一种基于多类型情景信息的兴趣点推荐模型*

胡德敏[a,b]，杨 晨[a]
（上海理工大学 a. 光电信息与计算机工程学院；b. 计算机软件技术研究所，上海 200093）

摘 要：当前最新的兴趣点推荐工作开始融合地理、文本和社交信息进行推荐，但是还存在信息挖掘不充分的情况。为此，提出了改进的多类型信息融合的联合概率生成的兴趣点推荐模型。首先提出了自动学习文档话题数目的分层狄利克雷过程主题模型，学习用户和兴趣点相关兴趣话题；同时，利用由签到分布决定带宽大小的核密度估计法，个性化地理信息对用户签到行为的影响，而且还融合了用户位置访问序列中已访问兴趣点对待访问兴趣点的影响，即序列模式的影响；然后综合考虑了用户社交关系的影响；最后基于联合概率生成模型，融合文本、地理、社会和序列信息，提出 TGSS-PGM 兴趣点推荐模型，依据计算结果从而生成兴趣点推荐列表推荐给用户。实验结果表明，该模型在推荐准确率等多种评价指标上都取得了更好的结果。

关键词：基于位置的社交网络；兴趣点推荐；隐马尔可夫链；核密度估计；话题模型；社交影响
中图分类号：TP181 **文献标志码**：A **文章编号**：1001-3695(2018)06-1636-05
doi:10.3969/j.issn.1001-3695.2018.06.008

Point-of-interest recommendation model based on multi-type contextual information

Hu Demin[a,b], Yang Chen[a]
(*a. School of Optical-Electrical & Computer Engineering, b. Institute of Computer Software & Technology, University of Shanghai for Science & Technology, Shanghai* 200093, *China*)

Abstract: The state-of-the-art studies started paying attention to comprehensively analyze geographical information, comment information and social information, but there is still insufficient information mining. To this end, this paper proposed a joint probabilistic generative model of multi-information fusion. Firstly, the framework learned interest topics of users and POIs through textual information, by exploiting the aggregated hierarchical Dirichlet process model, which could automatically learn the number of topics, to replace latent Dirichlet allocation model. Secondly, according to the kernel density estimation method, whose bandwidth depended on the check-in distribution, the framework conducted personalized modeling of geographic information. Thirdly, it also took consideration of sequential patterns, which was the impact of the visited location to the non-visited location. Then, it modeled social relevance comprehensively. At last, based on the joint probabilistic generative model, this paper proposed the TGSS-PGM model, exploiting multi-type contextual information and incorporating these factors effectively. Experimental results in real world social network show that the proposed model outperforms state-of-the-art recommendation algorithms in terms of precision and rating error.

Key words: location-based social network(LBSN); location recommendations; additive Markov chain; kernel density estimation; topic model; social influence

0 引言

随着移动设备、无线通信和位置采集技术的不断发展，基于位置的社交网站，如 Foursquare、Gowalla 和 Facebook places 等应运而生，使得基于位置的社交网络(location-based social network，LBSN)每天有多达上亿次的访问量。用户能够通过 LBSN 建立社交联系，以“签到”的形式与朋友分享他们当前的地理位置(兴趣点)和社交活动，并对已访问的兴趣点(point-of-interest，POI)，作出文字评价。基于位置服务的兴趣点推荐，能从大量的兴趣点中有效地帮助用户过滤掉不感兴趣的兴趣点，找到他们可能感兴趣的兴趣点，已经成为推荐领域的热点研究问题。

不同于传统推荐任务，兴趣点推荐是一个基于上下文信息的位置感知的个性化推荐[1]。一方面，兴趣点签到数据包含多种类型的情境信息，包括地理位置、文本内容、社会关系、签到时间信息等；另一方面，兴趣点的各种信息均是不完整且模糊的。针对兴趣点相关数据的不完整性和多类型性，选用联合概率生成模型对各类信息进行有效融合，避免了概率矩阵分解模型的矩阵稀疏性过高而造成的算法时间和空间复杂度高的问题。在对各个类型的情境信息分别进行处理时，由于兴趣点相关的文本信息具有不完整性，一般采用话题模型来处理文本信息挖掘。但是目前的文本内容信息挖掘模型通常选用潜在狄利克雷分配(latent Dirichlet allocation，LDA)模型及其派生模型。这些模型需事先指定主题数目，并且模型的主题挖掘效果直接依赖指定主题的数目是否合适。而在不具备任何先验知识的情况下，用户难以准确地估算出主题数[2]。因此，造成兴

收稿日期：2017-01-11；**修回日期**：2017-03-08 **基金项目**：国家自然科学基金资助项目(61170277，61472256)；上海市教委科研创新重点项目(12zz137)；上海市一流学科建设项目(S1201YLXK)

作者简介：胡德敏(1963-)，男，上海人，副教授，博士，主要研究方向为计算机网络、分布式计算、云计算；杨晨(1991-)，女，江苏徐州人，硕士，主要研究方向为推荐系统(776044543@qq.com).

