

# 基于多维上下文感知图嵌入模型的兴趣点推荐\*

陈劲松<sup>1,2</sup>, 孟祥武<sup>1,2</sup>, 纪威宇<sup>1,2</sup>, 张玉洁<sup>1,2</sup>



<sup>1</sup>(智能通信软件与多媒体北京市重点实验室(北京邮电大学), 北京 100876)

<sup>2</sup>(北京邮电大学 计算机学院, 北京 100876)

通讯作者: 孟祥武, E-mail: mengxw@bupt.edu.cn

**摘要:** 近些年来,兴趣点推荐系统已经逐渐成为移动推荐系统领域的研究热点之一.多种因素联合建模的方法逐渐深入,如时间、空间、序列、社会化和语义信息被引入统一模型,以建模多维情景下的用户偏好.其中,嵌入学习模型作为一种有效的多因素联合建模方法,在移动推荐领域有较好的性能.然而,多数嵌入学习的模型只是简单地将显式因素,如时间戳、项目、区域、序列等嵌入到相同的空间,由于缺乏对用户和项目的语义特征的深层次挖掘,在用户签到极端稀疏时,难以精准获取用户偏好.鉴于此,提出一种多维上下文感知的图嵌入模型——MCAGE.在MCAGE中,利用主题模型提取用户和项目间的潜在语义特征,并重新定义了一系列图的节点及关联规则,设计了更有效的用户偏好公式,以此提升刻画移动用户偏好的精准度.最后,通过在真实数据集上的实验分析,证明了该模型具有更好的推荐性能.

**关键词:** 移动推荐;语义特征;嵌入学习模型;主题模型

**中图法分类号:** TP18

中文引用格式: 陈劲松,孟祥武,纪威宇,张玉洁.基于多维上下文感知图嵌入模型的兴趣点推荐.软件学报,2020,31(12): 3700-3715. <http://www.jos.org.cn/1000-9825/5855.htm>

英文引用格式: Chen JS, Meng XW, Ji WY, Zhang YJ. POI Recommendation Based on Multidimensional Context-Aware Graph Embedding Model. Ruan Jian Xue Bao/Journal of Software, 2020,31(12):3700-3715 (in Chinese). <http://www.jos.org.cn/1000-9825/5855.htm>

## POI Recommendation Based on Multidimensional Context-aware Graph Embedding Model

CHEN Jin-Song<sup>1,2</sup>, MENG Xiang-Wu<sup>1,2</sup>, JI Wei-Yu<sup>1,2</sup>, ZHANG Yu-Jie<sup>1,2</sup>

<sup>1</sup>(Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia (Beijing University of Posts and Telecommunications), Beijing 100876, China)

<sup>2</sup>(School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

**Abstract:** In recent years, the point-of-interest (POI) recommendation system has gradually become one of the research hotspots in the field of mobile recommendation systems. The method of joint modeling of multiple factors, such as time, space, sequence, socialization, and semantic information, has been gradually introduced into a unified model to compute the user preferences under multidimensional scenarios. As an effective multi-factor joint modeling method, the embedding learning model has better performance in the mobile recommendation systems. However, many of the embedded learning models just simply embed the explicit factors, such as timestamps, items, regions, sequences, etc. into the same space. Due to the lack of deep mining of user and item semantic features, it is hard to accurately obtain user preferences when the users' check-in data is extremely sparse. In view of this, a multi-dimensional context-aware graph embedding model, called MCAGE, is proposed in this study. In MACGE model, the topic model is used to extract the potential semantic features between users and items. Then, a series of graph nodes and association rules are redefined. To enhance the accuracy of

\* 基金项目: 北京市教育委员会共建项目

Foundation item: Mutual Project of Beijing Municipal Education Commission, China

收稿时间: 2018-07-20; 修改时间: 2018-12-24; 采用时间: 2019-04-22

describing the user preferences, a more effective user preference formula is designed. Finally, the results of experiments based on the real-world dataset shows that the proposed model has better recommendation performance.

**Key words:** mobile recommendation; semantic feature; embedding learning model; topic model

随着通信网络和移动设备的快速发展,基于位置的社交网络方兴未艾,移动用户可以在各种社交平台上分享自己的经历,线上虚拟社会和线下物理社会紧密联通,产生海量用户行为数据<sup>[1]</sup>.一方面,国家经济的腾飞催生了大量有价值的兴趣点,如餐馆、商场、超市等,这些项目期望增加客户量,提高商业价值;另一方面,大量兴趣点的信息对移动用户形成了严重信息负载,造成了移动用户的选择困境.这种需求和供给的矛盾,促使兴趣点(point-of-interest,简称 POI)推荐逐渐成为推荐系统领域的研究热点之一.

鉴于花费和地理区域的限制,以及隐私保护的需求,多数移动用户仅仅能提供少量签到.这意味着相比于数字化项目,如音乐和电影,POI 推荐面临更加严峻的数据稀疏性问题.传统多数协同过滤方法的性能,均因用户-项目矩阵的稀疏而出现明显下降,难以满足移动用户对推荐精度的要求.

根据用户行为特征,我们将签到数量稀疏的移动用户进一步细分为如下几类.

- (1) 签到密度较高的移动用户.这类用户虽然整体签到数较少,但是在一段时间内签到较为密集.例如,在长达 1 年的时间中,用户小张只有 10 次签到,但是这 10 次签到集中在 1 天内或一周内;
- (2) 签到密度稀疏的移动用户.这类用户签到数较少,签到记录之间的时间间隔较大.例如,长达 1 年的时间中,用户小李签到 10 次,这 10 次签到每次间隔 1 个月以上.

通过在一个真实的基于位置的社交网络数据(Yelp)上的分析,我们发现稀疏性签到的用户在所有用户中占有很大比例(0.829),其中:签到密度较高的稀疏用户占比例 0.504,签到密度较低的稀疏用户占比例 0.326.由此可知,精准地为这两类稀疏用户实现精准的 POI 推荐,有巨大的经济价值和研究意义.

近期,大量的工作研究如何使用多维信息克服签到数据稀疏性问题,时间、空间、社会化、语义等因素被引入整体模型,挖掘多元情景对用户偏好的综合影响.文献[2,3]结合时间、空间和语义信息,挖掘用户偏好在这三维空间内的精准定位.文献[4]对序列因素进行建模,分析符合用户行为序列模型的个人偏好.文献[5]利用时间因素和空间因素来刻画 POI 特征.文献[6]将时间、空间、语义等引入建模.综上所述,时间、空间和语义因素是在 POI 推荐算法中常被使用的因素,序列因素则能够有效提升推荐的性能.这是因为签到密度较高的稀疏性用户占一定比例,序列因素的建模能够有效提升这部分用户的偏好获取.然而我们发现,还没有一种方法能够有效地为签到密度较低的稀疏用户提供有价值的推荐.

对于签到密度较低的稀疏用户,难以通过建模提取精准的序列效应.受到文献[2]的影响,我们假设用户选择 POI 和其本身的特征有关,我们首先使用 LDA 主题模型<sup>[7]</sup>从用户对 POI 的描述中提取出潜在主题,并且定义了选择相似度和随机相似度,其中,选择相似度是指同一个用户选择的 POI 间的相似度的平均值,随机相似度是指用户选择的 POI 和随机抽样的 POI 间的相似度的平均值,如图 1 所示.

从图中可以发现,用户选择的 POI 间有明显的内在相似度.也就是说,用户偏向于选择那些语义相似的 POI.

基于该现象,本文构建了一种图嵌入学习模型,该模型除了统一对时间、地区和描述文字进行统一建模,还利用潜在的主题模型 LDA 提取用户签到信息的语义信息,并由此定义了 POI-POI 二部图,用以对签到密度较低的稀疏性用户行为进行建模,以此计算用户偏好.进一步发现,用户的行为存在周期性特征,因此,并非签到时间越久的行为对本次行为的影响越低,而是用户对越类似的 POI 评分更相近.例如,如果用户还吃辣的食物,那么火锅类的餐馆评分就会远高于粤菜,这一偏好受到就餐时间的影响较低.考虑到该问题,我们重新定义了用户偏好公式,以此更加精准刻画移动用户偏好.

本文第 1 节介绍移动推荐领域的相关工作.第 2 节详细介绍基于多维上下文感知图嵌入学习模型.第 3 节结合该模型设计对比实验,并给出实验结果和相关分析.第 4 节总结全文并展望下一步研究.

