DOI: 10.11992/tis.201801048
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180625.0912.002.html

多特征融合的兴趣点推荐算法

涂飞
（重庆理工大学 计算机科学与技术学院，重庆 400054）

摘　要：基于位置社交网络的兴趣点推荐越来越受到工业界和学术界的关注。由于用户签到数据集的稀疏性以及签到地理位置的聚集性，使得目前的推荐算法效率普遍不高，特别是当用户外出到新的地点时，推荐效果更是急剧下降。因此本文提出了一种基于用户−区域−内容主题的多特征联合推荐算法 (UCRTM)，以隐主题模型为基础，在统一的框架下利用隐含因子关联性融合了用户的偏好、兴趣点的内容以及兴趣点所属地理区域主题等信息来进行推荐，使得用户无论身处何地，都能获得理想的推荐服务。本文在两种真实的数据集上进行了实验，结果表明该方法不仅能够克服数据的稀疏性以及弱语义性等问题，而且与其他方法相比具有更高的推荐准确率。

关键词：位置社交网络；兴趣点推荐；主题模型；困惑度；稀疏性；聚集性；协同过滤；特征融合
中图分类号：TP391.9　　**文献标志码**：A　　**文章编号**：1673−4785(2019)04−0779−08

中文引用格式：涂飞. 多特征融合的兴趣点推荐算法 [J]. 智能系统学报, 2019, 14(4): 779–786.
英文引用格式: TU Fei. A point of interest recommendation algorithm based on multi-feature fusion[J]. CAAI transactions on intelligent systems, 2019, 14(4): 779–786.

A point of interest recommendation algorithm based on multi-feature fusion

TU Fei
(School of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: The point of interest recommendation service is receiving increasing attention from the industry and academia. The sparsity of users' activity history datasets and aggregation of geological position prevent the current recommendation algorithm efficiency from being high, and especially, when a user goes out to a new city, the recommendation effect will fall sharply. Therefore, this paper presents a user-content-region topic model based on a joint recommendation algorithm, considering to the user's preferences, the content of the point of interest, and the geographical area, making users obtain an ideal recommendation service irrespective of their location. An experiment was carried out on two real datasets, and the results show that this method can not only overcome problems such as data sparseness, weak semantic performance, but also has a higher recommendation accuracy compared with other methods.

Keywords: location-based social networks; point of interest recommendation; topic model; perplexity; sparseness; aggregation; collaborative filtering; multi-feature fusion

基于位置的社交网络 LBSN[1](location based social network) 实现了用户对其访问地理位置的签到功能，并能够发布相应的评论、图片、视频信息与好友分享。通过 LBSN 可以提供好友推荐[2-3]、兴趣点推荐[4-5] 等多种个性化服务。兴趣点推荐的目标是向特定用户推荐满足其需求的、具有一定长度的未知兴趣点列表，来增强用户体验。一般来说，推荐的兴趣点包括地点 (如饭馆、商场、公园和影院等) 和活动 (如演唱会、公益活动等) 两类。与传统电子商务网站的推荐系统不同，兴趣点推荐有如下的特性：

收稿日期：2018−01−27.　**网络出版日期**：2018−06−26.
基金项目：国家自然科学基金项目 (61272277).
通信作者：涂飞. E-mail: tufeicq1979@163.com.

1) 数据的弱语义性：传统的推荐系统中，用户对商品的评分是显现的，可以表达其偏好程度，如五级评分制中 5 代表很喜欢，而 1 代表不喜欢。在兴趣点推荐系统中，只能获取用户对兴趣点的访问次数，而签到次数的多少并不能反映用户的偏好程度。换句话说传统的推荐数据中同时包含正例和负例，而兴趣点推荐数据仅包含正例，这就使得很多成熟的推荐方法如协同过滤算法等并不直接适用于兴趣点推荐。

2) 数据的稀疏性：用户对兴趣点的签到矩阵相比于传统的用户−商品评分矩阵更加稀疏，如国外著名的 LBSN 网络——Gowalla 的数据稀疏度为 2.08×10^{-4}。传统的推荐算法难以直接适用于兴趣点推荐。此外用户的历史活动记录具有聚集性，通常只集中在居住地或工作地等少数几个区域。当用户外出时，由于缺少该区域的历史签到记录无法做出准确的推荐。

3) 社交影响不大：传统在线社交网络上的“朋友”往往具有相似的兴趣爱好，因此很多推荐算法通过引入社交关系来处理数据的稀疏性问题，提高效率。但是在 LBSN 中的“朋友”不一定具有相同的兴趣爱好，引入社交关系对兴趣点推荐影响不大。