趣点的文本挖掘效果一般。还有在现实生活中,人们的移动顺序通常会蕴涵着序列模式,但是当前多融合的兴趣点研究工作却忽略了序列模式的影响。不难发现人们访问兴趣点的序列会受到行为模式的影响,一些固定的行为模式或生活习惯往往会影响用户访问的位置。具体来说,人们倾向于在运动健身之后去餐馆用餐补充能量,即已访问兴趣点对待访问兴趣点具有一定程度的决定作用。因此,序列模式是兴趣点推荐的重要信息,应当利用起来提高推荐精度。再者即使目前的研究开始利用核密度估计方法建模地理位置对用户访问兴趣点的影响,仅以已签到兴趣点间的距离计算带宽,评估签到分布,造成带宽大小与已获取的兴趣点签到密度无关,导致距离相同的高签到密度和低签到密度的带宽相同。因此,地理信息挖掘模型有待进一步优化。

因此本文针对以上所提出的问题进行改进,以便提高兴趣点推荐的准确性。首先,在处理兴趣点推荐的文本信息时,不是采用需要预先指定主题数目的主题模型——LDA 模型,而是基于无参主题模型的分层狄利克雷过程模型(hierarchical Dirichlet process,HDP),提出一种聚合 HDP 模型(aggregation hierarchical Dirichlet process,AHDP)学习用户兴趣。该模型的优势在于可以根据原始数据特征推断出最终的类别数量,而无须事先指定。其次,基于多阶隐马尔可夫链(multi-order additive Markov chain,M-AMC)建模用户签到行为的序列模式。由于兴趣点签到的概率分布不仅取决于用户最近访问的兴趣点,同时也与用户签到序列中之前访问的兴趣点相关,所以选用 M-AMC 模型预测用户访问新的兴趣点的概率分布。接着,利用核密度估计,基于已有的兴趣点签到密度分布学习出自适应带宽,为用户在地理坐标上评估一个个性化签到分布。然后,综合用户社交关系和居住距离,计算出用户社交相似度。最后,融合文本、序列、地理和社交信息,提出了一种联合概率生成模型 TGSS-PGM(probabilistic generative model based on textual geographical social and sequential influence)。

1　相关工作

相比于普通推荐系统,基于位置社交网络的兴趣点推荐中包含多种多源上下文信息(如文本类标签信息、地理位置和日期),在本章将按照位置社交网络中包含的以各类信息建模的方式分别介绍兴趣点推荐的相关研究工作。

位置社交网络中文本的常用处理方式有两种:一种是使用自然语言处理的方式[3];另一种是使用概率主题模型的方式来抽取用户的感兴趣主题[4]。自然语言处理方法会针对用户的点评文本计算出一个情感的极性(喜欢或者不喜欢),然后将该情感得分融入到模型中。而使用概率主题模型进行文本挖掘则相对更为直观[1,5~7]。概率主题模型以用户的点评文本作为输入,输出用户的兴趣主题,其中每个主题包含了一些表征该主题的词汇,同一类主题中的词汇具有意义上的相近性[6]。当前兴趣点文本信息挖掘主要利用 LDA 模型[1,5~7],然而该模型需要预先指定主题数目,并且主题数目的选取直接影响了主题挖掘效果。地理学第一定律(tobler's law)指出用户对位置的访问随着位置与用户间的距离成反比。事实上地理邻近性显著地影响用户在兴趣点上的签到行为,地理信息被集中用于兴趣点推荐[7]。Ye 等人[8]假设同一用户访问的两个位置的距离满足幂律分布(power law distribution,PD)。幂律分布保证了随着兴趣点与用户间距离的增大,用户对于兴趣点签到的概率越小,但却忽略了用户的签到在地理位置上的聚集效应。Cheng 等人[9]基于地理位置聚集效应,假设用户的签到位置在每个中心点服从高斯分布,采用多中心高斯模型拟合用户访问位置与中心点之间的距离分布。Zhang 等人[10]基于核密度估计,为每个用户采用固定带宽核密度评估方法建模兴趣点的地理签到分布。任星怡等人[7]提出了自适应带宽的核密度方法,是目前最为优秀的模型。因此,本文引用了文献[7]对于地理信息处理的部分模型。

人们的位置移动规律表现出显著的序列影响。目前研究兴趣点推荐中序列影响的建模方式主要有三种。Zheng 等人[11]从所有用户的签到记录中挖掘出签到最多的兴趣点序列,为用户推荐旅游路线。但是该方法没有考虑用户的个性化特点,为所有用户返回相同的兴趣点序列。Cheng 等人[12]通过有关用户的外部特征进行建模来推荐个性化位置序列。用户的这些外部特征一般从图片中抽取,然而,签到数据中经常缺少图片。另外是基于序列影响进行位置推荐的研究工作,Cheng 等人[13]做了相关工作,他们根据每个用户自身签到过的位置序列,为该用户学习个性化序列模型,进行兴趣点推荐。然而这种方法要求每个用户不少于 100 个签到位置,不适于实际应用中多数用户兴趣点签到数据稀疏的情况。Zhang 等人[14]提出了利用隐马尔可夫链挖掘序列信息,遗憾的是,在他们的兴趣点推荐模型中却忽略了文本内容信息。

事实上,朋友之间会有共同的兴趣,因此社交联系在基于位置的社交网络中广泛应用来提高推荐质量[14]。具体的做法是,根据用户之间的社交联系和居住距离得出他们之间的相似程度,并将这个相似性融合到协同过滤中。