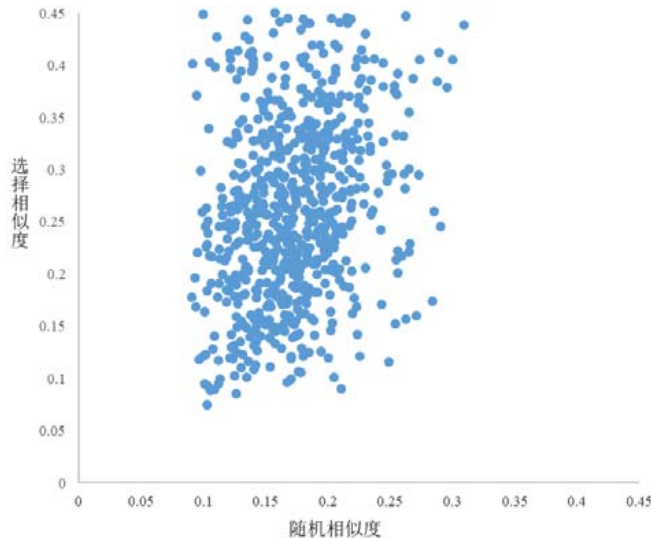


Fig.1 User check-in similarity calculation diagram

图 1 用户签到相似度计算图

## 1 相关工作

POI 推荐系统所应用的场景为:当用户到达一个地区时,为目标用户推荐一组可能感兴趣的 POI 列表.该地区可能是用户经常活动的区域,如办公点或者居住地等,也可能是一个陌生的区域,如用户去某地旅游等.文献[8]将 POI 推荐问题拆分为两个部分:本地推荐和外地推荐,并采用基于用户的协同过滤的方法,融合了地理因素和社交关系等因素,以缓解数据稀疏所带来的影响.文献[9]提出了一种基于位置感知的用户移动模型,主要解决社交网络中的影响最大化问题.文献[10]在针对地理因素的影响上,提出了一种地理稀疏填充生成模型,认为用户在外地推荐问题上的行为表现和目标地区的热门项目有很大关联.如用户在外地旅游时,最可能访问的就是当地的特色景点.即:用户在外地的表现行为不但由用户的个人偏好所影响,还受到当地项目流行度的影响.文献[11-14]都是采用了矩阵分解的方式研究地理因素对用户签到行为的影响:文献[12]提出了一种基于排名的地理因素分解模型;而文献[13]则是通过加入时间因素,探索地理因素与时间因素的依赖关系;文献[14]将社会化关系和地理因素一并融入到矩阵分解模型中,加强用户与地区之间的关联程度;文献[15]认为,只有 POI 自身的用户偏好会受到地理因素的影响,所以他们过滤掉了有关 POI 内容的用户偏好,只将 POI 自身与地区因素联合建模,用于“新城市”的推荐.

在签到数据丰富的情况下,一个用户的签到时间记录可以反映出该用户的行为习惯.为了从其中挖掘出用户的行为规律,文献[16]提出了一种协同检索建模的方法,融合了时空因素和社交关系,挖掘用户与朋友之间的签到记录的关联性.文献[17]提出了一种时间感知的协同过滤模型,认为在签到时间上有着相似规律的用户有着相似的行为模式.文献[18]认为,用户在同一天的不同时间段内签到行为不同,且在一段连续的时间域内,用户的签到行为具有一定的相似性.文献[19]提出了一种高阶张量分解的方法,分析用户签到时间的变化,以捕捉用户的长期偏好和短期偏好.文献[20]较文献[19]的不同之处在于:他们提出了一种基于张量分解的排序方法,通过研究张量因子分解以完成对 POI 的排名,并过滤掉那些排名较远的 POI,提高模型的推荐效率.文献[21]的作者认为:在研究时间因素的影响时,常常会由于对时间建模的处理不当而导致推荐效果不佳,如为每个用户在每个时间段内建模,会丢失时间相关性的影响.为此,他们放弃了对时间做离散处理然后建模,而是计算用户访问新地点的时间概率密度,以减少信息丢失.

由于在实际的移动推荐情景中,推送结果会受到地理、时间等因素的影响,为了更好地融合这些影响因素,文献[22]提出了一种嵌入学习模型,用户的签到集映射到一个低维空间中,以此来获得用户的潜在行为特征.文

献[23]提出一种基于图嵌入的学习模型,预先定义出一种图的结构和顶点关联规则,将用户的签到集合按照影响因素,如地理因素、时间因素等拆分开来,存放在对应的图中.其实验结果表明:这种图嵌入学习模型能够很好地融合多种影响因素,并提升移动推荐系统的性能.

在挖掘用户与项目的潜在主题特征方面,文献[24]认为,用户在不同地区所表现出的兴趣不同,即在空间上存在“兴趣漂移”的现象.他们结合了 LDA 主题模型,提出一种基于时空的用户主题特征模型,目的在于反馈用户在目标区域内的主题特征.文献[6]提出了一种时空主题模型,他们摒弃了离散的时间段,采用连续时间模拟签到数据,用于捕捉用户在不同时间段和不同地区的主题特征分布.然而,由于使用了连续的时间段去模拟实验,因此在签到集合较为稀疏的情况下,就不能够取得好的效果.文献[25]提出了一种基于作者主题模型的协同过滤算法,通过收集带有文字性描述的信息来提取出用户偏好主题.

从上述研究中可以得到,影响移动推荐模型性能的因素主要有 3 种:地理因素、时间因素和语义因素.然而,通过挖掘地理、时间等因素来获得用户行为规律的方式需要大量且稠密的签到数据的支撑.若提供的数据集稀疏或者对于那些签到记录不频繁的用户来说,这种方式所获得的收益将会变得很低.如果一个用户的签到记录很稠密,这就表示该用户活动很频繁.这种用户在选择能力上一般要强于那些签到记录较少的用户.而推荐系统的意义就是为这些选择能力弱的用户做推荐服务.在签到记录不足的情况下,可以通过挖掘用户和 POI 的语义特征的方式,降低稀疏数据给推荐模型带来的影响,同时也能够更好地为需要推荐的用户服务.

为了应对数据稀疏性的影响以及更好地融合以上 3 种因素,本文提出一种基于多维上下文感知的图嵌入模型的移动推荐方法.利用 LDA 主题模型挖掘用户和 POI 的潜在语义特征,并重新定义图中顶点的关联规则,以提升模型在数据稀疏情况下的推荐性能.

相较于传统的单一因素建模,引入多因素联合建模的方法能够提升对项目特征和用户偏好刻画的精准度.如在电影推荐系统中,就可以使用图嵌入学习模型,从多个维度,如电影被点击的时间段、电影的评论和电影自身内容等方面充分挖掘电影项目的特征,从而改善用户的使用体验.

## 2 一种基于多维上下文感知图嵌入模型的移动推荐方法

### 2.1 图的结构

在本文中,我们构建了 3 个二部图,分别是 POI-POI 图、POI-地区图和 POI-时间段图,将语义因素、地理因素和时间因素共同映射到一个维度为  $d$  的潜空间中,通过同时更新这 3 个图,最终,POI、地区和时间段表示为合适的  $d$  维向量.

Yin 等人在文献[23]提出的图嵌入学习模型的核心部分是 POI-POI 图,而该图是建立在一种序列效应的基础之上.在给定的一个时间间隔  $\Delta T$  内,若用户先访问了地点  $A$ ,后访问了地点  $B$ ,则认为  $A$  与  $B$  之间会存在一种关联.这种行为模式发生的次数越多,则认为  $A$  与  $B$  之间的关联性越强.即当一个用户访问地点  $A$  之后,则接下来会有很大的概率访问地点  $B$ .对于签到密度较高的用户,这种构图方法会有良好的性能,但对于签到密度较低的用户,签到间隔大,序列效应很难精准建模,该方法的性能会明显下降.由图 1 描述的实验结果可知,用户会倾向于选择语义关联度更高的项目.基于该现象,我们认为,利用语义信息能够在签到数据稀疏的情况下构建更加精准的 POI-POI 图.因此,我们重新制定了 POI-POI 图的构建规则,利用 LDA 主题模型计算 POI 潜在主题向量,并将其作为 POI 的语义特征.随后计算两个 POI 之间的语义特征相似度,当语义特征相似度高于一定阈值时,则认为两个 POI 间存在一条连接.

图的结构如图 2 所示.POI-地区图和 POI-时间段图中,节点之间符合多对一和多对多的关系.关于 POI-POI 图,选择一个 POI 点作为一个顶点集合,随后选择其余的顶点作为另一个顶点集合(忽略这些顶点的关联),那么它就是一个广义上的二部图.

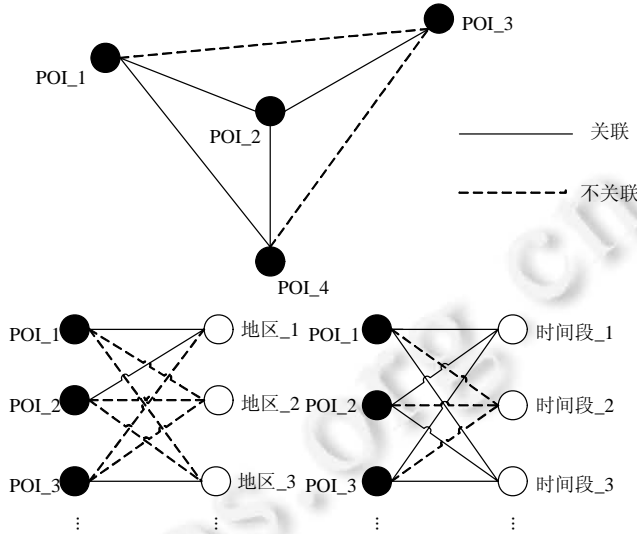


Fig.2 Structure of the graph  
图 2 图的结构

2.2 图的定义

定义 1(POI-POI 图). 设  $G_{VV}=(V \cup V, \epsilon_{vv})$  为 POI-POI 图,其中,  $V$  为 POI 的集合,  $\epsilon_{vv}$  是 POI 间边的集合.

