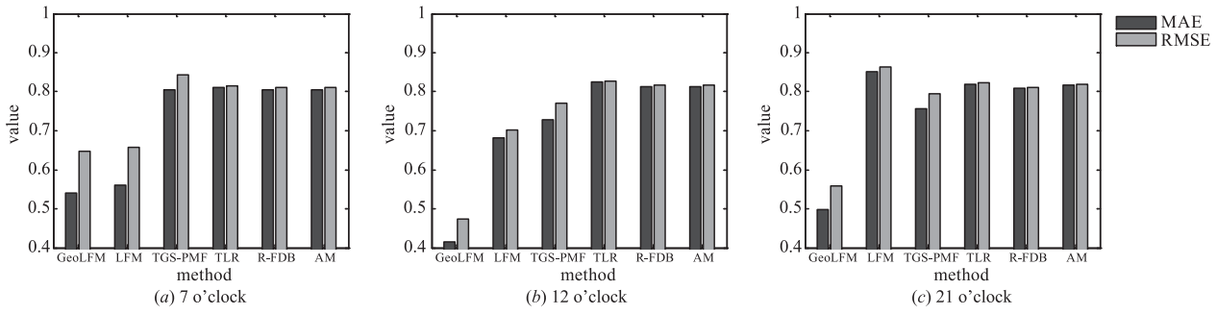
5.2.2 实验效果评价及参数影响

考虑到不同时段出租车载客情况的差异,图 3、图 4 分别展示在隐因子数 $K=20$ 的情况下,工作日与周末的 7~9 时、12~14 时与 21~23 时,6 种不同方法在测试集上,MAE 和 RMSE 的取值情况. 从图中可以看到,GeoLFM 无论是均方根误差还是平均绝对误差都是最小的. 周末与工作日的情况大体相同,以工作日为例,我们可以看到,与 LFM、TLR、TGS-PMF、R-FDB 和 AM 相比,GeoLFM 在 12~14 时段内推荐结果的 MAE 分别降低了 23.4%、49.8%、48.7%、49.1% 和 49.4%. 同时该时段下的 RMSE 分别下降了 19.8%、42.6%、41.3%、41.9% 和 42.0%. 同样,在 21~23 时段下的推荐结果也能体现我们方法的优越性. 7~9 时段的误差较大的原因是早上 7 点前的出租车载客数据太过稀少,此时如果仅依靠地理信息,则相当于加大了客观因



(*a*) 7 o'clock　(*b*) 12 o'clock　(*c*) 21 o'clock

图3　工作日分布情况



(*a*) 7 o'clock　(*b*) 12 o'clock　(*c*) 21 o'clock

图4　周末分布情况

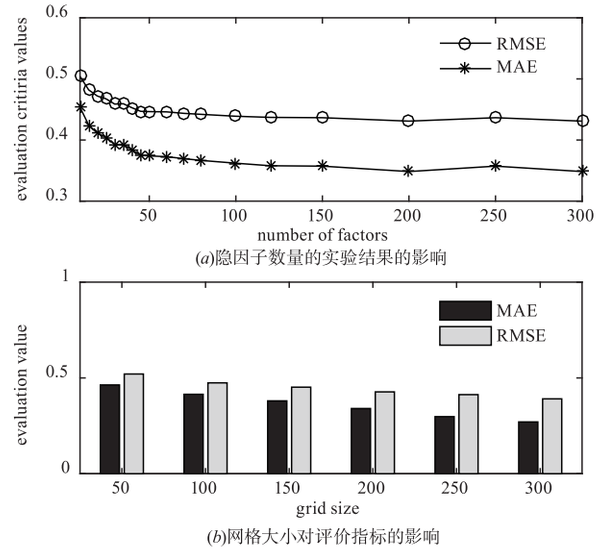
素的比重. 事实上,出租车载客推荐应该是结合司机偏好与客观地理信息的过程,如果只考虑某一方面的因素,可能会加大推荐的误差. 从全局来看,TLR、R-FDB与 AM 在应对稀疏数据时表现较差. 而 GeoLFM 则在一定程度上缓解了出租车推荐系统中的数据稀疏性问题,这充分说明融合地理信息的可行性与必要性.

