基于地理-社会关系的多样性与个性化兴趣点推荐

孟祥福 张霄雁 唐延欢 贾 迪 齐雪月 毛 月
(辽宁工程技术大学电子与信息工程学院 辽宁 葫芦岛 125105)

摘 要 当前的兴趣点推荐方法主要侧重于拟合用户-兴趣点评分矩阵来获取用户偏好,进而为用户推荐其满意度高的兴趣点集合.然而,该类方法得到的推荐结果之间通常比较相似,不具有多样性,实际上为用户推荐与其偏好相关但彼此之间又有一定差异性的兴趣点更有实际意义.针对上述问题,本文提出一种综合考虑兴趣点之间地理关系和社会关系的多样性与个性化推荐方法.首先,将兴趣点之间的地理关系与社会关系相融合,构建了兴趣点的地理-社会关系模型,以此评估兴趣点之间的相关度.然后,在兴趣点相关度矩阵基础上,提出了基于谱聚类的兴趣点多样性划分方法,从而得到若干个具有差异性的兴趣点集合.最后,提出了基于概率因子模型的兴趣点选取与个性化排序方法,从各聚类中选出满足用户偏好的兴趣点构成推荐列表.实验结果表明,本文方法不仅使兴趣点推荐列表具有多样性,同时也具有更高的推荐准确性,从而实现了兴趣点多样性与个性化推荐的有机融合.

关键词 地理-社会关系模型;兴趣点推荐;多样性;个性化
中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2019.02574

A Diversified and Personalized Recommendation Approach Based on Geo-Social Relationships

MENG Xiang-Fu ZHANG Xiao-Yan TANG Yan-Huan JIA Di QI Xue-Yue MAO Yue
(*School of Electronic and Information Engineering, Liaoning Technical University, Huludao, Liaoning* 125105)

Abstract With the universal use of mobile Internet and the increase of spatial Web objects, Point-of-Interest (POI) recommendation system has become a new research hot-spot. Recently, most of the existing POI recommendation algorithms mainly focus on meeting the user's personalized needs or fitting the User-POI rating matrix, and striving to provide the list of POIs that are most close to the user preferences for the user. However, these algorithms neglected the diversity of the recommendation list, which may result in the POIs in the recommendation list are very similar to each other in the aspects of location, name, and other relevant features, and thus cannot effectively broaden the user's perspectives. In fact, to provide a list of POIs that are relevant to user real preferences but different from each other is more meaningful. Moreover, although there is no explicit social correlations between POIs, the social relationships (such as friendship) already existed between users who have visited the same set of POIs would make the POIs correlate with each other in the social aspect. To deal with the problem mentioned above, this paper proposes a diversified and personalized POI recommendation approach which integrally considers both the geographical distance and social relationships between POIs. First, through considering the geographic distance between POIs and the social relationships between users who have visited the

收稿日期:2018-09-11;在线出版日期:2019-05-27.本课题得到国家自然科学基金(61772249)、辽宁省自然科学基金(20170540418)、辽宁省教育厅一般项目(LJYL018)资助.**孟祥福**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为空间数据管理、推荐系统、Web 数据库查询.E-mail:marxi@126.com.**张霄雁**,硕士,工程师,中国计算机学会(CCF)会员,主要研究方向为空间数据管理、城市计算、数据挖掘.**唐延欢**,硕士,主要研究方向为推荐系统.**贾 迪**,博士,副教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为图像处理、机器学习、推荐系统.**齐雪月**,硕士研究生,主要研究方向为基于 LBSN 的推荐系统.**毛 月**,硕士研究生,主要研究方向为短文本分析、推荐系统.

POIs, a geo-social relationship model is proposed by integrating the geographical distance and social relationships between POIs, which can be used for measuring the integrated geo-social correlation degree between POIs. Based on the POI geo-social correlation matrix, the spectral clustering algorithm is leveraged to partition the POIs into several clusters that are far from each other in geo-social distance while the POIs in the same cluster are close to each other in geo-social relationships. Lastly, the POIs relevant to user preferences are selected from each cluster by using the probabilistic factor model-based algorithm and these POIs are ranked according to their satisfaction degrees to the user preferences as the diversified and personalized recommendation list. The experimental results demonstrate that our method cannot only improve the diversity of the recommendation list but also achieve the higher ranking precision. The experiments conducted on the real datasets demonstrated that the diversity of recommendation results obtained by using our method is much higher than that of the state-of-the-art methods and the precision is also a little higher than the methods without diversity, which realizes the integration of the diversification and personalization of POI recommendation.