在给定的评论集  $D_{review}$  中,对每个 POI  $v$  都提取出对应的文本集  $w_v$ ,然后使用 LDA 主题模型的方法计算其主题特征向量  $\bar{w}_v$ ,并使用余弦距离公式来计算各个 POI 的主题特征向量之间的相似度.若  $v_i$  和  $v_j$  的主题特征向量的余弦相似度大于阈值  $\alpha$ ,则将  $v_i$  和  $v_j$  连上边,且将这条边的权重置为 1.具体做法如下.

首先统计全部 POI 的评论信息,建立一个语料库;然后将每个用户的评论看作一篇文档,同时也将一个 POI 的所有评论看成一篇文档.根据文献[7]所提出的方法,计算出每篇文档主题特征分布向量,即每个用户的主题特征向量  $\bar{w}_u$  和每个 POI 的主题特征向量  $\bar{w}_v$ .

在计算得到 POI 的主题特征向量  $\bar{w}_v$  后,为了能够表示 POI 与 POI 之间的关联关系,我们使用余弦公式计算两个 POI 的主题特征向量的空间距离,即使用余弦距离来表示 POI 之间的相似程度.最后,通过设置阈值  $\alpha$  来控制相似度的划分标准.这样就可以将 POI 间的相似关系保存在 POI-POI 图中.

定义 2(POI-地区图). 设  $G_{VR}=(V \cup R, \epsilon_{vr})$  为 POI-地区图,其中,  $V$  为 POI 的集合,  $R$  为地区的集合.根据 POI  $v$  的地理信息(数据集中是以经纬度的形式表达的),找到与之相对应的地区  $r$ ,将  $v$  与  $r$  连上边,并将这条边的权重置为 1.

POI-地区图用于处理 POI 与地理因素之间的关系.考虑的移动推荐系统的应用场景,当目标用户给出当前地理位置时,则与之相邻地区的 POI 被推荐的几率就大一些.如果推荐的 POI 与目标用户当前位置相隔太远,那么就会由于不切实际的推荐而失去系统的实际意义.为此,我们利用经纬度计算了用户访问记录中相邻的两个 POI 之间的距离偏移量及所占比例,其统计结果如图 3 所示.其中:纵坐标代表所占比例;横坐标代表距离偏移量,其中一个单位刻度为 5km.从图 3 可以看出:用户的活动范围是有限的,即受到地理因素的制约.所以在为用户提供移动推荐服务时,与用户当前位置相近的 POI 点被推荐的几率会更大.

在构建 POI-地区图时,我们按照行政区对城市进行划分,将城市表示成若干个地区组成的集合;随后,根据 POI 的地理位置信息为其匹配相应的地区,并将这种地理关系保存于 POI-地区图中.

定义 3(POI-时间段图). 设  $G_{VT}=(V \cup T, \epsilon_{vt})$  为 POI-时间段图,其中,  $V$  为 POI 的集合,  $T$  为时间段的集合.根据用户的签到集合,若一个 POI  $v$  在一个时间段  $t$  内被访问,则将  $v$  与  $t$  连上边,并将这条边的权重设置为时间段  $t$  内的访问频率( $v$  在时间段  $t$  内被访问的次数与  $v$  被访问总次数的比值).

POI-时间段图用于刻画 POI 和访问时间之间的关联,根据文献[18],用户的行为存在时间周期性.也就是说,在不同的时间段内,用户趋向于访问不同类型的 POI.为此,我们分析了 Yelp 数据集,从中抽取了多个项目,并将其在时间轴上的签到频数作图,获得图 4.

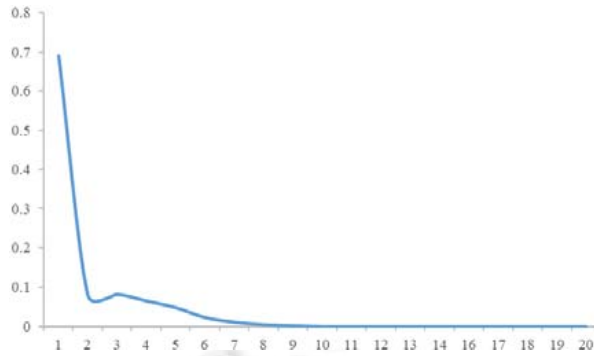


Fig.3 Distance offset between adjacent POIs

图 3 相邻 POI 之间的距离偏移量

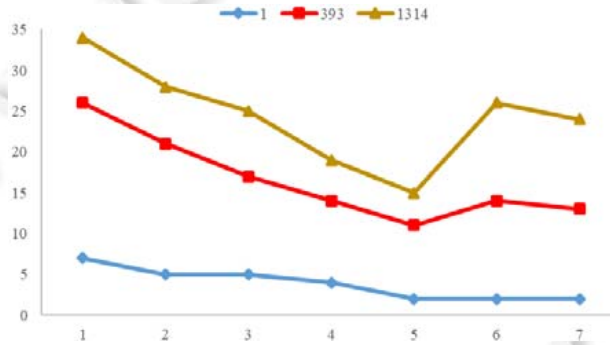


Fig.4 Visits at different times

图 4 不同时间段内的访问量

由图 4 所示,每个 POI 都存在不同的周期.例如,1 号 POI 是一所医院,由于医院通常工作日上班,因此,工作日就诊的用户较多.也就是说,一些功能性的 POI,如银行、医院等,有明显的时间周期,在不同时间段的访问人数有显著变化.目前,为了刻画时间周期性,通常会将签到时间划分成不同的时间段:一种是按照工作时间划分,将其划分为工作日和周末;一种是按照不同时间粒度划分,如月、日、小时.

在构建 POI-时间段图时,根据图 4 所显示的统计结果,我们按照一周 7 天的规则对时间进行划分.并依据 POI 被签到的时间,为其匹配对应的时间段,并将这种周期性关系保存于 POI-时间段图中.

### 2.3 图嵌入模型

给定一个二部图  $G_{VV}=(V_A \cup V_B)$ ,  $V_A$  和  $V_B$  是两个互不相交的顶点集.依据文献[23,26]中提出的方法,我们可以使用负采样的方式计算图中每个顶点在潜空间中的嵌入向量:

$$O = \arg \max \sum_{e_{ij} \in \mathcal{E}} w_{ij} \cdot \log p(v_j | v_i) \tag{1}$$

$$\log p(v_j | v_i) = \log \sigma(\vec{v}_j^T \cdot \vec{v}_i) + \sum_{n=1}^K E_{v_n \sim P_n(v)} [\log \sigma(-\vec{v}_n \cdot \vec{v}_i)] \tag{2}$$

目标函数如公式(1)所示,其训练的目标是为了让二部图中一个端点被选择时,另一边与之关联端点出现的概率即条件概率达到最大. $v_i$  和  $v_j$  是边  $e_{ij}$  的两个端点,其中, $v_i$  属于  $V_A$ , $v_j$  属于  $V_B$ , $v_n$  是通过负采样从  $V_B$  得到的顶

点,  $\vec{v}_i, \vec{v}_j$  和  $\vec{v}_n$  分别是其对应顶点的嵌入向量.  $\sigma(x)$  是 Sigmoid 函数,  $K$  是每次采样时选取负采样的边的数目. 从文献[27]中, 我们可以得到一组参数设置:  $K$  取 5, 且  $P_n(v) \propto d_v^{3/4}$ ,  $d_v$  是顶点  $v$  的出度. 在计算顶点的嵌入向量的梯度时, 我们采用文献[28]所提出的异步随机梯度的算法. 具体做法为: 如果边  $e_{ij}$  被采样, 假设此边的权重用  $w_{ij}$  表示, 那么可以利用公式(3)来计算顶点  $v_i$  的嵌入向量  $\vec{v}_i$  的梯度:

$$\frac{\partial O}{\partial \vec{v}_i} = w_{ij} \cdot \frac{\partial \log p(v_j | v_i)}{\partial \vec{v}_i} \quad (3)$$

在使用负采样方法进行嵌入向量的计算时, 对于图中的每条边而言, 一次计算的时间复杂度为  $O(\varphi \times (K+1))$ , 其中,  $\varphi$  表示每次采样所需要的时间,  $K$  表示负采样的边的数目. 假设一个图中含有  $E$  条边, 此时的时间复杂度将增加至  $O(\varphi \times K \times E)$ . 由于  $K$  在计算时取值为较小的整数, 因此计算过程中的时间复杂度取决于采样时间  $\varphi$  以及图中所含边数  $E$ . 为了降低计算复杂度, 我们使用了文献[29]中提出的别名表方法, 该方法能根据边的权重调整每条边被采样的概率, 将每次采样的时间复杂度将至  $O(1)$ . 最终, 学习二部图中顶点的嵌入向量的时间复杂度为  $O(K \times E)$ .