2　多类型信息融合的推荐模型

2.1　问题定义

从 LBSN 中的用户在 POI 上的历史签到数据提取出丰富的信息,其中包括 POI 描述和用户评论的文本信息、POI 和用户的地理信息、用户访问过的位置序列及用户的社交信息。为便于说明,表 1 列出了本文用到的关键数学符号所表示的含义。

表 1　文中关键数学符号

符号	意义	符号	意义
U	在 LBSN 上所有用户集合	$\theta_{u_i,i}$	某文档 d_{u_i} 的第 i 个词语所属主题
u_i	某用户:$u_i \in U$	W、V	文本相关唯一词集合、唯一词数量
L	在 LBSN 上所有 POI 集合	(H_1,H_2)	经度和纬度全局带宽
l_j	某 POI:对应经纬度坐标对(x_j,y_j)	r_{u_i,l_j}	用户 u_i 在兴趣点 l_j 处的签到频率
Φ	在 LBSN 上所有兴趣主题集合	S_u	位置序列表示为 $S_u=(l_1,l_2,\cdots,l_n)$
ϕ_k	主题—词语分布:$\phi_k \in \Phi$	$d_{u_i,i}$	文档 d_{u_i} 中的第 i 个词语

本文的整体框架如图 1 所示。图中文本内容挖掘模块充分利用了相关文本信息,有效评估了用户与 POI 之间兴趣相关性;序列模式挖掘模块通过已知的用户先前访问序列,推断用户访问某 POI 的概率;地理位置挖掘模块评估了用户的个性化签到分布,构建已访问兴趣点与未访问兴趣点地理相关性;社交关系挖掘模块聚集与用户有社交关系的朋友用户,并依据访问 POI 的相似性和居住距离因素得出社交相似度。TGSS-PGM 将兴趣、序列和地理相关性,加权用户社交相似度,得出最后的偏好分数,生成兴趣点推荐列表进行兴趣点推荐。

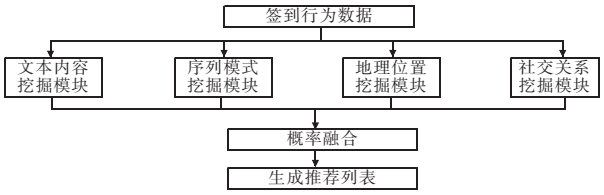

图 1 TGSS-PGM 整体框架

2.2 文本信息建模

兴趣话题提取根据签到文本信息,基于 HDP 算法,提出了 AHDP 模型学习用户兴趣。首先,聚集有关同一兴趣点的所有评论文本评论为兴趣点文档 d_{l_j},同样聚集所有同一用户的文本评论到一个用户文档 d_{u_i}。由此,得到一个大的文档集合,每个文档对应一个兴趣点或一个用户。AHDP 模型的主流程如图 2 所示。

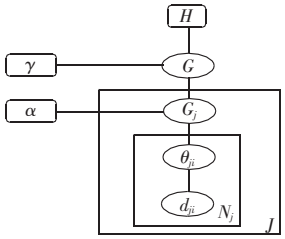

图 2 AHDP 模型流程

2.2.1 兴趣主题提取

选取参数是 λ 的随机过程 H 作为基分布,由基分布 H 和 concentration 参数 γ 构成 Dirichlet 过程 $G \sim DP(\gamma, H)$。接着以 G 为基分布,以 concentration 为参数,构造 Dirichlet 过程混合模型,为每个文档抽取主题分布 $G_j \sim DP(\alpha, G)$。从上述构造过程可以看出,各文档均服从统一的基分布 H,因此可保证各文档主题共享。每个主题 $\phi_k \in \Phi(k = 1, 3, \cdots, K)$ 是 H 的一个独立抽样。本质上每一用户或兴趣点的主题与文本词语的多项分布,对于文档 d_{u_i},$\theta_{u_i,i} \sim G_j(u_i = 1, 2, \cdots, N_{u_i})$ 是服从 G_j 的独立同分布的随机变量序列,$\theta_{u_i,i}$ 指示了文档 d_{u_i} 第 i 个词语 $d_{u_i,i}$ 所分配的主题 $\theta_{u_i,i} \in \Phi$。于是,文档 d_{u_i} 的生成过程如下:

$$d_{u_i,i} \mid \theta_{u_i,i} \sim F(\theta_{u_i,i})$$

其中:$F(\theta_{u_i,i})$ 表示在给定主题分布 $\theta_{u_i,i}$ 下,单词 $d_{u_i,i}$ 的分布,即为多项分布,与基分布 H 构成共轭分布。令 $F(\theta)$ 的概率密度为 $f(\cdot \mid \theta)$。分布 $\theta_{u_i,i}$ 有先验分布 H,其概率密度为 $h(\cdot)$。引入 $z_{i,j}$ 表示唯一词 w_i 所分配的主题。

AHDP 模型的构造过程是基于狄利克雷过程混合模型(Dirichlet process mixture model,DPMM)的。该 DPMM 由两层狄利克雷过程(Dirichlet process,DP)构成。第一层构造过程如下:

$$\beta \sim \mathrm{GEM}(\gamma), G(\phi) = \sum_{k=1}^{K} \beta_k \delta(\phi, \phi_k), \phi_k \sim H(\lambda), k = 1, 2, \cdots \quad (1)$$