Keywords　geo-social relationship model; POI recommendation; diversification; personalization

1　引　言

移动网络的普遍应用和空间 Web 对象(也称兴趣点，Point of Interest(POI)，如餐厅、电影院、旅馆、景点等用户感兴趣的地点)的日益增多，使得兴趣点推荐技术正成为当前推荐系统和空间数据库查询领域的研究热点. 现有兴趣点推荐算法主要是通过拟合用户-兴趣点评分矩阵来推测用户偏好，进而从兴趣点集合中获取与当前用户偏好最为相关的兴趣点集合[1-3]. 这些方法虽然具有较高的推荐准确性，但却忽略了推荐列表本身的多样性，使得推荐的兴趣点之间通常比较相似，而实际上用户希望系统能够为其推荐既贴近偏好需求且彼此之间又具有一定差异的兴趣点，即个性化与多样性的兴趣点推荐. 个性化技术在信息检索、数据库查询和推荐系统中已经有较多研究，多样性的思想近年来在信息检索和 Web 数据库查询领域逐渐得到关注. 例如，文献[4]提出了兼顾个性化与多样化的搜索引擎查询推荐方法，文献[5]研究了关系数据库的多样性查询方法. 本文将借鉴信息检索和 Web 数据库查询领域的个性化与多样性查询思想，并在作者以前关于兴趣点多样性推荐方面的研究工作[6]基础上，重点研究兴趣点之间的相关度评估和多样性与个性化的兴趣点推荐技术.

本文首先构建了兴趣点的地理-社会关系模型，从而对兴趣点之间的地理-社会关系相关度进行评估；在此基础上，然后提出了基于谱聚类的兴趣点聚类方法，使得不同类别兴趣点之间具有较远的地理-社会关系距离；最后，采用矩阵分解算法对每个聚类中的兴趣点进行用户满意度评分，选出每个集合中用户满意度最高的兴趣点，这些兴趣点根据用户对其满意度进行降序排列，从而形成多样性和个性化的兴趣点推荐列表.

本文第 2 节是相关工作；第 3 节描述总体解决方案；第 4 节提出兴趣点的地理-社会关系模型；第 5 节给出基于谱聚类的兴趣点划分方法；第 6 节描述兴趣点的多样性选取与个性化排序方法；第 7 节是效果与性能实验评价；第 8 节总结全文.

2　相关工作

近年来随着 GIS 和移动网络的迅速发展，与空间数据紧密相关的空间关键字查询和兴趣点推荐技术的研究逐渐增多. 查询与推荐最终都是要为用户返回一个结果列表，但二者的差别是：查询是根据用户明确给定的查询要求返回查询结果，推荐是通过推测用户偏好为用户主动推荐其可能感兴趣的项目列表. 因此，查询的重点在于满足用户需求，推荐的重点在于推测用户偏好和拓展用户视野. 本文方法主要涉及兴趣点相关度评估和兴趣点推荐算法，下面分别从这两个方面对国内外相关研究工作进行综述.

2.1　兴趣点相关度评估

与兴趣点相关的信息主要包括三类：一是位置

信息(如经纬度),二是文本描述信息(如名称、设施、类别等),三是签到信息(如签到用户 ID、签到时间、用户评论等).兴趣点之间的相关性评估是对兴趣点进行分析和聚类的前提,根据考虑的因素不同,可将相关度评估方法分为两大类:一类是考虑兴趣点之间的位置相近度和文本相似度,即地理-文本关系评估模型(Geo-text model);另一类是考虑兴趣点之间的位置相近度和社会关系紧密度,即地理-社会关系评估模型(Geo-social model).

2.1.1 地理-文本关系评估模型

在该模型中,兴趣点之间的位置相近度主要通过计算它们之间的欧式距离或路网(Road network)距离进行评估[2,7-9].文本相似度评估的基本思想是,先将兴趣点的文本信息进行关键字(Keyword)提取,然后计算每个关键字的 TF-IDF 权重,进而将每个兴趣点的文本信息表示成一个向量(向量的维数是所有兴趣点文本信息中不同关键字的个数),最后利用 Cosine 相似度方法计算一对兴趣点之间的文本相似度[7,9].此外,兴趣点的评论文本信息与兴趣点的名称信息不同,评论文本信息通常是短文本,缺乏统计信息和主题信息,因此当前的短文本分析方法(如文本分割、概念标注、词性标注)[10-11]通常被应用于评论文本之间的相似度评估.兴趣点的位置相近度和文本相似度,通过加权线性叠加形成最终的兴趣点之间的位置-文本相似度.该类相似度评估方法主要为空间关键字查询提供支持.