在本文提出的 GeoLFM 模型中,潜在因素的数量与网格划分大小都对推荐的结果有一定的影响. 为了测试其实际影响程度,本文分别设置了潜在因素的数量 K 从 0 到 300 递增,网格大小从 50 到 300 米变化,如图 5 所示. 这两个因素的取值范围均是根据实际经验设定,若隐因子数量设置过大,或者网格大小设置过小,会产生较大的计算量,但系统的准确率提升并不明显. 若隐因子数量设置过小,或网格大小设置过大,虽然计算量变小,但模型的表达能力与推荐的精度均不高.

由图 5 (*a*)看出,随着隐因子数量的增加,MAE 与 RMSE 值的分布逐渐收敛在 0.35 与 0.43 附近. 从图 5 (*b*) 中可以看出,随着网格变大,系统的准确率也变得越高,且越来越趋于平缓. 这是因为在一定的范围内,网格越大,网格内包含的载客点更多,原来分布在网格间边缘区域载客点的数量则相对变少,所以最后的推荐结果就会越好.

5.3　性能分析

本实验的物理环境为:Intel(R) Core(TM) i5-6500 CPU @3.2GHz 处理器,8GB 内存的单机. 软件环境为



(*a*)隐因子数量的实验结果的影响

(*b*)网格大小对评价指标的影响

图5　敏感参数K与δ对评价指标的影响

基于 Windows7 64 位操作系统的 Python 2.7. 实验中时间与空间的代价如表 1 所示. 其中展示了 GeoLFM 模型在 1~5 天时间跨度下的耗时与相应的结果. 从表 1 可见,随着天数的增加,无论是计算还是存储的消耗,都明显增加. 由于需要存储历史数据,所以存储资源的消耗与所需存储历史数据的时长呈正相关增长,但是推荐结果的误差并没有明显的降低. 这说明使用规模小的数据进行推荐具有可行性,而且还能减少云计算的存储空间,降低计算和通信开销,提高推荐效率和响应速度.

表 1 不同时间长度计算代价

时间/d	1	2	3	4	5
耗时/s	21	49	91	124	153
空间/m	34.91	72.75	110.62	147.71	187.17
MAE	0.428	0.441	0.425	0.423	0.422

6 总结

本文使用短时间内的出租车轨迹数据进行出租车载客区域推荐,减少了云平台中的计算与存储开销.针对短时间数据情况下存在的数据稀疏性问题,本文使用融合地理信息的隐语义模型,并使用高斯分布描述空间地理位置信息与可达度的关系,为司机提供了精准的出租车载客区域推荐.实验使用真实数据集对本文方法进行验证后,发现在平均绝对误差(MAE)与均方根误差(RMSE)两个常用评价指标上均有最优效果.相比于协同过滤等推荐算法,本文的方法在一定程度上缓解了数据稀疏性的问题.在与 LFM、PMF 等常见方法的比较中,验证了融合地理信息确实能在一定程度提升推荐效果,减少推荐误差.

下一步工作中,我们将考虑通过社交网络挖掘特殊日期和交通事件对出租车司机的偏好以及推荐结果的影响,以进一步提高载客区域推荐的准确率.

参考文献

[1] Kong X, Xia F, Wang J, et al. Time-location-relationship combined service recommendation based on taxi trajectory data[J]. IEEE Transactions on Industrial Informatics, 2017,13(3):1202 – 1212.

[2] 廖律超,蒋新华,邹复民,等.一种支持轨迹大数据潜在语义相关性挖掘的谱聚类方法[J].电子学报,2015,43(5):126 – 134.
LIAO Lü-chao, JIANG Xin-hua, ZOU Fu-min, et al. A spectral clustering method for big trajectory data mining with latent semantic correlation[J]. Acta Electronica Sinica,2015,43(5):126 – 134. (in Chinese)

[3] Li X, Gao C, et al. Rank-GeoFM: A ranking based geographical factorization method for point of interest recommendation[A]. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York: ACM, 2015. 433 – 442.

[4] 宋飞,李蓉,张思东,等.基于趋势预测的合乘收益研究[J].电子学报,2014,42(7):1353 – 1359.
SONG Fei, LI Rong, ZHANG Si-dong, et al. The research on taxi sharing benefit based on tendency estimation[J]. Acta Electronica Sinica, 2014, 42(7): 1353 – 1359. (in Chinese)

[5] Liu Y, Liu J, Wang J, et al. Recommending a personalized sequence of pick-up points[J]. Advances in Services Computing. Springer International Publishing, 2016, 10065: 278 – 291.