此处 $\beta \sim \mathrm{GEM}(\gamma)$ 表示权重系数的构造关系。在第一层 DP 的基础之上,第二层的构造过程是以 G_0 为基分布的 DP,其中 $\varphi_j \sim DP(\alpha, \beta)$。

$$G_j(\phi) = \sum_{k=1}^{K} \varphi_{jk} \delta(\phi, \phi_{jk}), \varphi_j = (\varphi_{jk})_{k=1}^{K} \quad (2)$$

2.2.2 AHDP 模型采样

本文使用吉布斯采样,并根据有限混合模型的近似过程,求解话题—词语分布与用户—话题分布。首先需要采样潜在变量 z 的条件分布:

$$p(z_{i,k} \mid z_{-i}) \propto \frac{n_i^{(k)} + \alpha}{\sum_{k=1}^{K} n_j^{(k)} + K\alpha} \times \frac{n_k^{(w)} + \beta}{\sum_{w=1}^{V} n_k^{(w)} + K\beta} \quad (3)$$

于是,使用 $\theta_{ik} = \frac{n_i^{(k)} + \alpha}{\sum_{k=1}^{K} n_j^{(k)} + K\alpha}$ 和 $\phi_{kw} = \frac{n_k^{(w)} + \beta}{\sum_{w=1}^{V} n_k^{(w)} + K\beta}$ 来评估 θ 和 ϕ。其中:$n_k^{(w)}$ 是关于话题 k 的词频率;$n_i^{(k)}$ 是关于用户 u_i 的文档 d_{u_i} 的话题观察计数;V 是唯一词的数量;K 是话题的数量。

进一步地,利用相对熵(Kullback-Leibler divergence,KLD)计算用户文档与兴趣点文档之间的多项话题分布之间的相似性,并根据话题距离计算出用户对于兴趣点的兴趣相关度。给出计算公式如下:

$$p^{\mathrm{tex}}(u_i, l_j) = 1 - D_{KL}(\theta_{ik} \parallel \pi_{jk}) \quad (4)$$

$$D_{KL}(Q \parallel P) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \quad (5)$$

2.3 地理信息建模

地理相关性模型利用核密度估计评估签到分布,并且在高签到密度区域生成峰值自适应核评估;反之,在低签到密度区域则生成平滑的自适应核评估,提升了本文的地理信息模型签到分布的预测能力。

2.3.1 全局带宽与核函数

首先,根据用户的位置坐标和签到频率,计算出全局带宽 (H_1, H_2) 如下:

$$A = \sum_{j=1}^{n} r_{u_i,l_j}$$

$$H_1 = 1.08n^{-\frac{1}{5}} \sqrt{\frac{1}{A}\sum_{i=1}^{n}(r_{u_i,l_j} \times x_j - \frac{1}{A}\sum_{k=1}^{n} r_{u_i,l_j} \times x_k)^2} \quad (6)$$

$$H_2 = 1.08n^{-\frac{1}{5}} \sqrt{\frac{1}{A}\sum_{i=1}^{n}(r_{u_i,l_j} \times y_j - \frac{1}{A}\sum_{k=1}^{n} r_{u_i,l_j} \times y_k)^2} \quad (7)$$

将全局带宽作为核函数中的平滑参数,并且采用核函数公式,以用户签到频率作为加权系数可得

$$K_H(l_j - l_i) = \frac{1}{2\pi H_1 H_2} \times \mathrm{e}^{\left(-\frac{(x_j - x_i)^2}{2H_1^2} - \frac{(y_j - y_i)^2}{2H_2^2}\right)} \quad (8)$$

$$f(l_j \mid u_i) = \frac{1}{2}\sum_{i=1}^{n}(r_{u_i,l_j} \times K_{Hh_i}(l_j - l_i)) \quad (9)$$

2.3.2 自适应带宽评估签到概率

进一步地,根据 $f(l_j \mid u_i)$ 得出用户在当地的自适应带宽和自适应带宽对应的核密度函数。

$$h_i = (d^{-1} f(l_j \mid u_i))^{-\tau} \quad (10)$$

$$K_{Hh_i}(l_j - l_i) = \frac{1}{2\pi H_1 H_2 h_i} \times \mathrm{e}^{\left(-\frac{(x_j - x_i)^2}{2H_1^2 h_i^2} - \frac{(y_j - y_i)^2}{2H_2^2 h_i^2}\right)} \quad (11)$$

敏感参数 $\tau = 0.5$。最后,根据全局带宽和自适应当地带宽,用户 u_i 在一个未签到兴趣点 l_j 自适应核评估签到的概率分布如下:

$$p^{\mathrm{geo}}(l_j \mid u_i) = \frac{1}{A}\sum_{i=1}^{n}(r_{u_i,l_j} \times K_{Hh_i}(l_j - l_i)) \quad (12)$$

2.4 序列信息建模

从各用户访问的位置序列中抽取出位置转移序列,并用位置—位置转移图(location-location transition graph,L^2TG)表示,如图 3 所示。