2.1.2 地理-社会关系评估模型

兴趣点之间的位置相近度评估与前述方法类似,社会关系紧密度主要通过访问兴趣点的用户群体之间的朋友关系或社会联系进行评估.随着诸多社交平台(如 QQ、微信、Facebook、Foursquare、Gowalla 等)的普及,用户之间社会联系的获取变得越来越容易.文献[12-13]通过挖掘用户的兴趣点签到行为规律和用户之间的信任关系来评估不同兴趣点之间的社会关系紧密度.文献[14]的研究发现,在社交网络中联系密切的两个用户,他们之间的签到行为也非常相似.文献[15]通过对两个基于位置的社会化网络数据集 Brightkite 和 Gowalla 中的用户签到地点和具有朋友关系的用户签到地点分布情况进行分析,发现在这两个数据集中用户首次签到地点与其朋友或朋友的朋友已签到地点的重合比例分别为 23%和 31%,说明用户之间的朋友关系对用户选择的签到地点具有较大影响,因此地点之间的社会关系紧密度可由用户之间的朋友关系来间接评估.Shi 等人[16]将兴趣点的位置关系和社会关系相融合,提出了一种地理-社会关系模型,该模型利用欧式距离计算兴趣点之间的位置相近度,利用访问兴趣点的用户群体之间的直接朋友关系进行计算(但没有考虑朋友之间的亲密程度).本文根据文献[15]的发现,并在文献[16]提出的地理-社会关系模型基础上,进一步考虑朋友关系的亲密度对兴趣点社会关系的影响,给出了兴趣点社会关系的量化方法,并深入研究了兴趣点的聚类、多样性选取和个性化排序方法.

2.2 兴趣点推荐

兴趣点推荐与传统商品推荐较为相似,但由于用访问兴趣点的概率会受到多种因素(如天气、交通等)影响,使得兴趣点推荐与商品推荐拥有更加丰富的语境,这导致兴趣点推荐系统需要考虑更多更为复杂的因素.结合传统的商品推荐方法分类,本文将近年来的兴趣点推荐研究工作分成两大类:第一类是协同过滤推荐,该类方法的基本思想是利用具有相似行为或相近兴趣偏好的用户群体的偏好,为相关用户推荐感兴趣的信息的方法;第二类是基于用户偏好和行为规律的个性化兴趣点推荐方法,该类方法通过对某个用户历史签到数据的分析,得到该用户偏好和行为规律,再结合兴趣点的时空属性,为该用户返回贴近其个性化偏好需求的兴趣点列表.

2.2.1 协同过滤推荐

协同过滤推荐有两大类方法,一类是基于内容的协同过滤推荐,另一类是基于模型的协同过滤推荐.

(1)基于内容的协同过滤推荐,文献[17]分别给出了基于用户和基于兴趣点的协同过滤推荐方法.在基于用户的协同过滤中,首先计算用户之间的相似度,然后根据相似用户对兴趣点的满意程度,为当前用户进行兴趣点推荐.文献[18-19]提出了根据用户之间的信任关系、社会联系、兴趣点的位置距离和文本关联等信息对兴趣点进行过滤推荐.基于内容的协同过滤通常会面临稀疏和扩展性问题.

(2)基于模型的协同过滤推荐,又可进一步分为三种.

第一种是基于隐语义模型(Latent Factor Model)的协同过滤推荐,该类方法根据已有用户对兴趣点的评分,构建用户-兴趣点评分矩阵,隐语义模型将预测的用户-兴趣点评分矩阵分解为两个矩阵,即用来评估用户对隐因子偏好程度的用户-隐藏因子矩阵和用来评估兴趣点贴近隐因子程度的兴趣点-隐因子矩阵,在此基础上利用矩阵分解方法拟合用户-

兴趣点的实际评分矩阵，最终将预测评分高的兴趣点集合作为推荐列表. 当前在推荐系统中被广泛采用的矩阵分解系列算法是隐语义模型的典型应用. 文献[20]分别将概率矩阵分解算法[21]和概率因素模型[22]应用于兴趣点推荐，首先利用多中心点高斯模型对用户在兴趣点的签到概率进行建模，提取出兴趣点地理特征对用户签到行为的影响，然后融合兴趣点的地理特征和社会信息，进而采用泛矩阵分解算法进行兴趣点推荐. 需要指出的是，基于矩阵分解的协同过滤方法也同样面临数据稀疏和效率不高等问题. 针对上述问题，重庆大学罗辛等人[23]在确保推荐准确率和提升算法效率等方面进行了深入研究，提出了一系列创新性研究成果. 例如，文献[23]提出了一种基于非负矩阵分解的协同过滤推荐模型，该模型利用非负单元素更新规则，对 Tikhonov 正则化项进行了整合，提出了正则化的基于单元素的非负矩阵分解模型(RSNMF)，使得整个特征矩阵的非负更新过程仅依赖于参与特征而不是全部特征，从而确保了算法具有较高的推荐准确率和较低的计算复杂度. 文献[24]针对当前矩阵分解模型不能动态融合增量式用户反馈信息的问题，提出了一种动态-静态相融合的矩阵分解推荐模型，该模型的优点是每次根据增量信息进行动态更新时只有一小部分特征被重新训练，从而大大提高了计算效率. 文献[25]针对一阶优化算法存在的问题，提出了一种基于 Hessian-free 优化算法的隐因子协同过滤模型，该模型能够通过二阶优化过程从给定的不完全矩阵中提取潜在因子. 与基于一阶优化算法的潜在因子协同过滤模型相比，该方法具有较高的预测精度和较高的计算效率.