为了将 POI-POI 图、POI-地区图和 POI-时间段图嵌入到同一个低维潜空间中, 我们使用文献[23]中提出的联合嵌入学习方法, 通过联合训练的方式得到 POI、地区和时间段在共享低维空间中的的表述向量:  $\vec{v}, \vec{r}$  和  $\vec{t}$ .

## 2.4 计算用户偏好

通过迭代学习参数, 当图形中节点的结构稳定后, 我们能够获得 POI、地区和时间段的表述向量. 然而, 嵌入图模型不能直接学习用户的表述向量, 需要利用合理的设计, 将用户的偏好用 POI 来表述. 其中, Yin 等人在文献[23]中提出了一种基于时间延迟的动态用户偏好模型, 其计算公式如下:

$$\vec{u}_\tau = \sum_{(u, v_i, \tau_i) \in D_u \cap (\tau_i < \tau)} \exp^{-(\tau - \tau_i)} \cdot \vec{v}_i \quad (4)$$

其中,  $D_u$  表示目标用户的签到记录集合,  $\tau$  表示当前时间节点,  $\vec{v}_i$  表示 POI  $v_i$  的嵌入向量. 由公式(4)可知: 该方法将用户偏好视作用户历史访问记录的加权累积, 而每一次访问对用户偏好的权重随时间衰减, 时间越早的签到, 对目前的用户偏好的影响越小.

按照公式(4)对用户偏好进行建模时, 最近一次访问的 POI 对用户偏好的影响最大. 通过对 Yelp 数据集中用户最近一次访问记录的评分进行统计, 其结果如图 5 所示, 可以清晰地发现: 在最后一次访问中, 用户对项目的评分较为均匀地分布在 1 分~5 分的间隔中, 其中被评 1 分和 2 分的项目比例(18%)和(16%), 而如果认为最近的行为对用户偏好的影响最大, 则计算出的用户偏好中会夹杂大量低评分的项目, 明显不能真实反映用户的偏好.

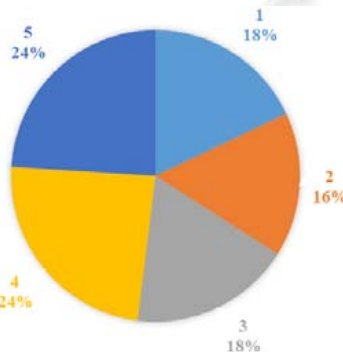


Fig.5 Last score distribution

图 5 最近一次的评分分布

此外, 在数据非常稀疏时, 则获取的用户偏好会极端接近上一次的用户偏好. 例如, 假设用户访问集合  $\{(t_1, v_1), (t_2, v_2), (t_3, v_3)\}$ , 其中,  $t_1$  和  $t_2$  距离当前时间较远,  $t_3$  距离当前时间较近, 则目前的用户偏好会和  $t_3$  时刻的偏好非常相似. 可见, 这种方法虽然突出了用户偏好的时间衰减, 但却忽略了用户偏好的周期性和连续性. 而在日常



生活中,用户的偏好往往存在周期性.例如,有的用户会在每周末去小区附近的餐馆就餐;或者连续性,例如,一些用户偏好吃火锅,可能外出就餐中火锅的比例最大.对于这些用户,POI 的类型对用户偏好的影响远甚于时间的衰减效应.

基于上述的数据和客观现实的分析,我们重新定义了用户偏好的计算公式.我们认为,用户的偏好受两方面因素的影响:一方面,显式评分能够反映用户的偏好;另一方面,用户的偏好在语义上有一定的趋向性(图 1 所示).也就是说,用户会给那些喜爱的 POI 较高的评分.同时,用户也会喜欢一类 POI,如一些四川的用户可能对火锅的偏好要远甚于其他类型的餐馆.具体的公式如下所示:

$$\bar{u} = \sum_{(u,v_i) \in D_u} \left( \frac{S_{v_i}}{S_u} + \cos(\bar{w}_u, \bar{w}_{v_i}) \right) \cdot \bar{v}_i \quad (5)$$

其中,括号内第 1 项是指 POI  $v_i$  受到用户  $u$  的喜爱程度. $S_{v_i}$  表示用户  $u$  对 POI  $v_i$  的评分, $S_u$  表示用户  $u$  的所有评分记录中的最大值. $\frac{S_{v_i}}{S_u}$  的值越大,说明用户对该 POI 的偏爱程度越大;第 2 项是指用户  $u$  和 POI  $v_i$  在语义上的相似性,旨在挖掘 POI  $v_i$  在语义上与用户  $u$  偏好的契合度. $\bar{w}_u$  表示用户  $u$  的主题特征向量; $\bar{w}_{v_i}$  表示 POI  $v_i$  的主题特征向量; $\cos(\cdot)$  表示余弦相似度,相似度越大,则表示用户  $u$  与 POI  $v_i$  的契合程度越高.

值得注意的是,两项的取值范围均为(0,1].即,这里我们认为,用户评分和语义相似性在用户偏好的计算上有相同的影响力.

展开来说,我们认为,用户对 POI 的偏好由显示反馈和语义相似性共同组成.而这样的机制还有助于揭示用户对 POI 的偏好原因.例如,假设通过计算发现,特定用户偏爱火锅,但是一家具体的火锅店评分却较低,则可能分析出其环境不能满足用户需求.因此,通过融合这两种因素,能够更准确地刻画用户偏好.另外,在缺乏评分记录时,我们依然可以通过计算语义特征的相似性描述用户偏好,一定程度上克服数据缺失的问题.

## 2.5 生成推荐列表

在为用户提供移动推荐服务时,需要知道用户当前位置  $l$  和当前时间节点  $\tau$ ,即为模型提供一个输入三元组  $q=(u,l,\tau)$ .

为了减少模型的计算量,参考图 3 的统计结果,绝大部分的用户活动范围没有超过 50km,所以我们将设定一个阈值  $\rho=50\text{km}$ .以用户的当前位置为圆心,计算半径范围中所有地区的 POI.在计算 POI 与用户的关联评分之前,先进行一次距离筛选,过滤掉那些距离用户当前位置过远的 POI,避免无意义的计算,提高模型效率.

筛选出符合用户当前位置的 POI 后,需要对这些 POI 做关联评分计算.通过上文给出的训练方法,我们可以计算得到 POI-POI、POI-地区和 POI-时间段的二部图中顶点的嵌入向量  $\bar{v}$ ,  $\bar{r}$  和  $\bar{t}$ .然后,我们根据文献[23]中提到的公式(6)计算出 POI 的相关评分.这里的评分代表 POI 与用户偏好的相似程度,评分越高,则代表用户可能会更喜欢:

$$\text{Score}(q,v) = \bar{u}^T \cdot \bar{v} + \bar{r}^T \cdot \bar{v} + \bar{t}^T \cdot \bar{v} \quad (6)$$

其中, $\bar{u}$  表示用户的偏好特征向量,由公式(5)计算得到; $\bar{r}$  是代表用户当前位置所对应地区的嵌入向量; $\bar{t}$  是代表当前时间节点所对应的时间段的嵌入向量; $\bar{v}$  是在阈值距离范围内 POI 的嵌入向量.

通过公式(6)计算出推荐列表后,按照评分降序排列,并过滤掉用户已经访问过的 POI,最后将这个列表推荐给用户.

## 2.6 方法描述

输入数据:签到集(包括用户、POI、签到时间和签到地理位置)、评论集(包含用户、POI、评分和评论)、一个包含用户实时信息的三元组  $(u,l,\tau)$ ,其中, $u$  表示目标用户, $\tau$  表示当前时间节点, $l$  当前用户所在的位置;

输出数据:一个按预测值递减排序的 POI 推荐列表.

方法流程:

步骤 1. 利用 LDA 主题模型,结合评论集中的用户对 POI 的评论数据,提取出用户和 POI 的主题特征向



量  $\bar{w}_u$  和  $\bar{w}_v$ . 根据定义 1 描述的构图方法, 构建 POI-POI 图  $G_{VV}$ .

- 步骤 2. 根据定义 2 描述的构图方法, 结合签到集中 POI 的地理位置信息, 构建 POI-地区图  $G_{VR}$ .
- 步骤 3. 根据定义 3 描述的构图方法, 结合签到集中的签到时间等信息, 构建 POI-时间段图  $G_{VT}$ .
- 步骤 4. 根据公式(1)~公式(3)以及第 2.3 节中描述的联合嵌入学习方法, 将  $G_{VV}$ ,  $G_{VR}$  和  $G_{VT}$  嵌入到同一潜在空间中, 并通过学习得到 POI、地区和时间段在共享低维空间中的表述向量:  $\bar{v}$ ,  $\bar{r}$  和  $\bar{t}$ .
- 步骤 5. 根据公式(5)对目标用户  $u$  的偏好建模, 得到目标用户  $u$  的偏好向量  $\bar{u}$ .
- 步骤 6. 依当前时间节点  $t$  和用户当前所在位置  $l$ , 并结合公式(6)计算目标用户对每个 POI 的预测分数.
- 步骤 7. 根据预测分数生成推荐列表, 并结合签到集中目标用户的签到记录, 过滤掉用户曾访问过的 POI, 得到最终的推荐列表.