[6] 王晓文.基于载客热点区域的出租车巡游路径推荐方法的研究与实现[D].山东,青岛:中国海洋大学,2015.
Wang Xiaowen. Research and Implementation on Taxi Cruise Path Recommendation Method Based on Pick-up Hotspots Areas[D]. Shangdong, Qingdao: Ocean University of China, 2015. (in Chinese).

[7] Yang Q, Gao Z, Kong X, et al. Taxi operation optimization based on big traffic data[A]. IEEE International Conference on Ubiquitous Intelligence and Computing[C]. Beijing: IEEE, 2015. 127 – 134.

[8] Yuan J, Zheng Y, Zhang L, et al. Where to find my next passenger[A]. Proceedings of the 13th International Conference on Ubiquitous computing[C]. Beijing: ACM, 2011. 109 – 118.

[9] 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[J].计算机学报,2017,40(4):824 – 841.
Ren Xing-Yi. Song Mei-Na. Song Jun-De. Context-aware point-of-interest recommendation in location-based social networks[J]. Chinese Journal of Computers, 2017, 40(4): 824 – 841. (in Chinese)

[10] Liu L, Wu C, Zhang H, et al. Research on taxi drivers′ passenger hotspot selecting patterns based on GPS data: A case study in Wuhan[A]. International Conference on Transportation Information and Safety[C]. Banff, Canada: IEEE, 2017. 432 – 441.

[11] Jiang W, Lian J, Shen M, et al. A multi-period analysis of taxi drivers′ behaviors based on GPS trajectories[A]. International Conference on Intelligent Transportation Systems[C]. New York: IEEE, 2018. 1 – 6.

[12] Zhang D, Sun L, Li B, et al. Understanding taxi service strategies from taxi GPStraces[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1): 123 – 135.

[13] Wang Y, Zheng Y, Xue Y. Travel time estimation of a path using sparse trajectories[A]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM, 2014. 25 – 34.

[14] 向隆刚,邵晓天.载体轨迹停留信息提取的核密度法及其可视化.测绘学报,2016,45(9):1122 – 1131.
Xiang Long-Gang, Shao Xiao-Tian. Visualization and extraction of trajectory stops based on kernel-density[J]. Acta Geodaetica et Cartographica Sinica, 2016, 45(9): 1122 – 1131. (in Chinese)

[15] Lian D, Zhao C, Xie X, et al. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation[A]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New

York:ACM,2014. 831 – 840.
[16] Zhang W,Wang J,Feng W. Combining latent factor model with location features for event-based group recommendation[A]. The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. USA:ACM,2013. 910 – 918.
[17] 李玉省. 个性化推荐系统关键技术研究[D]. 北京:北京邮电大学,2016.
Li Yu-Sheng. Research on Some Key Technologies of Personalization Recommendation System. [D]. Beijing: Beijing University of Posts and Telecommunications, 2016. (in Chinese)
[18] Tang J,Jiang H,Li Z,et al. A two-layer model for taxi customer searching behaviors using GPS trajectory data [J]. IEEE Transactions on Intelligent Transportation Systems,2016,17(11):3318 – 3324.
[19] Qi H,Liu P. Mining taxi pick-up hotspots based on spatial clustering[A]. 2018 IEEE Smart World,Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications,Cloud & Big Data Computing,Internet of People and Smart City Innovation [C]. Guangzhou:IEEE,2018. 1711 – 1717.

作者简介

廖祝华　男,1977 年 9 月生,湖南株洲人. 副教授,分别于 2004 年和 2012 年在中科院研究生院和中科院计算所获硕士和博士学位,主要研究方向为数据挖掘、网络数据获取和计算机网络.

张　健　男,1994 年 8 月生,湖南岳阳人. 2019 年于湖南科技大学获硕士学位,主要研究方向为轨迹数据挖掘和大数据分析技术.

刘毅志(通信作者)　男,1973 年 9 月生,湖南衡阳人. 副教授,分别于 2003 年和 2011 年在湘潭大学和中科院计算所获硕士和博士学位,主要研究方向为多媒体内容分析、智慧城市和智慧医疗.