第二种是基于聚类模型的协同过滤推荐. 空间聚类模型主要包括基于划分的聚类、基于层次的聚类和基于密度的聚类. k-means、k-medoids 和 Clarans 等[26]属于基于划分的聚类方法，该类方法通过计算兴趣点之间的空间距离对数据进行划分. 基于层次的聚类算法主要有 BIRCH、Chameleon、CURE 等[27]，该类方法对数据集进行层次分解，分解粒度的大小需要满足预先设定标准. DBSCAN[28]和 OPTICS[29]是基于密度的聚类方法的代表，基本思想是根据数据集在空间上的稠密程度进行聚类，DBSCAN 需要预先设定聚类中对象的个数阈值. 需要指出的是，传统的 DBSCAN 模型只考虑了空间对象的位置特征，没有考虑它们之间的社会关系，因此只能得到地理空间上密度较大的聚集区域. Zhao 等人[30]采用 k-means 算法将每个地理区域的兴趣点进行聚类，将其中当地用户访问量高但游客访问量低的聚类中的兴趣点推荐给游客. 文献[31]在用户签到行为数据上对用户进行聚类划分，利用概率生成模型模拟用户行为，提升了协同过滤的预测准确率.

第三种是基于贝叶斯模型的兴趣点推荐. 该类方法先对用户-项目的评分数据进行统计，然后根据贝叶斯公式原理，运用条件概率预测用户对项目的评分[32]. Ye 等人[33]提出了利用幂概率模型来评估兴趣点之间的位置影响关系的建模方法，在此基础上通过朴素贝叶斯方法进行兴趣点的协同过滤推荐. 基于朴素贝叶斯分类方法的缺点是假设兴趣点类别的各特征之间是相互独立的，而现实中各特征之间很可能存在依赖关系. 基于贝叶斯信念网络的协同过滤假设各个特征之间存在依赖关系，在计算每一类的概率时考虑了各个特征之间的依赖关系，从而使得预测结果更为准确. 例如，He 等人[34]提出了一种基于贝叶斯信念网络的最优化兴趣点类别维度向量的方法，用来预测用户下一个到访的兴趣点类别.

需要指出的是，以上协同过滤推荐方法是根据相似用户的偏好为特定用户返回推荐结果，在个性化方面尚有不足.

2.2.2　个性化推荐

对于兴趣点的个性化推荐，现有的研究工作大致可分为 3 种. 第一种是通过从用户访问记录、签到数据等历史数据中分析和学习用户偏好，对用户进行兴趣点的个性化推荐. 文献[35]提出了一种综合考虑用户对兴趣点的内在兴趣度(如用户个人偏好)和外在兴趣度(如环境影响)的融合模型，获取用户对兴趣点的偏好程度，根据外在与内在的综合偏好对兴趣点进行过滤和排序；文献[36]根据用户签到数据，利用用户对兴趣点类别的偏好变迁，提出了一种新的兴趣点推荐模型，该模型通过深入解析用户偏好的转变模式，逐步修剪候选兴趣点列表，从而提高推荐的准确率. 第二种是根据兴趣点的时空属性和用户行为规律进行推荐，该类方法通过分析兴趣点在各个时段被用户访问的频率，可以得到兴趣点的时间属性. 文献[3]通过分析用户、时间、空间之间的两两相互作用关系，提出了连续兴趣点的时空排列方法；文献[37]根据用户访问兴趣点的时间偏好，提出了时间敏感的兴趣点推荐算法，为用户在某个时段上推荐合适的兴趣点. 第三种是协同过滤和用户行为分析相融合的个性化推荐方法. 例如，文

献[1]提出了一种基于矩阵分解的上下文感知 POI 推荐算法,该方法在矩阵分解模型中融合了用户的签到行为、POI 的地理位置和用户的社会关系等信息,有效解决了数据稀疏性问题,同时具有较高推荐的准确性.由于个性化兴趣点推荐方法需要先在用户历史行为数据上分析行为规律和偏好然后才能进行推荐,如果没有历史数据,推荐方法将不能获取到用户的行为规律和偏好模式,因此冷启动和用户偏好模型是个性化推荐面临的主要问题.

此外,对于基于社会化网络的兴趣点推荐系统,刘树栋等人[38]从分析基于位置的社会化网络的结构特征入手,对基于位置的社会化网络推荐系统的基本框架、基于不同网络层次数据挖掘的推荐算法及应用类型等进行了详细综述,并对未来的研究难点和热点进行了分析和展望.