### 3 实验结果与分析

首先介绍实验所用的数据集; 然后给出评价指标以及对比方法; 最后根据实验结果, 给出相应的结果分析并得出结论.

#### 3.1 数据集介绍

Yelp 数据集是由美国点评网站 Yelp 公开的内部数据集, 它包括了诸如英国的爱丁堡、德国的卡尔斯鲁厄、加拿大的滑铁卢、美国的菲尼克斯等 10 个城市, 约 366 000 个用户、61 000 个项目和 1.6 亿条签到的数据. 每条签到记录都包括了 5 个属性: 用户 ID、项目 ID、经纬度、项目内容和签到时间. 但在这个数据集中的签到时间并没有具体到时分秒, 而是以年月日的形式展现的, 如 2018-3-17. 我们将数据集中关于美国拉斯维加斯的数据记录截取出来用于本次实验.

Foursquare 是当前最流行的基于位置的社交网络之一, 其用户的活动范围包含了诸如东京、纽约和洛杉矶等全球知名城市. 与 Yelp 数据集不同的是, Foursquare 公开的数据集中, 其每条签到记录的签到时间精确到了时分秒. 我们将数据集中关于洛杉矶的数据记录截取出来用于本次实验.

实验所使用的数据集中包含的具体数据见表 1.

Table 1 Content of the dataset

表 1 数据集的内容

| 数据集名称  | Yelp          | Foursquare    |
|--------|---------------|---------------|
| 用户数    | 21 718        | 31 540        |
| 项目数    | 12 920        | 22 412        |
| 签到数    | 617 538       | 273 703       |
| 签到日期范围 | 2008.1-2015.1 | 2009.4-2013.9 |

为了验证算法的准确性, 我们将数据集拆分为训练集和测试集, 划分方式为将每个用户的签到记录依照签到时间排序, 取签到记录时间较早的 80% 作为训练集, 其余的作为测试集.

#### 3.2 评价指标

为了验证模型的推荐性能, 本文采用精确率  $Precision@N$  和召回率  $Recall@N$  为评价指标来进行测试. 各评价指标的定义如下:

$$Precision@N = \frac{|D_{test} \cap Top\_N|}{|Top\_N|} \quad (7)$$

$$Recall@N = \frac{|D_{test} \cap Top\_N|}{|D_{test}|} \quad (8)$$

其中,  $|D_{test}|$  表示测试集,  $|Top\_N|$  表示为用户生成的大小为  $N$  的推荐列表. 在后续的实验模块中, 我们采用所有用户的精确率均值和召回率均值来评价推荐模型的性能.

### 3.3 对比方法

- JIM<sup>[2]</sup>是一种融合时间因素、地理因素、内容因素和口碑因素的联合概率生成模型;
- GeoMF<sup>[11]</sup>是一种融合地理因素的矩阵分解模型;
- RankGeoFM<sup>[12]</sup>是一种基于排序的地理因子分解模型;
- GE<sup>[23]</sup>是一种融合序列效应、地理因素和时间因素的图嵌入学习模型;
- TAD-FPMC<sup>[19]</sup>是一种基于时间感知的四阶张量分解模型.

### 3.4 实验设计与结果分析

#### 3.4.1 参数设置

在 MCAGE 模型中,有 3 个参数需要调整并优化,以使得模型的推荐性能达到最优.在生成 POI-POI 图时,涉及到两个待优化的参数:POI 主题特征向量的维度,即主题数 *Topic* 和关联阈值  $\alpha$ ;在计算图模型的嵌入向量时,嵌入向量维度 *d* 的设定也需要寻求一个最优值.故在参数设置的实验模块,我们设计了两个实验,用于探索主题数与阈值、嵌入向量的维度对模型的推荐性能所带来的影响.由于 Foursquare 数据集中没有用户对 POI 的评分数据,因此在该数据集上进行实验时,我们忽略了用户评分数据对用户偏好的影响.

#### (1) 主题数与阈值的设定

当设置的主题数不同时,LDA 主题模型所挖掘得到的主题特征向量的精确度也会产生差异.而在构建 POI-POI 图时,是利用 POI 的主题特征向量之间的空间距离来表示 POI 之间的相似程度,关联阈值  $\alpha$  的设定将会影响到相似 POI 之间的筛选.为了确定这两个参数的最优值,我们固定空间向量的维度,调整主题数 *Topic* 和关联阈值  $\alpha$ ,并观测模型的推荐性能.其实验结果见表 2~表 5.

**Table 2** Precision rate under different topic numbers and thresholds on Yelp

表 2 在 Yelp 数据集上不同主题数与阈值下的精确率

| $\alpha \backslash$ Topic | 10      | 20      | 30      | 40             | 50      | 60      | 70      |
|---------------------------|---------|---------|---------|----------------|---------|---------|---------|
| 0.80                      | 0.014 8 | 0.015 3 | 0.012 9 | 0.015 8        | 0.012 7 | 0.012 1 | 0.010 3 |
| <b>0.825</b>              | 0.016 7 | 0.016 9 | 0.016 5 | <b>0.017 1</b> | 0.016 6 | 0.016 5 | 0.016 2 |
| 0.85                      | 0.016 5 | 0.016 6 | 0.016 6 | 0.016 6        | 0.016 4 | 0.016 2 | 0.016 2 |
| 0.875                     | 0.016 5 | 0.016 6 | 0.016 6 | 0.016 9        | 0.016   | 0.015 5 | 0.016 1 |
| 0.90                      | 0.016 2 | 0.014 5 | 0.014 8 | 0.016 6        | 0.011 9 | 0.010 9 | 0.010 2 |

**Table 3** Precision rate under different topic numbers and thresholds on Foursquare

表 3 在 Foursquare 数据集上不同主题数与阈值下的精确率

| $\alpha \backslash$ Topic | 10      | 20      | 30      | 40      | 50             | 60      | 70      |
|---------------------------|---------|---------|---------|---------|----------------|---------|---------|
| 0.80                      | 0.011 4 | 0.011 1 | 0.012 2 | 0.012 6 | 0.012 7        | 0.012 6 | 0.012 3 |
| <b>0.825</b>              | 0.012 6 | 0.013 1 | 0.012 8 | 0.013 5 | <b>0.013 6</b> | 0.013 6 | 0.013 5 |
| 0.85                      | 0.012 6 | 0.012 9 | 0.013 0 | 0.012 8 | 0.013 3        | 0.013 4 | 0.013 3 |
| 0.875                     | 0.012 5 | 0.012 7 | 0.013 1 | 0.013 5 | 0.013 5        | 0.013 5 | 0.013 4 |
| 0.90                      | 0.011 6 | 0.012 0 | 0.012 1 | 0.012 1 | 0.012 0        | 0.012 0 | 0.011 9 |

**Table 4** Recall rate under different topic numbers and thresholds on Yelp

表 4 在 Yelp 数据集上不同主题数与阈值下的召回率

| $\alpha \backslash$ Topic | 10      | 20      | 30      | 40             | 50      | 60      | 70      |
|---------------------------|---------|---------|---------|----------------|---------|---------|---------|
| 0.80                      | 0.050 2 | 0.051 8 | 0.043 7 | 0.053 5        | 0.043 1 | 0.041 1 | 0.034 5 |
| <b>0.825</b>              | 0.056 8 | 0.057 5 | 0.055 9 | <b>0.058 1</b> | 0.056 4 | 0.056 2 | 0.054 9 |
| 0.85                      | 0.055 9 | 0.056 4 | 0.056 4 | 0.056 4        | 0.055 7 | 0.055 1 | 0.054 9 |
| 0.875                     | 0.055 9 | 0.056 4 | 0.056 4 | 0.057 2        | 0.054 4 | 0.052 7 | 0.054 6 |
| 0.90                      | 0.055 1 | 0.049 3 | 0.050 3 | 0.056 4        | 0.040 4 | 0.037 2 | 0.034 5 |

**Table 5** Recall rate under different topic numbers and thresholds on Foursquare

表 5 在 Foursquare 数据集上不同主题数与阈值下的召回率

| $\alpha$ \ Topic | 10      | 20      | 30      | 40      | 50             | 60      | 70      |
|------------------|---------|---------|---------|---------|----------------|---------|---------|
| 0.80             | 0.034 5 | 0.032 4 | 0.037 2 | 0.037 9 | 0.038 1        | 0.038   | 0.037 6 |
| <b>0.825</b>     | 0.037 8 | 0.038 2 | 0.038 1 | 0.038 8 | <b>0.038 9</b> | 0.038 8 | 0.038 8 |
| 0.85             | 0.037 7 | 0.038 4 | 0.038 6 | 0.038 8 | 0.038 8        | 0.038 7 | 0.038 8 |
| 0.875            | 0.037 8 | 0.038   | 0.038 3 | 0.038 7 | 0.038 7        | 0.038 8 | 0.038 7 |
| 0.90             | 0.035 5 | 0.035 3 | 0.035 7 | 0.035 9 | 0.036          | 0.036   | 0.035 8 |

从实验的结果上,我们可以得到以下结论.