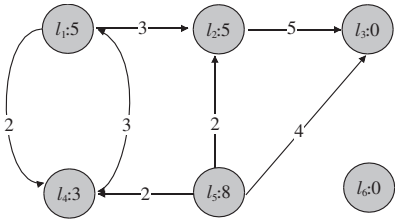

图 3 位置—位置转移图

利用 M-AMC 模型进行顺序概率的预测。更进一步地,在位置访问序列中,签到时间越晚的位置对新的可能访问位置的影响越大。因此,对 l_i 到 l_{n+1} 的转移概率加权衰减系数 $W(l_i)$ 中一般选择衰减参数 $\eta=0.1$,访问兴趣点越早,对应的衰减因子 $W(l_i)$ 则越小。最终得到潜在因子的表达式为

$$W(l_i)=2^{-\eta\times(n-i)}$$

$$f(l_{n+1},l_i,n+1-i)=\frac{W(l_i)\times TP(l_i\rightarrow l_{n+1})}{\sum_{j=1}^{n}W(l_j)} \tag{13}$$

最后,在位置序列 S_u 条件下,得出访问新位置 l_{n+1} 的顺序概率分布如下:

$$p^{\text{seq}}(l_{n+1}\mid S_u)=\sum_{i=1}^{n}\frac{W(l_i)\times TP(l_i\rightarrow l_{n+1})}{\sum_{j=1}^{n}W(l_j)}=\frac{\sum_{j=1}^{n}W(l_j)\times TP(l_i\rightarrow l_{n+1})}{\sum_{i=1}^{n}W(l_i)} \tag{14}$$

2.5　社交信息建模

U 是用户集合;L 是兴趣点集合;$u,u'\in U$ 是用户集合中的用户;X_u 和 $X_{u'}$ 是 0/1 向量,向量的每个分量表示用户是否访问过某个兴趣点。利用向量的余弦相似度,计算用户的社交相似度如下:

$$\text{sim}(u,u')=\frac{X_u\times X_{u'}}{|X_u|\times|X_{u'}|} \tag{15}$$

又有,$\text{disSim}(u,u')$ 是用户 u 与 u' 之间的距离相似度,从用户之间的社交联系和居住距离得出,具体计算公式如下:

$$\text{disSim}(u,u')=\begin{cases}1-\dfrac{\text{dis}(u,u')}{\max_{u''\in F(u)}\text{dis}(u,u')} & u'\in F(u)\\ 0 & u'\in F(u)\end{cases} \tag{16}$$

其中:$F(u)$ 是在社交网站上与用户 u 存在社交关系的用户集合;$\text{dis}(u,u')$ 是用户 u 与 u' 之间的居住距离。综合用户之间的社交相似度和居住距离相似度,得出最终的相似度为

$$\text{POISim}(u,u')=\frac{X_u\times X_{u'}}{|X_u|\times|X_{u'}|}\times\text{disSim}(u,u') \tag{17}$$

2.6　融合多类型情境信息的 TGSS-PGM

在 2.1 ~ 2.4 节中,针对文本内容、序列模式、地理位置和社交关系分别建模。本节基于联合概率生成模型,提出了融合文本、序列地理、社交的 TGSS-PGM,进行用户下一个访问位置的评分预测。融合函数的定义如下:

$$\hat{r}_{u_i,l_j}=\frac{\sum_{u'\in U\wedge u'\neq u_i}\text{SocSim}(u_i,u')\times r_{u',l_j}}{\sum_{u'\in U\wedge u'\neq u_i}\text{SocSim}(u_i,u')}\times p^{\text{seq}}(l_j\mid S_{u_i})\times p^{\text{geo}}(l_j\mid u_i)\times p^{\text{tex}}(u_i,l_j) \tag{18}$$

将所有在社交网站上与用户 u_i 社交关系的用户对兴趣点 l_j,加权上与用户 u_i 相似度之后,求出评分的平均值,接着融合式(4)(11)(13)所求的各项概率分布,最后得出用户 u_i 访问位置 l_j 测评分。

$$\hat{r}_{u_i,l_j}=\frac{\sum_{u'\in U\wedge u'\neq u_i}\text{SocSim}(u_i,u')\times r_{u',l_j}}{\sum_{u'\in U\wedge u'\neq u_i}\text{SocSim}(u_i,u')}\times p^{\text{seq}}(l_j\mid S_{u_i})\times p^{\text{geo}}(l_j\mid u_i)\times p^{\text{tex}}(u_i,l_j) \tag{19}$$

3　实验结果与分析

本章主要说明评价本文提出的 TGSS-PGM 的实验设置。实验过程中选取了最先进的兴趣点推荐模型与 TGSS-PGM 进行比较,主要包括一阶马尔可夫链模型[11~13,15,16]、地理社交模型[3,9,17~21]和混合模型 LORE[14]等。

3.1　数据集描述

Foursquare 是一个基于位置的社交网站,拥有大规模的用户量。Foursquare 网站允许用户在不同的位置进行签到。因此,利用 Foursquare 的签到数据,通过建模分析,可以定量评估人们的移动模式。本文使用由文献[7]提供的 Foursquare 数据集的一部分作为实验数据集。实验数据集的统计如表 2 所示。

表 2　数据集统计

数据集	Foursquare	数据集	Foursquare
用户的数量	5 468	社会关系的数量	35 216
POI 的数量	7 286	签到或者评价的数量	512 643

本文随机选择实验数据集的 20% 作为测试数据集,其余 80% 的数据集作为训练数据集。在本文实验数据集中,有 5 468个用户总共访问 7 286 个 POI。