综上可见,兴趣点推荐方法的研究已经从多方面展开,然而上述推荐方法研究的重点在于如何给用户推荐满意度最高的兴趣点(也就是追求推荐的准确性)和推荐算法的执行效率,却忽略了推荐列表本身的多样性和个性化,这将导致推荐的兴趣点之间可能非常相似而不能扩大用户视野.对于上述问题,本文综合考虑了兴趣点的地理位置和社会联系属性,建立了地理-社会关系模型,用来度量兴趣点之间的地理-社会关系相关度,在此基础上提出了一种新的多样性与个性化兴趣点推荐方法,该方法在充分考虑推荐列表多样性的情况下兼顾了用户个人偏好,目标是拓宽用户视野和提升用户体验.需要指出的是,虽然文献[39]提出了一种兴趣点多样性推荐方法,但本文方法与该文有以下两点不同:(1)聚类标准不同.本文构建了地理-社会关系模型,用来评估兴趣点之间的地理-社会关系相关度,在此基础上对兴趣点进行聚类;而文献[39]仅根据兴趣点自带的类别标签划分兴趣点的类别,没有考虑兴趣点的地理关系和社会关系.(2)推荐目标不同.本文目标是在兴趣点的地理-社会关系基础上进行聚类,然后从每个聚类中找出满足用户偏好的对象并按其对用户偏好的贴近度排序,得到的结果综合体现了多样性与个性化;文献[39]的目标是最大化推荐结果对兴趣点类别的覆盖率和对用户偏好的相关性.

3 总体解决方案

本文的总体解决方案如图 1 所示,共分 3 步.

第 1 步,构建兴趣点的地理-社会关系模型.首先根据兴趣点之间的位置关系和访问兴趣点的用户之间的社会关系,构建兴趣点的地理-社会关系模型,计算兴趣点的地理-社会关系相关度矩阵.

第 2 步,兴趣点聚类.在兴趣点相关度矩阵基础上,采用谱聚类方法对兴趣点进行聚类,使得地理-社会关系相关度较高的兴趣点聚成一类且不同聚类之间具有较低的相关度.

第 3 步,兴趣点多样性选取与个性化排序.采用矩阵分解算法(如概率因子模型、非负矩阵分解、奇异值分解等),预测出每个用户访问各兴趣点的次数,以此评估该用户对各兴趣点的偏好程度,形成用户满意度矩阵;从第 2 步生成的每个聚类中,各选取一个当前用户满意度最高的兴趣点,形成兴趣点推荐列表,这些兴趣点按当前用户满意度进行降序排列,从而得到兼顾多样性与个性化的兴趣点推荐列表.

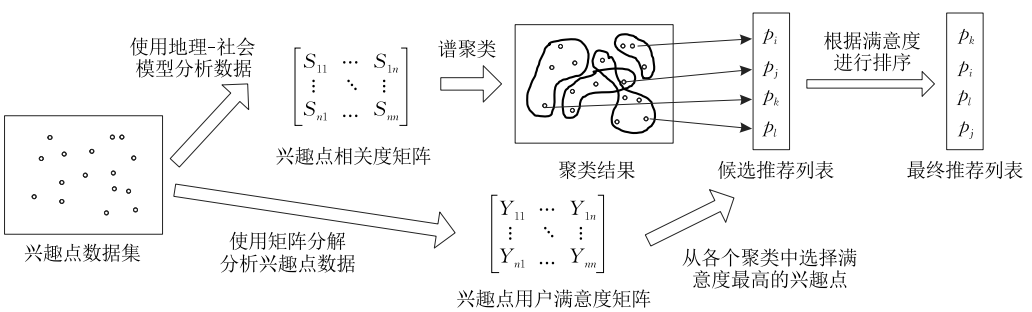


图 1 解决方案总体框架图

4 地理-社会关系模型

本节给出兴趣点的相关定义和地理-社会关系模型.

4.1 相关定义

定义 1. 兴趣点集合.令 $p_i=(p_i.loc, p_i.doc)$ 表示一个兴趣点,其中 $p_i.loc$ 代表 p_i 的位置信息(通常由经纬度表示),$p_i.doc$ 表示 p_i 的文本信息(如 ID、名称、种类、设施等描述信息),n 个不同的兴趣

点构成了集合 $P=\{p_1,p_2,p_3,\cdots,p_n\}$.

定义 2. 用户社会关系网络图. 令 $G=(U,E)$ 代表用户社会关系网络图，其中 U 为顶点(用户)集合，E 为边集合，顶点 $u_i\in U$ 代表一个用户，边 $(u_i,u_j)\in E$ 表示用户 u_i 和 u_j 具有直接朋友关系.