1. 当主题数 *Topic* 一定时,模型的精确率与召回率并不会随着阈值  $\alpha$  的增大而一直增大,它们呈现出一种先增加后减少的趋势.当阈值较小时,对 POI 的语义特征筛选效果较差,不能够有效地区分出存在关联的 POI;而当阈值较大时,筛选条件变得苛刻,此时相关联的 POI 数量会变得稀少,这就直接导致了 POI-POI 图的稀疏性变大,进而影响了模型的推荐性能;
2. 当阈值  $\alpha$  一定时,模型的精确率与召回率会随着主题数 *Topic* 的增大同样呈现出一种先增加后减少的趋势.主题数 *Topic* 代表着 POI 语义特征向量的维度.就向量性质而言,随着维度的增加,其所能表示的信息就越丰富,但考虑到在模型中的实际应用场景,模型的推荐性能并未随着主题数 *Topic* 的增加而不断增大.这是由于对于单个用户而言,其评论集中所包含的单词量是有限的.在使用 LDA 模型挖掘语义特征向量时,如果主题数 *Topic* 设置得太大,那么一个单词可能会在多个主题维度中占据较大比重,则此时的特征向量就不能准确地反映用户或者 POI 的语义特征,进而会影响到 POI-POI 图的构建以及用户偏好特征的获取,最终影响到推荐模型的性能;
3. 从整体的实验结果上来看,在 Yelp 数据集上,当阈值  $\alpha$  取 0.825、主题数 *Topic* 取 40 时,模型能取得一组较优的结果;在 Foursquare 数据集上,当阈值  $\alpha$  取 0.825、主题数 *Topic* 取 50 时,模型能取得一组较优的结果.考虑到模型的推荐性能以及计算开销等因素,在保证模型取得较好效果的情况下,还要尽量减小模型的计算开销,故我们将  $\alpha$  设置为 0.825,主题数 *Topic* 设置为 40.将这一组参数设定为最优的参数取值.

## (2) 空间向量维度的设定

空间向量维度  $d$  的设定,在一定程度上也会影响到嵌入向量对数据特征的代表能力,即不同维度的嵌入向量对数据特征描述的精确程度不同.从直观上说,空间维度越大,所表示的数据特征越精确.但这是人的直观感觉,并不意味着维度值越大,模型的推荐性能越强.

为了确定一种最优的维度设置值,我们测试了在不同空间向量维度下模型的推荐性能,其实验结果如图 6 和图 7 所示.

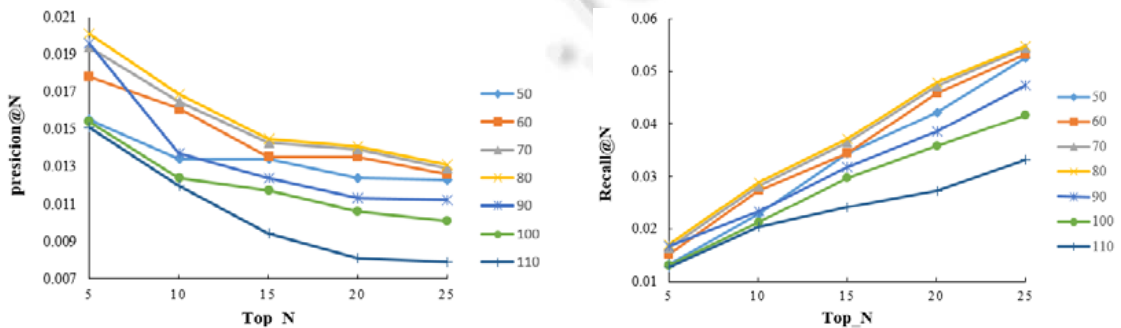


Fig.6 Precision rate and recall rate under different dimensions on Yelp

图 6 在 Yelp 数据集上不同维度下的精确率和召回率

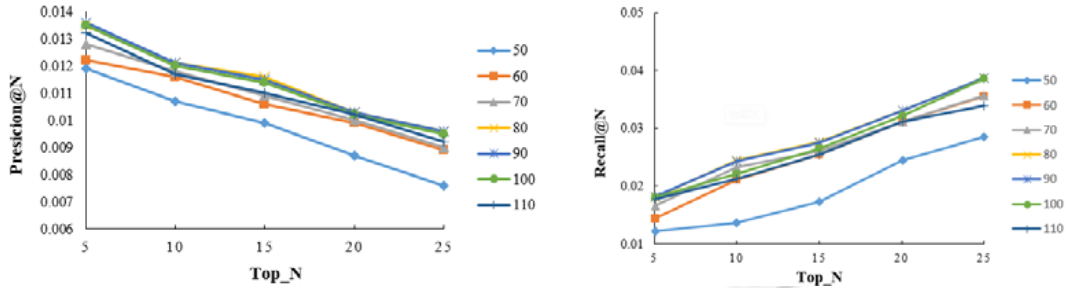


Fig.7 Precision rate and recall rate under different dimensions on Foursquare

图 7 在 Foursquare 数据集上不同维度下的精确率和召回率

从实验结果中可以看出:在不同的  $Top\_N$  的推荐下,模型的推荐性能并没有随着空间向量维度的增加而不断提高.在 Yelp 数据集上的实验结果显示:当空间向量维度  $d$  的取值在 50~80 之间时,模型所计算得到的精确率与召回率随着  $d$  的增大而不断增加;当维度  $d$  的取值在 80~110 之间时,这种增长模式却呈现出一种相反的姿态——精确率与召回率随着  $d$  的增大而不断减小;在 Foursquare 数据集上也有着相同的波动趋势,且当维度  $d$  取值为 90 时,能得到最优的实验效果.但在维度设置较小时,所得到的实验结果较差,这与在 Yelp 数据集上得到的实验结果不同.这种实验结果在一定程度上说明:空间向量的表示能力会随着维度的增加而不断变大,但同时,空间向量的表示能力具有一定的不稳定性,即维度过大或者过小都将对空间向量的表示能力造成消极影响.并且空间向量维度的增大会导致另外一个问题——收敛变慢,训练模型的耗时增加.由于在训练过程中,为了尽可能降低时间损耗,我们采用的训练方法为设置固定的迭代次数,即当模型训练的迭代次数固定时,若向量维度过大,在迭代过程中图节点的结构并没有完全达到稳定的状态,则训练出来的空间向量表示能力将会变弱.所以,综合时间损耗和推荐性能这两个方面的考虑,我们将空间向量维度  $d$  的值设置为 80.

至此,模型的参数调整实验告一段落,后续模块中,将会使用上述所确定的参数值进行相关实验.

### 3.4.2 模型对比

在得到一组最优的参数设定之后,接下来我们将 MCAGE 模型与包括 JIM,GE,GeoMF,RankGeoFM 以及 TAD-FPMC 在内的 5 种推荐模型在两个数据集上作比较,并使用精确率和召回率来评价模型的推荐性能.具体的实验结果如图 8、图 9 所示.

从整体的实验结果上来看,我们所提出的 MCAGE 模型的推荐性能都要优于其他 5 种模型.因为在 MCAGE 模型中不仅考虑到了地理因素、时间因素和语义因素,还通过诸如评分等信息来捕捉用户的动态偏好特征,所以在实验中能表现出较好的推荐性能,并且该实验结果还能说明:在引入多因素建模方面,图嵌入模型是一种高效的选择.JIM 模型的推荐性能位列第 2,这是因为在 JIM 模型中也融入了包括语义因素、地理因素以及口碑因素等,不但充分挖掘了 POI 的语义信息,还将其与 POI 的地理特征联合建模.但是在 JIM 模型中并没有考虑到用户个性化特征的计算,由于忽略了此点,它在推荐的表现不及 MCAGE 模型.TAD-FPMC 模型的效果位列第 3,虽然在该模型中引入了 POI 的分类数据,通过对用户签到记录之间的 POI 分类转移进行建模来为用户推荐,但由于数据集中存在大量的签到稀疏用户,使用这种方法对用户行为进行建模会受到较大的影响.GE 模型的效果次于 MCAGE,JIM 和 TAD-FPMC,虽然 GE 也是使用的图嵌入模型,但在 GE 中,图模型构建的核心是 POI 间的序列效应,而在数据集中,用户签到记录的时间跨度很大,不同用户签到记录的密度也有很大的差距,这些因素对需要利用序列效应来构建二部图的 GE 模型造成了较大的挑战.并且在 GE 模型中,计算用户偏好特征时使用的是一种基于时间延迟的方法,而依据我们收集统计的 Yelp 数据集中的评分数据,用户最近的签到记录并不能够准确地刻画用户的偏好特征,甚至在有些情况下,这些签到记录所体现出的偏好特征会与用户的真实偏好间呈现出一种相对立的状况.所以受到这两种因素的制约,GE 模型的推荐性能比起 MCAGE,JIM 和 TAD-FPMC 就逊色了很多.GeoMF 模型与 RankGeoFM 模型较为相似,二者都是融入了地理因素,但 RankGeoFM 采用了基于排名的因子分解,虽然二者在处理手法上有很多相似之处,但加上了基于排名的 RankGeoFM 模型在最终的推荐结

果上还是要优于只是考虑地理因素的 GeoMF 模型.而由于在 GeoMF 和 RankGeoFM 没有考虑到诸如时间因素和语义因素的影响,所以它们的推荐性能要低于之前的 3 种模型.这也从侧面说明:在数据集里包含的属性信息足够丰富的情况下,比起只考虑单因素建模,引入多因素建模的方法所展现出的推荐性能要更优越.