3.2　评价指标

推荐准确率:通常,推荐技术为每个推荐候选项计算出相应的评分。为每个目标用户返回最高评分的 top-k 兴趣点列表作为推荐结果。在评价兴趣点推荐模型的推荐质量时,找到目标用户实际访问位置中,有多少位置被该推荐模型所发现是至关重要的。为了实现这一目标,本文采用准确率和召回率两个标准评价指标。

a)准确率定义为用户实际访问兴趣点的个数占 k 个推荐兴趣点的比例,如下所示:

$$\text{precision}=\frac{\text{number of discovered locations}}{k}$$

b)召回率定义为用户实际访问兴趣点的个数占用户实际访问的所有位置个数的比例,如下所示:

$$\text{precision}=\frac{\text{number of discovered locations}}{\text{number of positive locations}}$$

3.3　实验参数设置

在文本内容挖掘模型中,本文设置先验分布 $\alpha=50/T$,T 为话题数目。在地理位置挖掘模型中,参数 τ 不是自由参数,当 $\tau=0.5$ 时,地理概率分布达到最佳。此外,除非另外约定,序列影响挖掘模型中的临界值 ΔT 设置为 1 d,衰减率参数 η 设置为 0.05。

3.4　模型对比

本文选定了四种模型作为对比模型。

a)AMC 模型[11~13,15,16]利用一阶马尔可夫链(first-order Markov chain,FMC),仅根据签到数据的序列影响,计算用户访问新位置的可能性。

b)iGSLR 模型[21]是位置—社交感知的推荐模型的代表,该模型融合了地理位置和社交关系对用户访问兴趣点可能性的影响。

c)LCARS 模型[6,22]基于话题模型构建位置—内容感知的推荐系统来推断用户个性化兴趣和位置偏好,融合了地理位置和文本内容的混合模型。

d)LORE 模型[14]综合利用了 M-AMC 模型和二维签到数据的概率分布,推断用户访问兴趣点的概率分布。该模型是综合考虑了地理位置、社交关系和序列影响的多维融合模型。

3.5　实验结果分析

TGSS-PGM 基于 Foursquare 数据集与其余兴趣点模型的推

荐性能对比如图 4 所示。从表 2 及图 4 可知，由于 TGSS-PGM 融合了社交关系、地理位置因素、顺序因素以及评论信息影响，与其他四个对比推荐模型相比，该模型在准确率和召回率上表现出了最好的推荐质量。随着兴趣点个数 k 的增加，使得准确率不断下降和召回率不断上升。这是由于给用户推荐更多的兴趣点有助于用户发现更多的兴趣点，这样会促进用户兴趣点签到频率的增加。

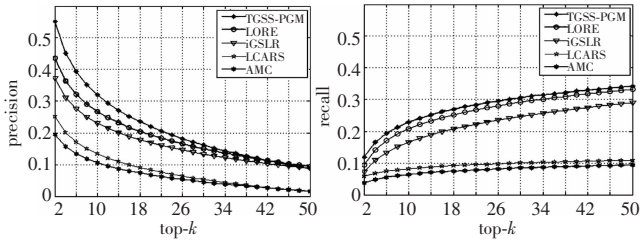

图 4 TGSS-PGM 基于 Foursquare 数据集与其余兴趣点模型的推荐性能对比

AMC 模型在预测访问新兴趣点的可能性时，仅利用最近访问的位置，而忽略早前访问兴趣点的影响，因此，会存在顺序模式挖掘不深入的问题，并且没有考虑相关评论信息、地理以及用户社会关系的影响。AMC 最终体现出第 5 优秀的推荐精度。

LCARS 模型利用 LDA 推断用户的个性化兴趣和区域的当地偏好。当地偏好或者个性化兴趣表现为话题的混合物，每个话题是 POI 上的分。但是 LCARS 忽略了 POI 上用户签到行为的地理特征、社会特征与序列影响。因此，LCARS 最终体现出第 4 优秀的推荐精度。

iGSLR 模型融合了地理位置和社交关系对用户访问兴趣点可能性的影响，却忽略了文本内容信息和序列影响。因此，iGSLR 模型最终体现出第 3 优秀的推荐精度。

LORE 模型的缺陷主要有两个方面：a）忽略了用户评价信息中所包含的文本信息；b）以固定带宽核评估方法计算地理位置的影响，没有考虑签到分布的影响。然而 LORE 模型综合了顺序影响、地理位置和社交关系，并且与 iGSLR 模型不同，其个性化了地理位置的影响。因此，相对于 iGSLR 模型，最终，其推荐精度提高较大。

TGSS-PGM 基于 Foursquare 数据集在推荐质量上表现最好，相对于 LORE 模型取得了较大的提高。原因在于 TGSS-PGM 相对于 AMC、LCARS、iGSLR、LORE 来说，全面考虑用户基于兴趣点的评论文本内容、用户社会关系、基于地理因素的影响以及用户移动序列中的顺序模型。TGSS-PGM 的推荐精度最高。