鉴于此，本文提出了一种新的推荐模型——用户−内容−区域主题模型 (user-content-region topic model, UCRTM)。该模型同时分析了用户兴趣、地点特定主题以及所属地理区域主题等多个特征，以隐含主题为基础，用统一的框架将各种特征进行融合，在一定程度上克服了用户签到数据的稀疏性和弱语义性等问题，实验证明能获得较好的用户体验。

1　相关工作

目前基于位置社交网络的兴趣点推荐算法可归纳为 3 类：

1) 传统推荐算法的直接运用：这类方法认为用户对兴趣点的签到次数能够代表其偏好程度，构造用户−兴趣点签到矩阵并利用传统的推荐算法进行推荐。如：文献 [6] 提出的基于用户和兴趣点的混合协同过滤算法；文献 [7] 提出的基于正则化矩阵分解 (RMF) 算法和文献 [8] 提出的概率矩阵分解 (PMF) 算法等。这类方法的本质是尽量完善推荐模型，但由于数据集本身过度稀疏，以及数据的弱语义性导致推荐质量并不高。

2) 引入地理因素的推荐算法：在 LBSN 中，用户与兴趣点的地理距离也是推荐的重要因素，这也是有别于商品推荐的重要特征，这类算法将地理信息融入到模型中，如文献 [6] 认为用户签到的兴趣点在地理位置上是符合幂律分布，文献 [8] 则认为用户的活动区域是围绕多个中心点展开的，进而引入了多中心高斯分布模型。事实证明对地理信息建模有助于推荐效果的提升。

3) 引入社交影响的推荐算法：这类方法认为社交网络上的“朋友”拥有相同的兴趣爱好，结合朋友的签到访问历史记录进行推荐，如文献 [9] 利用相似用户进行推荐时，直接利用好友进行推荐，而忽略 LBSN 中其他用户。文献 [8] 将社交关系直接融入到矩阵分解 (PMF) 算法中，但是实验证明社交关系对推荐准确率的影响不大。

还有一些方法同时考虑了地理因素、用户的偏好以及社交关系。如文献 [10] 设计了一种 UPS (user, proximity and social-based) 算法，将社交影响因子融合到基于用户偏好的协同过滤算法中，以此来提高用户相似度计算的准确性。实验证明该算法在稀疏数据环境下的推荐效果并不是很好，而且该算法没有考虑用户在不同的地理位置的影响。文献 [11] 提出了 USG (user, social and geographical influence based recommendation) 推荐算法，综合考虑了用户偏好、社交影响和地理影响，采用线性融合技术集成这 3 种因素,以此来提高算法的准确率。该方法虽然考虑了地理因素，但是只考虑了用户常驻地特征，推荐的地点都是常驻地附近区域，而且算法的参数不能自适应地调节。此外还有很多学者利用概率产生式模型对位置社交网络的推荐系统进行研究，将影响用户签到决策的各种因素进行综合考虑和集成，比如文献 [12] 提出的 LCARS 系统从用户兴趣、地理位置、兴趣点当地特色 3 个方面分析，来对用户的签到行为进行建模，文献 [13] 提出的 JUMAI 系统更是从用户兴趣、兴趣点所在区域与用户所在区域的距离、签到时间，以及兴趣点类别 4 个角度来指导签到决策。文献 [14] 在此基础上还考虑了用户在新的地点会产生兴趣漂移情况。但是这些模型均没有考虑兴趣点本身的内容，其次上述模型在对各因素进行建模时，没有体现自适应的特性，即针对不同的兴趣点，何种因素对决策起支配作用。本文提出了用户−区域−内容主题模型，真实地模拟了用户对兴趣点签到的决策过程，实验证明在稀疏的数据集下有较理想的推荐效果。

2　用户−区域−内容主题模型

2.1　模型介绍

用户是否会对特定的兴趣点签到, 会受到以下 3 种因素的影响:

1) 用户自身偏好的影响: 一般来说, 只有兴趣点满足用户的喜好, 用户才会欣然前往并产生签到行为。比如球迷可能去看 CBA 联赛, 而音乐爱好者可能去听演唱会。

2) 兴趣点自身内容的影响: LBSN 中基本包含了对兴趣点的介绍信息, 图 1 是豆瓣活动网站的页面信息, 该页面显示了活动的时间、地点以及主题。当用户浏览该页面时, 可能被活动的主题信息中某个特征所吸引, 才促使了用户的签到行为。

图 1　兴趣点简介

Fig. 1　A brief introduction of interest points

3) 兴趣点所属区域的影响。用户根据自身爱好或是事先知晓兴趣点的内容而产生的签到行为可认为是有目的, 有主观倾向性的。但并不是所有的访问签到行为都是如此。用户的某次签到行为可能开始是漫无目的的, 只是随机选择某一地理区域的某一兴趣点。但是此处的随机也受以下两点约束: 1) 兴趣点所属区域离用户的距离。当该区域离用户较近时, 被用户访问的概率较大, 否则访问概率较小。2) 区域的主题。当用户外出到新的区域时, 对该区域一无所知, 也无法从其“相似用户”获得信息, 在做决策是否访问某一兴趣点时, 往往会受到该区域主题的影响。比如该区域的风俗习惯、当地人的兴趣喜好, 或是当地比较著名的人文景点等。