定义 3. 用户签到记录. 用户集合 U 内所有用户的签到记录用 $CK=\{\langle u_i,p_k,t_r\rangle \mid u_i\in U,p_i\in P\}$ 表示，其中 $\langle u_i,p_k,t_r\rangle$ 表示用户 u_i 在时刻 t_r 访问过地点 p_k；对于地点 p_k，访问过 p_k 的用户集合表为 $U_{p_k}=\{u_i \mid \langle u_i,p_k,*\rangle\in CK\}$，其中“$*$”表示任意时间.

4.2 地理-社会关系模型

根据兴趣点之间的位置关系和社会联系，本文在文献[16]构建的地理-社会关系模型基础上，充分考虑了访问兴趣点的用户之间的社会关系紧密度，进一步完善了兴趣点之间的地理-社会关系模型. 该模型中，兴趣点 p_i 和 p_j 之间的地理-社会关系距离计算方法为

$$D_{gs}(p_i,p_j)=\omega\cdot D_P(p_i,p_j)+(1-\omega)\cdot D_S(p_i,p_j) \tag{1}$$

其中，$\omega\in[0,1]$ 为可调参数，用来调节位置距离 D_P 和社会关系距离 D_S 在计算的地理-社会关系距离中所占的比重；$D_P(p_i,p_j)$ 为兴趣点 p_i 和 p_j 之间的位置距离，计算方法为

$$D_P(p_i,p_j)=\frac{E(p_i,p_j)}{maxD} \tag{2}$$

其中，$E(p_i,p_j)$ 为 p_i 和 p_j 之间的欧式距离，$maxD$ 为兴趣点集合 P 中任意两点之间的最大距离.

$D_S(p_i,p_j)$ 为兴趣点 p_i 和 p_j 之间的社会关系距离，计算方法为

$$D_S(p_i,p_j)=1-\frac{CU_{ij}}{|U_{p_i}\cup U_{p_j}|} \tag{3}$$

其中，U_{p_i} 和 U_{p_j} 分别表示访问过兴趣点 p_i 和 p_j 的用户集合；CU_{ij} 表示用户之间的社会关系紧密度，计算方法为

$$CU_{ij}=\begin{cases}\sum\limits_{u_a\in U_a}\dfrac{\sum\limits_{u_b\in U_b}S_{u_a,u_b}}{|U_b|}, & U_b\neq\varnothing\\ 0, & U_b=\varnothing\end{cases} \tag{4}$$

其中，

$$U_a=\{u_a \mid u_a\in\{U_{p_i}\cup U_{p_j}\}\} \tag{5}$$

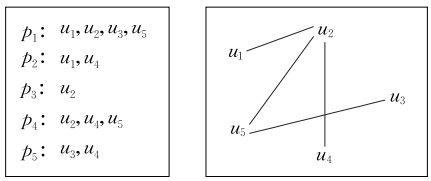
$$U_b=\begin{cases}\{u_b\in U_{p_j} \mid (u_a,u_b)\in E\}, & u_a\in U_{p_i}\\ \{u_b\in U_{p_i} \mid (u_a,u_b)\in E\}, & u_a\in U_{p_j}\end{cases} \tag{6}$$

其中，$(u_a,u_b)\in E$ 指的是 u_a 与 u_b 具有直接朋友关系. 假设一对朋友访问过的地点集合越相似，则该对朋友之间的兴趣偏好越相似，他们之间的社会关系也就越紧密. 因此，用户 u_a 与 u_b 之间社会关系紧密度 S_{u_a,u_b} 定义为

$$S_{u_a,u_b}=\frac{|P_{u_a}\cap P_{u_b}|}{|P_{u_a}\cup P_{u_b}|} \tag{7}$$

在式(7)中，P_{u_a} 和 P_{u_b} 分别表示用户 u_a 和 u_b 访问过的地点集合. 对于一对具有直接朋友关系的用户 u_a 和 u_b，式(4)中 CU_{ij} 的值会加上用户 u_a 和 u_b 的社会关系紧密度(根据式(7)计算)，这与文献[16]仅根据访问两个兴趣点的重叠朋友数来评估社会关系距离的方法不同，从而使得两个地点间的社会关系距离更加合理.

来看一个例子. 图 2(a)和(b)分别给出了用户签到和用户直接朋友关系图.



(a) Check-in records　　(b) User friendship graph

图 2　用户签到记录和直接朋友关系图

假设要计算地点 p_1 和 p_2 的社会关系距离. 如图 2(a)所示，访问过兴趣点 p_1 的用户集合为 $U_{p_1}=\{u_1,u_2,u_3,u_5\}$；如图 2(b)所示，与用户 u_2 具有直接朋友关系的用户有 u_1，u_4 和 u_5. 如果根据文献[16]提出的社会关系距离计算方法，则有 $CU_{12}=\{u_1,u_2,u_4\}$，又因为 $U_{p_1}\cup U_{p_2}=\{u_1,u_2,u_3,u_4,u_5\}$，因此兴趣点 p_1 和 p_2 之间的社会关系距离为：$D_S(p_1,p_2)=1-3/5=0.4$.