从 MCAGE 模型在不同的数据集上的表现来看,它在 Yelp 数据集上能发挥出更好的相对性能(同其他模型相比).其原因在于:Yelp 数据集中包含用户对 POI 的评分记录,而评分能够直接反映出用户对 POI 的偏爱程度,结合评分与语义信息能够更好地对用户偏好进行建模.但在 Foursquare 数据集中没有用户评分数据,导致对用户偏好刻画得不够准确.这也表明,融合用户评分与语义特征的偏好获取方法能够进一步提升模型的推荐性能.

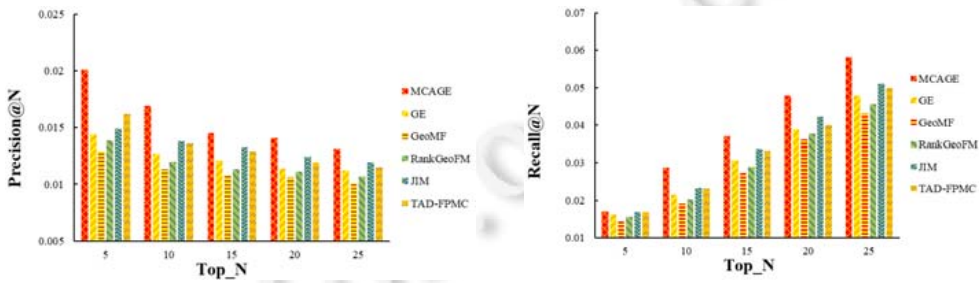


Fig.8 Precision rate and recall rate of different models on Yelp

图 8 在 Yelp 数据集上不同模型的精确率和召回率

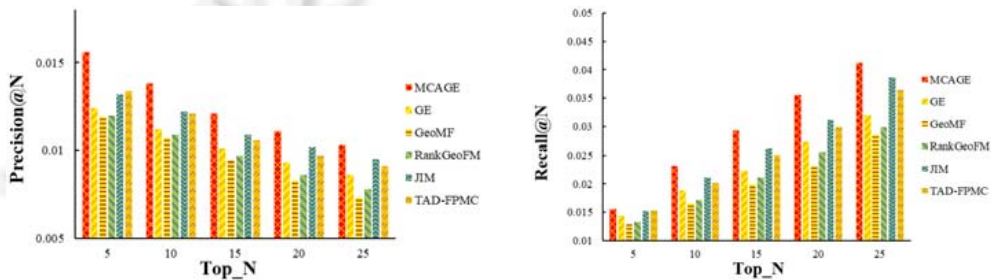


Fig.9 Precision rate and recall rate of different models on Foursquare

图 9 在 Foursquare 数据集上,不同模型的精确率和召回率

### 3.4.3 可解释性实验

语义信息特征挖掘的精确与否,将会直接影响到模型的推荐性能.如果从用户评论中所提取的语义特征信息不能与 POI 的分类信息相匹配,那么就说明用这种方式去计算 POI 的语义特征是不合理的,进而导致推荐的结果无法得到合理的解释.在此实验模块,我们随机抽取了博物馆、咖啡厅、酒店和餐馆等几种类别的下的 POI,采样出的 POI 编号与属性信息见表 6.

Table 6 POI's category information

表 6 POI 的类别信息

| POI 编号 | 类别信息                               |
|--------|------------------------------------|
| 393    | Museums, Arts & Entertainment      |
| 1888   | Food, Coffee & Tea                 |
| 2772   | Nightlife, Hotels & Travel, Hotels |
| 4396   | Wine Bars, Nightlife, Restaurants  |

然后,我们用主题模型计算它们的主题特征的分布情况,统计结果如图 10 所示.其中,横坐标代表主题编号,纵坐标表示目标主题所对应的频率.由统计结果可以看出,频率最高的几个主题编号分别为 3,11,29,33 和 36.最后,我们通过收集这些主题中频率最高的 5 个单词,其统计结果见表 7.

观察编号为 393 的 POI 的主题特征分布图,发现编号为 36 的主题所占频率较高,而该主题中频率最高的前



5个单词中包括了诸如展品、博物馆等词汇,在一定程度上能够反映出 POI393 的类别特征;在 POI1888 的主题特征分布中,主题 29 和主题 33 都占有较高的比例,查找这两个主题中的高频词可知,都是表现的餐饮行业的特征,如主题 29 中的“星巴克”一词,就具有很高的区分度;POI2772 的主题特征分布比较有特点,主题 3 占据了很高的比例,而该主题中的高频词“hotel”也很好体现该 POI 作为酒店这一类别的特征;POI4396 与 POI1888 同属于餐饮行业,同时,在其类别信息中还兼顾了诸如夜店和酒吧等词汇,这些特征信息在它的主题特征分布中也能够得到很好的表现,除了占据较大比例的主题 11 之外,主题 3 与主题 33 也都占据了较高的比例,说明该 POI 所包含的属性信息比较丰富。

从实验结果中我们可以看出,从用户评论集中提取语义特征的方法在一定程度上能够体现出 POI 的属性特征,这种方法的优点在于:能够将文本类的信息转换成特征向量的形式,利于模型的计算,并且具有较好的可解释性。

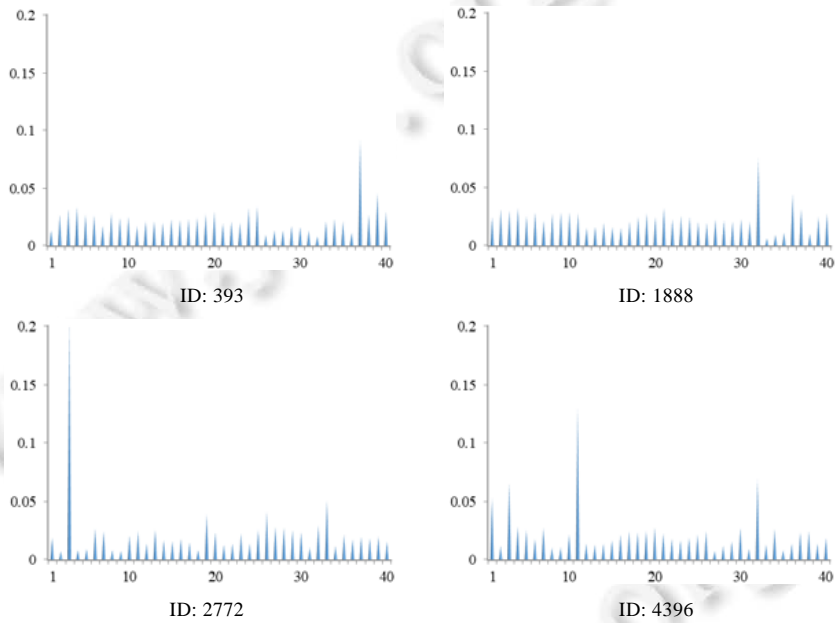


Fig.10 Thematic features distribution of different categories of POIs

图 10 不同类别 POI 的主题特征分布

Table 7 High-frequency words in the topic

表 7 主题中的高频词

| 主题编号 | 频率最高的前 5 个单词                                  |
|------|---|
| 3    | rest, car, hotel, dark, wine                  |
| 11   | food, fresh, wine, dinner, server             |
| 29   | Starbucks, service, milk, food, experience    |
| 33   | nice, taste, brought, staff, Vegas            |
| 36   | exhibits, museums, wonderful, diversion, kids |

#### 4 总结与展望

本文提出一种基于多维上下文感知的图嵌入模型的推荐方法.在引入多因素联合建模方面,我们使用主题模型挖掘用户和 POI 的语义特征,同时融入了地理因素和时间因素,用构建图模型的方式将语义、地理和时间等因素联合建模;在捕捉用户动态偏好方面,我们提出一种评分与语义特征相结合的偏好获取方法,通过用户的历史评分习惯以及用户与 POI 间的语义特征相似度作为筛选条件,突出了用户的个性化特征.实验结果表明:深层次挖掘用户和 POI 的语义特征,同时引入多因素联合建模的方法,能够有效地提高模型的推荐性能。

用户和 POI 的语义特征会随着时间的变化而发生改变,即,不同时间节点所对应的用户与 POI 的语义特征也会有不同.并且细致地区分不同时间段的语义特征,会使模型在提取用户特征和捕捉用户动态偏好时更加精确,在一定程度上能够进一步提升模型的推荐性能.如何有效地提取不同时间段内用户与 POI 的语义特征,是进一步研究时需要解决的问题.