用户对兴趣点签到, 必定是受到以上 3 种因素其中之一的影响。因此本文提出了一种基于用户−区域−内容的联合推荐模型, 利用隐主题因子表示上述 3 种因素, 将用户对 3 种因素的选择过程进行建模。

2.2　模型的形式化

图 2 为用户−区域−内容联合推荐模型对应的概率图。该图右边的部分是一个简单的 LDA 模型, 构造了兴趣点描述文档的生成过程。当用户 u 对兴趣点 l_u 签到时 l_u 的介绍文档 d 已经存在,文档和单词的主题分布可分别独立计算。当用户对兴趣点 l_u 签到时, 首先要确定 l_u 的主题 z, z 有 3 种来源, 分别为兴趣点介绍文档 d 中出现过的主题、用户的兴趣以及兴趣点所属地理区域的主题。采用选择变量 X 来控制兴趣点的主题 z 的来源, X 满足多项式分布, 其值分别为 user、region 和 content。

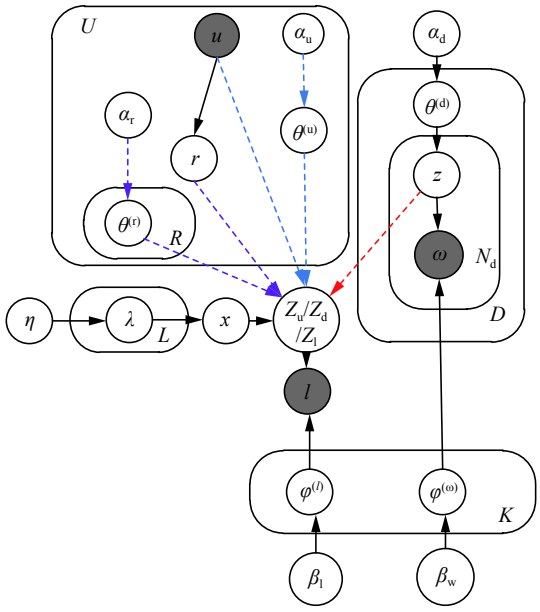

图 2　用户-区域-内容联合推荐模型

Fig. 2　User-content-region based joint recommendation model

假设模型中用户集合为 U, 兴趣点集合为 L, 介绍文档集合为 D, 单词集合为 W,区域集合为 R, 以及主题集合为 K, 具体生成过程如下:

1) 对于任意文档 $d(d \in D)$, 根据 Dirichlet(α_d) 分布得到文档 d 在主题上的多项式分布 $\theta^{(d)}{}_K$。

2) 对于任意主题 $k(k \in K)$, 根据 Dirichlet(β_w) 分布可得到词 w 在主题 k 上的概率分布 $\varphi^{(w)}{}_k$; 根据 Dirichlet(β_l) 分布可得到兴趣点 l 在主题 k 上的分布 $\varphi^{(l)}{}_k$。

3) 对于任意用户 $u(u \in U)$,根据 Dirichlet(α_u) 得到用户 u 的主题分布 $\theta^{(u)}{}_K$,根据 Dirichlet(α_r) 得到区域 r 的主题分布 $\theta^{(r)}{}_K$,根据 Dirichlet(α_{ur}) 得到用户 u 在区域 r 上分布 $\theta^{(ur)}{}_r$。

4) 长度为 N_d 的文档 d 中, 词 w_i 的生成过程为:

①根据文档 d 的主题分布 $\theta^{(d)}{}_K$ 抽样获得主题 z_i;

②利用单词在 z_i 上的概率分布 $\varphi^{(w)}{}_{z_i}$ 抽样产生单词 w_i。

5) 用户 u 对兴趣点 l 的签到过程如下：

① 根据 Dirichlet(η) 抽样得到 λ_1，根据 Multinomial(λ_1) 分布得到控制值 x；

②如果 $x=$ document，说明兴趣点的主题由文档 d 生成。此前已抽样得到文档 d 中所有单词的主题集合 $\{z_{w_1},z_{w_2},\cdots,z_{w_{Nd}}\}$，根据 Uniform($z_{w_1},z_{w_2},\cdots,z_{w_{Nd}}$) 分布抽样生成兴趣点的主题 z_l；