采用本文所提的计算兴趣点社会关系的方法(如式(3)～(6))，沿用图 2 的例子，以计算兴趣点 p_1 和 p_2 的社会关系为例，则有 $U_a=U_{p_1}\cup U_{p_2}=\{u_1,u_2,u_3,u_4,u_5\}$. 接下来，计算用户 u_1,u_2,u_3,u_4,u_5 两两之间的社会关系紧密度，结果表 1 所示.

表 1　用户社会关系紧密度矩阵

	u_1	u_2	u_3	u_4	u_5
u_1	1	1/4	1/3	1/4	1/3
u_2	1/4	1	1/4	1/5	2/3
u_3	1/3	1/4	1	1/4	1/3
u_4	1/4	1/5	1/4	1	1/4
u_5	1/3	2/3	1/3	1/4	1

根据图 2 和式(6)，当 $u_a=u_1$ 时，$U_b=\{u_1\}$；当 $u_a=u_2$ 时，$U_b=\{u_1,u_2,u_4\}$；当 $u_a=u_3$ 时，U_b 为空集；当 $u_a=u_4$ 时，$U_b=\{u_2,u_4\}$；当 $u_a=u_5$，U_b 为空集. 因

此，结合表 1 的社会关系紧密度，可得

$$CU_{12}=1/1+(1+1/4+1/5)/3+0+(1+1/5)/2+0=25/12,$$

进而，兴趣点 p_1 和 p_2 的社会关系距离为

$$D_S(p_1,p_2)=1-\frac{25/12}{5}\approx 0.583.$$

根据兴趣点 p_i 和 p_j 之间的地理-社会关系距离(式(1))，可将其转换为 p_i 和 p_j 之间的地理-社会关系相关度，即

$$S(p_i,p_j)=1-D_{gs}(p_i,p_j) \tag{8}$$

通过计算兴趣点集合 P 中任意一对兴趣点之间的地理-社会关系相关度，可得到一个 $n\times n$ 的相关度矩阵 $\boldsymbol{W}$，矩阵中的元素 w_{ij} 表示 p_i 和 p_j 之间的地理-社会关系相关度. 根据兴趣点之间的相关度矩阵，可以得到兴趣点之间的地理-社会关系网络图(简称兴趣点关系网络图)，顶点为兴趣点，边代表地理-社会关系，边的权重代表地理-社会关系相关度.

对兴趣点的地理-社会关系相关度评估算法的时间复杂度分析：假设有 n 个兴趣点，则计算所有兴趣点的位置关系的复杂度为 $O(n^2)$；对于兴趣点之间的社会关系，令访问过每对兴趣点的平均用户数为 m 人，每个用户的平均直接朋友数为 l 个，则计算每对用户社会关系的复杂度为 $O(m^2)$，计算式(6)中 U_b 的复杂度为 $O(ml)$，进而计算所有 n 个兴趣点两两之间的社会关系紧密度的复杂度为 $O((m^2+ml)n^2)$，这也是整个算法的时间复杂度. 虽然复杂度较高，但该部分可在离线阶段定期计算，并不影响在线推荐算法的执行效率.

5 兴趣点聚类划分

本节描述如何在兴趣点关系网络图上使用谱聚类算法对兴趣点进行聚类划分，从而得到多个具有差异性的兴趣点聚类.

5.1 基于谱聚类的兴趣点聚类方法

谱聚类(Spectral clustering)是一种基于图论的聚类方法，主要用于将带权无向图划分为两个或两个以上的最优子图，使得子图内部尽量相似而各个子图之间距离较远. 由于谱聚类仅需图的顶点之间的相似度矩阵(也就是本文的兴趣点之间的地理-社会关系相关度矩阵)，并适合高维空间数据的划分，因此本文采用谱聚类算法对兴趣点进行聚类划分. 谱聚类算法可区分为二路和多路谱聚类算法. 由于二路谱聚类也可实现多个聚类的划分，因此本文使用二路谱聚类算法. 二路谱聚类算法的损失函数将全图划分为两个子图，常见的划分准则包括最小割集准则(Minimum cut)、规范割集准则(Normalized cut)、比例割集准则(Ratio cut)和平均割集准则(Average cut)[40]等.

图 3 是一个兴趣点关系网络图，顶点代表兴趣点，边的权重为兴趣点之间的相关度. 本文用 $\boldsymbol{W}$ 代表兴趣点的相关度矩阵，其中的元素 w_{ij} 代表兴趣点 p_i 和 p_j(即顶点 i 与 j)之间的相关度. 下面结合图 3 描述利用二路谱聚类对兴趣点关系网络图进行划分的过程.