## References:

- [1] Meng XW, Hu X, Wang LC, Zhang YJ. Mobile recommender systems and their applications. *Ruanjian Xuebao/Journal of Software*, 2013,24(1):91–108 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/4292.htm> [doi: 10.3724/SP.J.1001.2013.04292]
- [2] Yin H, Zhou X, Shao Y, Wang H, Sadiq S. Joint modeling of user check-in behaviors for point-of-interest recommendation. In: *Proc. of the 24th ACM Int'l on Conf. on Information and Knowledge Management*. New York: ACM, 2015. 1631–1640. [doi: 10.1145/2806416.2806500]
- [3] Ren XY, Song MN, Song JD. Context-Aware point-of-interest recommendation in location-based social networks. *Chinese Journal of Computers*, 2017,40(4):824–841 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2017.00824]
- [4] Meng XW, Wang S, Ahmad W, Chen Y. A sequential multi-type POI recommendation service based on location. In: *Proc. of the 5th Int'l Conf. on Identification, Information and Knowledge in the Internet of Things*. Beijing: IEEE, 2016. 325–331. [doi: 10.1109/IKI.2016.5]
- [5] Lian D, Zhang Z, Ge Y, Zhang F, Yuan NJ, Xie X. Regularized content-aware tensor factorization meets temporal-aware location recommendation. In: *Proc. of the 16th IEEE Int'l Conf. on Data Mining*. Barcelona: IEEE, 2016. 1029–1034. [doi: 10.1109/ICDM.2016.0131]
- [6] Liu Y, Ester M, Qian Y, Hu B, Cheung DW. Microscopic and macroscopic spatio-temporal topic models for check-in data. *IEEE Trans. on Knowledge and Data Engineering*, 2017,29(9):1957–1970. [doi: 10.1109/TKDE.2017.2703825]
- [7] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003,3(Jan):993–1022.
- [8] Ferenc G, Ye M, Lee WC. Location recommendation for out-of-town users in location-based social networks. In: *Proc. of the 22nd ACM Int'l Conf. on Information and Knowledge Management*. New York: ACM, 2013. 721–726. [doi: 10.1145/2505515.2505637]
- [9] Zhu WY, Peng WC, Chen LJ, Zheng K, Zhou XF. Modeling user mobility for location promotion in location-based social networks. In: *Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*. New York: ACM, 2015. 1573–1582. [doi: 10.1145/2783258.2783331]
- [10] Wang W, Yin H, Chen L, Sun Y, Sadiq S, Zhou X. Location recommendation: A geographical sparse additive generative model for spatial item recommendation. In: *Proc of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*. New York: ACM, 2015. 1255–1264. [doi: 10.1145/2783258.2783335]
- [11] Lian D, Zhao C, Xie X, Sun G, Chen E, Rui Y. GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In: *Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*. New York: ACM, 2014. 831–840. [doi: 10.1145/2623330.2623638]
- [12] Li X, Cong G, Li X L, Pham TAN, Krishnaswamy S. Rank-GeoFM: A ranking based geographical factorization method for point of interest recommendation. In: *Proc. of the 38th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*. New York: ACM, 2015. 433–442. [doi: 10.1145/2766462.2767722]
- [13] Griesner JB, Abdessalem T, Naacke H. POI recommendation: Towards fused matrix factorization with geographical and temporal influences. In: *Proc. of the 9th ACM Conf. on Recommender Systems*. New York: ACM, 2015. 301–304. [doi: 10.1145/2792838.2799679]
- [14] Cheng C, Yang H, King I, Lyu MR. Fused matrix factorization with geographical and social influence in location-based social networks. In: *Proc. of the 26th AAAI Conf. on Artificial Intelligence*. AAAI, 2012. 17–23.
- [15] Zhang C, Wang K. POI recommendation through cross-region collaborative filtering. *Knowledge and Information Systems*, 2016, 46(2):369–387. [doi: 10.1007/s10115-015-0825-8]
- [16] Zheng HT, Zhou Y, Liang N, Xiao X, Sangaiah AK, Zhao C. Exploiting user mobility for time-aware POI recommendation in social networks. *IEEE ACCESS*, 2017,PP(99):1. [doi: 10.1109/ACCESS.2017.2764074]
- [17] Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM. Time-Aware point-of-interest recommendation. In: *Proc. of the 36th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*. New York: ACM, 2013. 363–372. [doi: 10.1145/2484028.2484030]



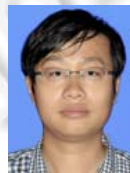
- [18] Gao H, Tang J, Hu X, Liu H. Exploring temporal effects for location recommendation on location-based social networks. In: Proc. of the 7th ACM Conf. on Recommender Systems. New York: ACM, 2013. 93–100. [doi: 10.1145/2507157.2507182]
- [19] Li X, Jiang M, Hong T, Liao L. A time-aware personalized point-of-interest recommendation via high-order tensor factorization. ACM Trans. on Information Systems, 2017,35(4):Article 31. [doi: 10.1145/3057283]
- [20] Zhao X, Li X, Liao L, Song D, Cheung WK. Crafting a time-aware point-of-interest recommendation via pairwise interaction tensor factorization. In: Proc. of the 8th Int'l Conf. on Knowledge Science, Engineering and Management. Chongqing: Springer, 2015. 458–470. [doi: 10.1007/978-3-319-25159-2\_41]
- [21] Zhang JD, Chow CY. TICRec: A probabilistic framework to utilize temporal influence correlations for time-aware location recommendations. IEEE Trans. on Services Computing, 2016,9(4):633–646. [doi: 10.1109/TSC.2015.2413783]
- [22] Feng S, Li X, Zeng Y, Cong G, Chee YM, Yuan Q. Personalized ranking metric embedding for next new POI recommendation. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. California: AAAI, 2015. 2069–2075. [doi: 978-1-57735-738-4]
- [23] Xie M, Yin H, Wang H, Xu F, Chen W, Wang S. Learning graph-based POI embedding for location-based recommendation. In: Proc. of the 25th ACM Int'l on Conf. on Information and Knowledge Management. New York: ACM, 2016. 15–24. [doi: 10.1145/2983323.2983711]
- [24] Yin H, Zhou X, Cui B, Wang H, Zheng K, Nguyen QVH. Adapting to user interest drift for POI recommendation. IEEE Trans. on Knowledge and Data Engineering, 2016,28(10):2566–2581. [doi: 10.1109/TKDE.2016.2580511]
- [25] Jiang S, Qian X, Shen J, Fu Y, Mei T. Author topic model-based collaborative filtering for personalized POI recommendations. IEEE Trans. on Multimedia, 2015,17(6):907–918. [doi: 10.1109/TMM.2015.2417506]
- [26] Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. LINE: Large-scale information network embedding. In: Proc. of the 24th Int'l Conf. on World Wide Web. New York: ACM, 2015. 1067–1077. [doi: 10.1145/2736277.2741093]
- [27] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. of the 27th Annual Conf. on Neural Information Processing Systems. New York: Curran Associates, 2013. 3111–3119.
- [28] Niu F, Recht B, Re C, Wright SJ. HOGWILD: A lock-free approach to parallelizing stochastic gradient descent. In: Proc. of the 25th Annual Conf. on Neural Information Processing Systems. New York: Curran Associates, 2011. 693–701.
- [29] Li AQ, Ahmed A, Ravi S, Smola JA. Reducing the sampling complexity of topic models. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2014. 891–900. [doi: 10.1145/2623330.2623756]

#### 附中文参考文献:

- [1] 孟祥武,胡勋,王立才,张玉洁. 移动推荐系统及其应用. 软件学报,2013,24(1):91–108. <http://www.jos.org.cn/1000-9825/4292.htm> [doi: 10.3724/SP.J.1001.2013.04292]
- [3] 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. [doi: 10.11897/SP.J.1016.2017.00824]



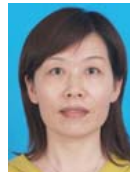
陈劲松(1995—),男,硕士生,主要研究领域为推荐系统.



纪威宇(1987—),男,硕士,CCF 学生会员,主要研究领域为数据挖掘,机器学习,推荐系统.



孟祥武(1966—),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络服务,用户需求,推荐服务.



张玉洁(1969—),女,硕士,副教授,主要研究领域为网络服务,用户需求,推荐服务.