图 5 反映了判断是否是位置转移的时间间隔临界值 ΔT 对序列相关推荐技术的推荐精度的影响。需要注意的是，ΔT 对不考虑序列因素的推荐技术没有影响。如图 5 所示，可以看出 TGSS-PGM 总是优于 LORE 和 AMC。不仅如此，当 ΔT 从 0.01 增长至 100 d 时，TGSS-PGM 的准确率和召回率快速增长，接着保持相对平衡。主要因为是在开始时随着时间间隔 ΔT 的增大，转移序列的数目变多；当 ΔT 增大到接近于用户签到序列中两个连续位置的最优时间间隔时，转移序列的数目保持平衡。基于这一发现，本文倾向于将 ΔT 取值较大，并将位置序列分割成数目适当的子序列，因为多数用户的出行周期不会非常短，会具有一定的延续性。

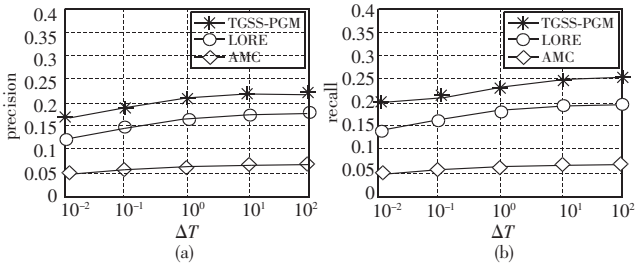

图 5 ΔT 对基于序列模式影响的兴趣点推荐性能的影响

4 结束语

如何将多种类型的上下文信息和用户的文本内容信息等多种异构数据应用到兴趣点推荐问题中，对于传统的推荐系统而言是一个挑战。本文提出一种新的推荐模型 TGSS-PGM，该模型将多源的上下文信息进行融合，并有效地进行推荐。TGSS-PGM 较好地解决了基于位置社交网络研究中利用多种情景和文本内容信息进行兴趣点推荐的问题。其最大的优势和创新点在于将用户签到文本信息、社交关系、地理位置、序列影响等多个方面的影响因素都融合到一个统一的模型中，这也是一个全新的工作。真实数据集的实验结果表明，TGSS-PGM 相对于其他的主流推荐模型在准确率和召回率两个评估指标上有着明显的提高。近年来，深度学习已经广泛应用于自然语言学习和各种情景信息的挖掘中。因此，接下来的研究工作将着眼于基于深度学习的兴趣点推荐工作。

参考文献：

[1] 任星怡，宋美娜，宋俊德. 基于用户签到行为的兴趣点推荐[J]. 计算机学报，2017，40(1)：28-51.

[2] 刘少鹏，印鉴，欧阳佳，等. 基于 MB-HDP 模型的微博主题挖掘[J]. 计算机学报，2015，38(7)：1408-1419.

[3] Liu Bin, Fu Yanjie, Yao Zijun, *et al*. Learning geographical preferences for point-of-interest recommendation[C]//Proc of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2013：1043-1051.

[4] Yang Dingqi, Zhang Daqing, Yu Zhiyong, *et al*. A sentiment-enhanced personalized location recommendation system[C]// Proc of the 24th ACM Conference on Hypertext and Social Media. New York：ACM Press，2013：119-128.

[5] 李鑫. 基于位置社交网络的地点推荐方法及应用研究[D]. 合肥：中国科学技术大学，2015.

[6] Yin Hongzhi, Sun Yizhou, Cui Bin, *et al*. LCARS：a location-content-aware recommender system[C]// Proc of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2013：221-229.

[7] 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.

[8] Ye Mao, Yin Peifeng, Lee W C, *et al*. Exploiting geographical influence for collaborative point-of-interest recommendation[C]// Proc of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM Press，2011：325-334.

[9] Cheng Chen, Yang Haiqin, King I, *et al*. Fused matrix factorization with geographical and social influence in location-based social networks[C]//Proc of the 26th AAAI Conference on Artificial Intelligence. Palo Alto, CA：AAAI Press，2012：17-23.

[10] Zhang Jiadong, Chow C, Li Yanhua. iGeoRec：a personalized and efficient geographical location recommendation framework[J]. IEEE Trans on Services Computing，2015，8(5)：701-714.

（下转第 1675 页）

L_1 重构和 PCA 重构结果,L_1 图比 PCA 重构误差更小。

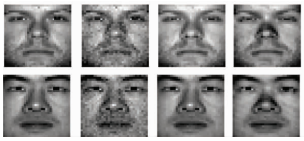

图 5　高斯噪声(强度 0.05)重构结果

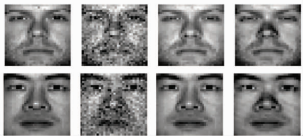

图 6　高斯噪声(强度 0.1)重构结果

4　结束语

基于稀疏编码在谱聚类图的构造中提出一种改进 L_1 稀疏表示图模型。将待测样本表示为样本集中其他样本的稀疏线性组合,优化该表示的平方误差,同时对表示系数进行 L_1 正则化。该模型构造的边权对数据噪声有很好的鲁棒性,同时能够反映数据局部线性结构。在两个标准图像数据库上对算法进行了谱聚类实验,并与几种经典聚类算法进行了比较。实验结果表明该算法具有更好的聚类性能,对噪声数据具有较好的鲁棒性。所提出模型有一个正则化参数 λ,该参数的选择对稀疏表示较为关键,如何选择仍有待进一步研究。此外,图的构造需要求解若干优化问题,因而相对一些经典算法如 PCA 耗时较长,对大规模问题如何加速算法是一个重要问题。

参考文献:

[1] Xu Rui, Wunsch D. Survey of clustering algorithms[J]. IEEE Trans on Neural Networks, 2005, 16(3): 645-678.