③如果 $x=$ user，说明兴趣点的主题来自于的用户的偏好：根据兴趣主题分布 $\theta^{(u)}{}_K$ 抽样获得兴趣点的主题 z_l；

④如果 $x=$ region，兴趣点的主题由兴趣点所属区域的主题生成，首先利用用户 u 在区域 r 上分布 $\theta^{(ur)}{}_r$ 得到区域 r，根据区域 r 在主题 k 上的概率分布 $\theta^{(r)}{}_k$ 获得主题 z_l。

6) 最后利用兴趣点在主题 z_l 上的概率分布 $\varphi^{(l)}{}_{z_l}$ 得到 l。

2.3　确定参数值

模型中变量的联合概率分布为

$$\begin{aligned}&p(u,r,l,z,x,w)=\\&p(w)p(l|z,x,r,u)p(z,r,u|x)p(x)=\\&p(w)(p(l|z,x)p(z|x,\theta^{u})p(x=\text{user})+\\&p(l|z,x)p(z|x,\theta^{d})p(x=\text{document})+\\&p(l|z,x)p(r|u)p(z|x,\theta^{r})p(x=\text{region}))\end{aligned}\tag{1}$$

由式 (1) 可知该模型需要估计以下 6 个参数：1) 文档的主题分布 $\theta^{(d)}$；2) 主题−词分布 $\varphi^{(w)}$(参数 1)、2) 为基本 LDA 模型对应的参数，用于求 $p(w)$)；3) 兴趣点的主题分布 $\varphi^{(l)}$(即 $p(l|z,x)$)；4) 用户兴趣分布 $\theta^{(u)}$；5) 用户活动区域分布 $\theta^{(ur)}$；6) 选择概率 λ 的多项式分布 (即 $p(x)$)。

文中采用 Gibbs 抽样方法，过程如下，具体的参数说明见表 1。

表 1　模型参数说明

Table 1　Model parameter description

参数	含义
d、u、w、l、x、z	变量的实例：d为文档，u为用户，w为词，l为兴趣点，x为控制开关，z为主题，r为区域
K、U、D、W、R、L	主题的个数、用户的个数、文档的个数、词的个数、区域的个数、兴趣点的个数
N_d	d中的单词总数
$C_{kd,-i}^{KD}$	主题k在文档d出现的次数
$C_{vk,-i}^{WK}$	词v属于主题k的次数
$C_{lk,-j}^{LK}$	兴趣点l属于主题k的次数
$C_{ku,-i}^{KU}$	用户u的兴趣为主题k的次数
$n_{l_j,\text{user},-j}$	兴趣点 l_j 的主题来源于用户兴趣的次数
$n_{l_j,\text{document},-j}$	兴趣点 l_j 的主题来源于文档主题的次数
$n_{l_j,\text{region},-j}$	兴趣点 l_j 的主题来源于区域主题的次数
$\theta^{(d)}$	文档主题的分布
$\varphi^{(w)}$	主题词的分布
$\theta^{(u)}$	用户兴趣的分布
$\theta^{(ur)}$	用户活动区域的分布
$\varphi^{(l)}$	兴趣点的主题分布
$\theta^{(r)}$	区域的主题分布
λ_1	兴趣点的主题选择概率分布
α_u、α_d、α_{ur}、α_{r}、β_w、β_{l}、η	Dirichlet分布的超参数。

1) 利用式 (2) 计算单词在主题上的后验概率，进而对单词的主题进行抽样，

$$P(z_i=k|w_i,z_{-i},w_{-i},\alpha_d,\beta_w)\propto\frac{C_{kd,-i}^{KD}+\alpha_d}{\sum_{k'}C_{kd,-i}^{KD}+K\alpha_d}\cdot\frac{C_{wk,-i}^{WK}+\beta_w}{\sum_{w'}C_{wk,-i}^{WK}+V\beta_w}\tag{2}$$

2) 计算兴趣点主题的后验概率，分 3 种情况：

①当选择变量 x=user 时，抽样方程为

$$P(z_{\text{j}}=k,x=\text{user}|l_j,z_{-j},\alpha_u,\beta_l,\eta)\propto\frac{C_{ku,-j}^{KU}+\alpha_u}{\sum_{k'}C_{k'u,-j}^{KU}+K\alpha_u}\cdot\frac{C_{lk,-j}^{LK}+\beta_l}{\sum_{l'}C_{l'k,-j}^{LK}+L\beta_l}\cdot\frac{\eta_{\text{user}}+n_{l_j,\text{user},-j}}{L+\sum_{x\in\{\text{user,region,document}\}}\eta_x}\tag{3}$$

②当选择变量 x=document 时，抽样方程为

$$P(z_j=k,x=\text{document}|l_j,z_{-j},\beta_l,\eta)\propto \frac{C_{kd}^{KD}}{N_d}\cdot\frac{C_{lk}^{LK}+\beta_l}{\sum_{l'}C_{l'k}^{LK}+L\beta_l}\cdot\frac{\eta_{\text{document}}+n_{l_j,\text{document},-j}}{L+\sum_{x\in\{\text{user,region,document}\}}\eta_x} \tag{4}$$

③当选择变量 x=region 时, 抽样方程为

$$P(z_j=k,r_j=r,x=\text{region}|l_j,z_{-j},\alpha_u,\alpha_{ur},\beta_l,\eta)\propto \frac{C_{ru,-j}^{RU}+\alpha_{ur}}{\sum_{r'}C_{r'u,-j}^{RU}+R\alpha_{ur}}\cdot\frac{C_{kr,-j}^{KR}+\alpha_r}{\sum_{k'}C_{k'r,-j}^{KR}+K\alpha_r}\cdot \frac{C_{lk,-j}^{LK}+\beta_l}{\sum_{l'}C_{l'k,-j}^{LK}+L\beta_l}\cdot\frac{\eta_{\text{region}}+n_{l_j,\text{region},-j}}{L+\sum_{x\in\{\text{user,region,document}\}}\eta_x} \tag{5}$$

当式 (2)~(5) 迭代一定次数后状态稳, 可用式 (6)~(14) 近似计算模型的参数值。

$$\theta^{(d)}=\frac{C_{kd}^{KD}+\alpha_d}{\sum_{k'}C_{k'd}^{KD}+K\alpha_d} \tag{6}$$

$$\varphi^{(w)}=\frac{C_{wk}^{WK}+\beta_w}{\sum_{w'}C_{w'k}^{WK}+W\beta_w} \tag{7}$$

$$\theta^{(u)}=\frac{C_{ku}^{KU}+\alpha_u}{\sum_{k'}C_{k'u}^{KU}+K\alpha_u} \tag{8}$$

$$\varphi^{(l)}=\frac{C_{lk}^{LK}+\beta_l}{\sum_{l'}C_{l'k}^{LK}+L\beta_l} \tag{9}$$

$$\theta^{(ur)}=\frac{C_{ru}^{RU}+\alpha_{ur}}{\sum_{r'}C_{r'u}^{RU}+R\alpha_{ur}} \tag{10}$$

$$\theta^{(r)}=\frac{C_{kr}^{KR}+\alpha_r}{\sum_{k'}C_{k'r}^{KR}+K\alpha_r} \tag{11}$$

$$\lambda_{l,\text{user}}=\frac{\eta_{\text{user}}+n_{l,\text{user}}}{L+\sum_{x\in\{\text{user,region,document}\}}\eta_x} \tag{12}$$

$$\lambda_{l,\text{document}}=\frac{\eta_{\text{document}}+n_{l,\text{document}}}{L+\sum_{x\in\{\text{user,region,document}\}}\eta_x} \tag{13}$$

$$\lambda_{l,\text{region}}=\frac{\eta_{\text{region}}+n_{l,\text{region}}}{L+\sum_{x\in\{\text{user,region,document}\}}\eta_x} \tag{14}$$

2.4　模型的推荐

估计模型参数后便可用于在线推荐阶段。对于特定的用户 u、兴趣点 l 以及其对应的介绍文档 d, 根据式 (15) 可得 u 对 l 签到概率:

$$p(l|d,u)=p(x=d)\sum_{k=1}^{K}p(l|z_k)p_{\text{test}}(z_k|d)+ p(x=u)\sum_{k=1}^{K}p(l|z_k)p(z_k|u)+p(x=r)\sum_{r=1}^{R}\sum_{k=1}^{K}p(r|u)p(z_k|r) \tag{15}$$

式中 $p(l|z_k)$、$p(x)$、$p(z_k|u)$、$p(r|u)$、$p(z_k|r)$、$p_{\text{test}}(z_k|d)$ 可以通过模型训练获得。如果是新的兴趣点 l, 可以先将其对应文档 d 的主题分布 $p_{\text{test}}(z_k|d)$ 在线计算, 即将文档 d 加入到测试集 D 中, 重新用 Gibbs 抽样的方法获得。如果是新的用户, 可以直接根据式 (16) 来计算:

$$p(l|d,u)=\sum_{k=1}^{K}p(l|z_k)p_{test}(z_k|d) \tag{16}$$

这也在一定程度上解决了用户或资源的冷启动问题。

3　实验结果与分析

3.1　数据集

豆瓣活动是我国最大的社交网络, 用户可以在该平台上发布和参与各类活动并签到。该数据集包含了 100 000 多个用户, 300 000 个事件, 以及 3 500 000 条签到记录。本文经过预处理后选择了其中 20 000 个用户、15 000 个活动的 150 000 条签到记录作为实验数据集。

Foursquare 是一个大型的公开数据集, 该数据集包含 11 326 个用户, 182 968 个兴趣点, 实验中通过筛选选择其中 10 000 个用户、25 000 个兴趣点进行分析。