[2] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.

[3] 蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述[J]. 计算机科学, 2008, 35(7): 14-18.

[4] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6): 1373-1396.

[5] Wu Zhenyu, Leahy R. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1993, 15(11): 1101-1113.

[6] Shi Jianbo, Malik J. Normalized cuts and image segmentation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[7] Hagen L, Kahng A B. New spectral methods for ratio cut partitioning and clustering[J]. IEEE Trans on Computer-Aided Design of Integrated Circuits and Systems, 1992, 11(9): 1074-1085.

[8] Ding C, He Xiaofeng, Zha Hongyuan, *et al*. Spectral min-max cut for graph partitioning and data clustering[R]. Berkeley: Lawrence Berkeley National Laboratory, 2001.

[9] Zhuang Liansheng, Gao Shenghua, Tang Jinhui, *et al*. Constructing a nonnegative low-rank and sparse graph with data-adaptive features[J]. IEEE Trans on Image Processing, 2014, 24(11): 3717-3728.

[10] Cheng Bin, Yang Jianchao, Yan Shuicheng, *et al*. Learning with L_1-graph for image analysis[J]. IEEE Trans on Image Processing, 2009, 19(4): 858-866.

[11] Donoho D L. For most large underdetermined systems of linear equations the minimal L_1-norm solution is also the sparsest solution[J]. Communications on Pure and Applied Mathematics, 2006, 59(6): 797-829.

[12] Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso[J]. Annals of Statistics, 2006, 34(3): 1436-1462.

[13] Wright J, Yang A Y, Arvind G, *et al*. Robust face recognition via sparse representation[J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2009, 31(2): 210-227.

[14] Chen S S, Donoho D L, Saunders M A. Atomic decomposition by basis pursuit[J]. SIAM Review, 2001, 43(1): 129-159.

[15] Donoho D L. Compressed sensing[J]. IEEE Trans on Inform Theory, 2006, 52(4): 1289- 1306.

[16] Candès E J, Romberg J K, Tao T. Stable signal recovery from incomplete and inaccurate measurements[J]. Communications on Pure and Applied Mathematics, 2006, 59(8): 1207-1223.

[17] Tibshirani R. Regression shrinkage and selection via the Lasso[J]. Journal of the Royal Statistical Society, 2011, 73(3): 273-282.

[18] Zou Hui, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society, 2005, 67(2): 301-320.

[19] Hull J. A database for handwritten text recognition research[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1994, 16(5): 550-554.

[20] Lee K C, Ho J, Kriegman D J. Acquiring linear subspaces for face recognition under variable lighting[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(5): 684-698.

[21] Chen S S, Saunders M A, Donoho D L. Atomic decomposition by basis pursuit[J]. SIAM Review, 2001, 43(1): 129-159.

(上接第 1640 页)

[11] Zheng Yantao, Zha Zhengjun, Chua T S. Mining travel patterns from geotagged photos[J]. ACM Trans on Intelligent Systems & Technology, 2012, 3(3): 338-343.

[12] Cheng Anjung, Chen Yanying, Huang Yenta, *et al*. Personalized travel re-commendation by mining people attributes from community-contributed photos[C]//Proc of the 19th ACM International Conference on Multimedia. New York: ACM Press, 2011: 83-92.

[13] Cheng Chen, Yang Haiqin, Lyu M R, *et al*. Where you like to go next: successive point-of-interest recommendation[C]//Proc of the 23rd International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2013.

[14] Zhang Jiadong, Chow C Y, Li Yanhua. LORE: exploiting sequential influence for location recommendations[C]//Proc of the 2nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2014: 103-112.

[15] Chen Zaiben, Shen Hengtao, Zhou Xiaofang. Discovering popular routes from trajectories[C]//Proc of IEEE International Conference on Data Engineering. Piscataway, NJ: IEEE Press, 2011: 900-911.

[16] Kurashima T, Iwata T, Irie G, *et al*. Travel route recommendation using geotags in photo sharing sites[C]//Proc of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2010: 579-588.

[17] Bao Jie, Zheng Yu, Mokbel M F, *et al*. Location-based and preference-aware recommendation using sparse geo-social networking data[C]//Proc of the 20th International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2012: 199-208.

[18] Gao Huiji, Tang Jiliang, Liu Huan. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks[C]//Proc of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2012: 1582-1586.

[19] Kurashirna T, Iwata T, Hoshide T, *et al*. Geo topic model: joint modeling of user's activity area and interests for local recommendation[C]//Proc of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2013: 375-384.

[20] Wang Hao, Terrovitis M, Mamoulis N. Location recommendation in location-based social networks using user check-in data[C]//Proc of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2013: 374-383.

[21] Zhang Jiadong, Chow C Y. iGSLR: personalized geo-social location recommendation: a kernel density estimation approach[C]//Proc of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2013: 334-343.

[22] Yin Hongzhi, Cui Bin, Sun Yizhou, *et al*. Pollari K. LCARS: a spatial item recommender system [J]. ACM Trans on Information Systems, 2014, 32(3): Article No. 11.