3.2　实验结果

为了验证算法的准确性, 本文采用了文献 [15] 提出的评估方法,具体如下:

1) 对于任意用户 u, 随机选择其签到数据中的 90% 作为训练集 S, 剩余的 10% 作为测试数据集 T。由于本文要分别计算算法对本地兴趣点和外地兴趣点推荐的准确率, T 根据不同的情况划分为本地数据和外地数据 (以兴趣点所属城市来区分)。

2) 在测试过程中, 随机选择用户 u 尚未签到的 200 个活动构成集合 E, 假设这些活动是用户不感兴趣的。

3) 将包含用户 u 的测试集中任意活动 e 加入到 E 中构成 201 个新的活动集合, 根据推荐算法选择评分最高的前 200 个活动作为 top-200 推荐列表, 如果活动 e 出现在推荐列表中, 将 hits 增 1, 否则 hits 保持不变 (hits 为评分常量)。

4) 评估标准查全率为

$$\text{Recall}=\frac{\#\text{hits}}{|T|} \tag{17}$$

本文选择以下 4 种算法进行比较:

1) 文献 [17] 提出的 IKNN 算法 (item-based k-nearest neighbors algorithm), 该算法利用 “近邻用户”来推荐感兴趣的活动, 然后根据活动地点离用户的远近进行过滤, 优先选择离用户较近的感兴趣的活动。

2) 文献 [16] 提出了 CKNN 算法 (category-based k-nearest neighbors algorithm), 该方法实质上

也是协同过滤，将用户的兴趣映射到具体的主题，进而进行推荐。

3) 文献 [11] 提出的 USG 推荐算法，该算法的核心思想还是协同过滤，线性框融合用户偏好、社交影响和地理影响这 3 种因子。

4) User-Content Topic Model(UCTM) 模型和 User-Region Topic Model (URTM) 模型，这两种模型可看作 UCRTM 模型的子模型。当 $\lambda_{l,\text{document}}$=0 时，此时模型忽略兴趣点介绍文档的内容信息，UCRTM 模型退化为 URTM 模型。当 $\lambda_{l,\text{region}}$=0 时，此时模型忽略兴趣点所处区域的主题信息，UCRTM 模型退化为 UCTM 模型。

3.3　实验结果

该模型有 9 个超参数需要设置，对于主题模型来说，超参数的值对最后的输出结果影响不大，但是会影响模型的收敛速度，这里设置 α_u、α_d、α_{ur}、α_r 为 0.1，β_w、β_l 为 0.05，所有 η 的值为 0.01。

1) UCRTM 模型为概率产生式模型，本文使用困惑度 (Perplexity) 作为评价标准，对本模型的预测能力进行评估，判断测试集 Dtest 中兴趣点生成的不确定性，Perplexity 的值越小，表示模型生成兴趣点的性能越好。Perplexity 的计算式为

$$\text{Perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{D_{\text{test}}}\log(p(l_d))}{|D_{\text{test}}|}\right\} \tag{18}$$

式中 $p(l_d)$ 由式 (15) 或 (16) 得出。由于本模型中包含了两个隐含变量 (主题数 K 和区域数 R)，为了分析这两个变量对模型生成能力的影响，首先固定隐含区域数，来观察 Perplexity 随不同主题数的变化情况。

从图 3 可以看出，当区域个数固定为 R=30 时，对于不同的主题数，Perplexity 均随着迭代次数的增加不断减小，当迭代次数达到 40 次后，Perplexity 趋于收敛。而且 Perplexity 还随着主题数 K 的增大不断减小，当主题数增加到一定程度后，Perplexity 不会持续下降，反而会有一些回升。如当 K=160 时，Perplexity 的值相比于 K=80 时反而增大了，这也在一定程度上说明，合适的主题数 K 可以提高模型的推荐效果。同理固定主题数 K=80，来观察隐含区域数 R 对 Perplexity 的影响。如图 4 所示，区域数与主题数的变化情况类似，当 R=30 时，可以得到最小的 Perplexity 值。因此本实验中主题数 K 设置为 80，而区域数 R 为 30。

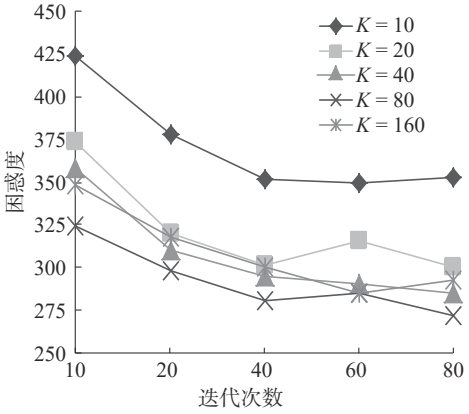

图 3　困惑度在不同隐含主题下的变化情况

Fig. 3　The perplexity changes in the number of different hidden themes

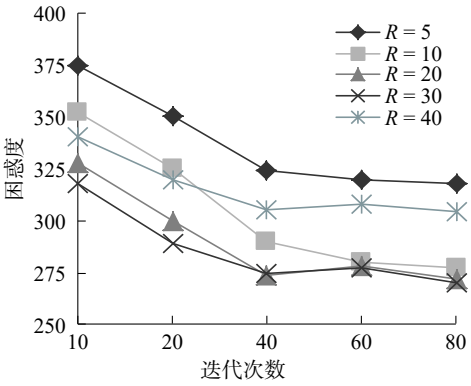

图 4　困惑度随不同隐含区域下的变化情况

Fig. 4　The perplexity changes in the number of different hidden region

2) 其次比较了各种算法的推荐准确率，因为用户的签到具有地域聚集性，本文将测试集分为两类：用户的本地活动测试集、用户的外地活动测试集。对豆瓣数据集和 Foursquare数据集进行了分析。图 5~8 分别给出了 6 种算法在两种数据集下的 top-N 推荐准确率，推荐列表的长度 N 在 2~20 变化。

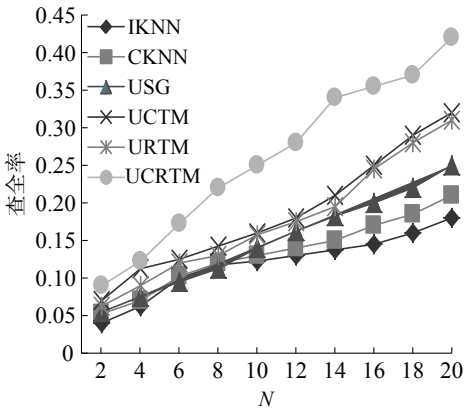

图 5　豆瓣数据集外地活动的推荐准确率比较

Fig. 5　Comparison of recommended accuracy out of town for Douban dataset

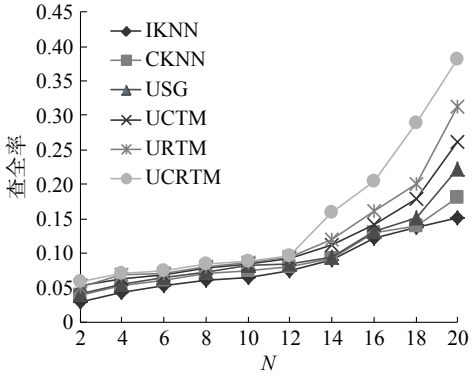

图 6　Foursquare 数据外地活动的推荐准确率比较

Fig. 6　Comparison of recommended accuracy out of town for Foursquare dataset

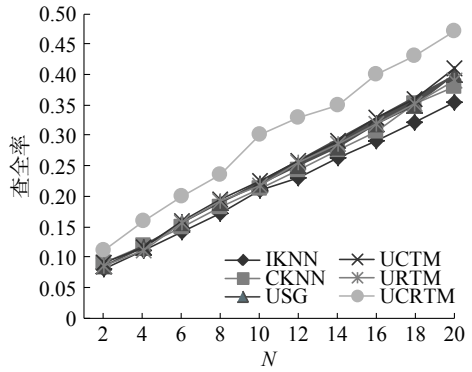

图 7　豆瓣数据集本地活动的推荐准确率比较

Fig. 7　Comparison of recommended accuracy in locality for Douban dataset

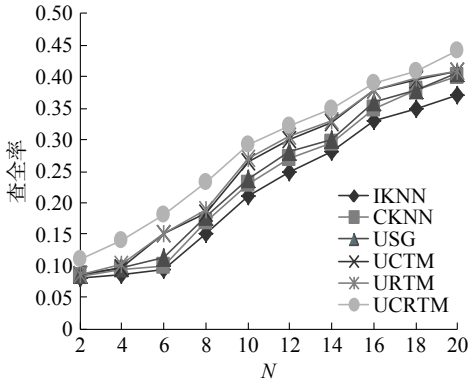

图 8　Foursquare 数据本地活动的推荐确率比较

Fig. 8　Comparison of recommended accuracy in locality for Foursquare dataset

由图 5 和图 6 可以看出，随着 N 的不断增加，各种算法的准确率都是不断提高的。对于外地活动的推荐,UCRTM、UCTM、URTM 优于 USG、CKNN、IKNN 算法，因为后 3 种方法为协同过滤算法，数据的稀疏性对其影响较大，用户或地点相似性在稀疏的环境下计算不准确，导致推荐准确率不高。由于 USG 算法考虑了社交好友的影响，推荐效果略好于 CKNN 和 IKNN 算法。而隐含主题模型受数据稀疏性的影响较小，在模型中兴趣点的隐含主题同时由用户兴趣分布、兴趣点介绍文档主题分布以及兴趣点所属区域的主题分布的影响，这些信息是对用户签到数据的有益补充。UCTM 和 URTM 均只考虑了其中两方面的影响，所以其推荐的准确程度不如 UCRTM 模型。

由图 7 和图 8 可以看出，在本地活动推荐中，UCRTM 模型同样优于其他各种方法，但考虑到用户本地签到的数据较多，采用协同过滤类的算法本身能够准确计算用户的相似性，不需要其他补充信息也能获得较高的准确率，因此最终各种方法的性能差距不大。但是本模型能够扩展更多的上下文信息，可靠性更高。

4　结束语

本文提出的用户−区域−内容联合推荐模型能够克服数据稀疏性以及弱语义性的影响，与其他方法相比有较高的推荐的准确率。以后还将进一步改善模型，增加环境、时间等上下文因素。其次该模型除了应用于兴趣点推荐外，还能将学习出的重要参数 (如用户的兴趣爱好、用户的活动特性、地理区域的主题等) 用于其他的 web 服务中。

参考文献:

[1] 罗军舟, 吴文甲, 杨明. 移动互联网: 终端、网络与服务[J]. 计算机学报, 2011, 34(11): 2029–2051.
LUO Junzhou, WU Wenjia, YANG Ming. Mobile internet: terminal devices networks and services[J]. Chinese Journal of Computers, 2011, 34(11): 2029–2051.

[2] YU Fei, CHE Nan, LI Zhijun, et al. Friend recommendation considering preference coverage in location-based social networks[C]//Proceedings of the 21st Pacific-Asia Conference, Advances in Knowledge Discovery and Data Mining. Jeju, South Korea, 2017: 91–105.

[3] ZHAO Yan, ZHU Jia, JIA Mengdi, et al. A novel hybrid friends recommendation framework for twitter[C]//Proceedings of the First International Joint Conference, Web and Big Data. Beijing, China, 2017: 83–97.

[4] YU Yonghong, WANG Hao, SUN Shuanzhu, et al. Exploiting location significance and user authority for point-of-interest recommendation[C]//Proceedings of the 21st Pacific-Asia Conference, Advances in Knowledge Discovery and Data Mining. Jeju, South Korea, 2017: 119–130.

[5] INTERDONATO R, INTERDONATO A. Personalized recommendation of points-of-interest based on multilayer local community detection[C]//Proceedings of the 9th International Conference, Social Informatics. Oxford, 2017:

552–571.

[6] YU Yonghong, GAO Yang, WANG Hao, et al. Joint user knowledge and matrix factorization for recommender systems[C]//Proceedings of the 17th International Conference, Web Information Systems Engineering. Shanghai, China, 2016: 77–91.

[7] BERJANI B, STRUFE T. A recommendation system for spots in location-based online social networks[C]//Proceedings of the 4th Workshop on Social Network Systems. Salzburg, Austria, 2011: 4.

[8] CHENG Chen, YANG Haiqin, KING I, et al. Fused matrix factorization with geographical and social influence in location-based social networks[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 17–23.

[9] YE Mao, YIN Peifeng, LEE W C. Location recommendation for location-based social networks[C]//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose, USA, 2010: 458–461.

[10] FERENCE G, YE Mao, LEE W C. Location recommendation for out-of-town users in location-based social networks[C]//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, USA, 2013: 721–726.

[11] YE Mao, YIN Peifeng, 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. Beijing, China, 2011: 325–334.

[12] YIN Hongzhi, CUI Bin, SUN Yizhou, et al. LCARS: A spatial item recommender system[J]. ACM Transactions on Information Systems, 2014, 32(3): 11.

[13] YIN Hongzhi, CUI Bin, Zhou Xiaofang, et al. Joint modeling of user check-in behaviors for real-time point-of-interest recommendation[J]. ACM Transactions on Information Systems, 2016, 35(2): 11.

[14] YIN Hongzhi, ZHOU Xiaofang, CUI Bin, et al. Adapting to user interest drift for poi recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2566–2581.

[15] CREMONESI P, KOREN Y, TURRIN R. Performance of recommender algorithms on top-n recommendation tasks[C]//Proceedings of the 4th ACM Conference on Recommender Systems. Barcelona, Spain, 2010: 39–46.

[16] BAO Jie, ZHENG Yu, 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. Redondo Beach, USA, 2012: 199–208.

[17] LINDEN G, SMITH B, YORK J. Amazon. com recommendations: item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1): 76–80.

作者简介：

涂飞，男，1979 年生，讲师，主要研究方向为服务计算、推荐系统。主持并参研省部级以上科研项目 7 项。