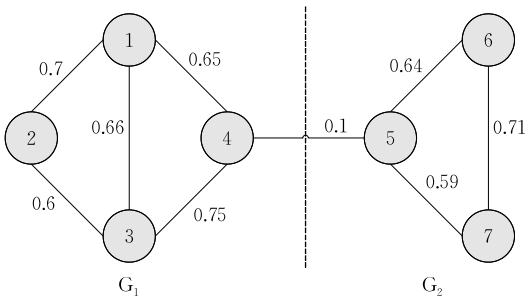


图 3 兴趣点关系网络图实例

采用二路谱聚类划分方法，可将图 3 的无向图分为两个最优子图 G_1 和 G_2，划分结果以 N 维向量(这里 $N=7$，因为有 7 个顶点)$\boldsymbol{q}=[q_1,q_2,\cdots,q_N]$ 表示. 以最小割集划分准则为例，若顶点 i 属于 G_1，则令 $q_i=c_1$；若顶点 i 属于 G_2，则令 $q_i=c_2$，其中 c_1、c_2 用于表示顶点 i 的聚类标签，可设为常数. 据此，将图 3 所示的兴趣点关系网络图 G 划分成两个最优子图 G_1 和 G_2 的划分方案，表示为

$$\boldsymbol{q}=[c_1,c_1,c_1,c_1,c_2,c_2,c_2].$$

按此划分方案，划分最优子图时所截断的兴趣点关系图中边的权重之和的函数，即损失函数可表示为

$$Cut(G_1,G_2)=\sum_{i\in G_1,j\in G_2}w_{ij}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(q_i-q_j)^2}{2(c_1-c_2)^2} \tag{9}$$

又因为

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(q_i-q_j)^2&=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(q_i^2-2q_iq_j+q^2)\\&=-\sum_{i=1}^{n}\sum_{j=1}^{n}2w_{ij}q_iq_j+\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(q_i^2+q^2)\\&=-\sum_{i=1}^{n}\sum_{j=1}^{n}2w_{ij}q_iq_j+\sum_{i=1}^{n}2q_i^2\sum_{j=1}^{n}w_{ij}\\&=2\boldsymbol{q}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{q}\end{aligned} \tag{10}$$

其中，$\boldsymbol{W}$ 为兴趣点的相关度矩阵，$\boldsymbol{D}$ 为对角矩阵，且有 $D_{ii}=\sum_{j=1}^{n}w_{ij}$. 合并算式(9)和式(10)可得

$$Cut(G_1,G_2)=\frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{q}}{(c_1-c_2)^2} \tag{11}$$

其中，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$. 由式(11)可知，当 $\boldsymbol{q}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{q}$ 取得最小值时，该图划分的损失函数取得最小值. 以图 3 为例，可计算得到相似度矩阵 $\boldsymbol{W}$、对角矩阵 $\boldsymbol{D}$ 以及对应的拉普拉斯矩阵 $\boldsymbol{L}$(分别如表 2(a)、(b)、(c)所示).

表 2　谱聚类矩阵示例

(a) 相似度矩阵 $\boldsymbol{W}$

	1	2	3	4	5	6	7
1	0	0.7	0.66	0.65	0	0	0
2	0.7	0	0.6	0	0	0	0
3	0.66	0.6	0	0.75	0	0	0
4	0.65	0	0.75	0	0.1	0	0
5	0	0	0	0.1	0	0.64	0.59
6	0	0	0	0	0.64	0	0.71
7	0	0	0	0	0.59	0.71	0

(b) 对角矩阵 $\boldsymbol{D}$

	1	2	3	4	5	6	7
1	2.01	0	0	0	0	0	0
2	0	1.3	0	0	0	0	0
3	0	0	2.01	0	0	0	0
4	0	0	0	1.5	0	0	0
5	0	0	0	0	1.33	0	0
6	0	0	0	0	0	1.45	0
7	0	0	0	0	0	0	1.3

(c) 拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$

	1	2	3	4	5	6	7
1	2.01	−0.7	−0.66	−0.65	0	0	0
2	−0.7	1.3	−0.6	0	0	0	0
3	−0.66	−0.6	2.01	−0.75	0	0	0
4	−0.65	0	−0.75	1.5	−0.1	0	0
5	0	0	0	−0.1	1.33	−0.64	−0.59
6	0	0	0	0	−0.64	1.45	−0.71
7	0	0	0	0	−0.59	−0.71	1.3

由于采用最小割集划分准则的谱聚类结果容易出现歪斜分割问题，即偏向小区域分割现象，因此本文采用规范割集准则(Normalized cut)对兴趣点进行划分. 沿用图 3 的例子，其损失函数表示为

$$Ncut(G_1,G_2)=Cut(G_1,G_2)\times\left(\frac{1}{d_1}+\frac{1}{d_2}\right) \tag{12}$$

划分方案表示为

$$q_i=\begin{cases}\sqrt{\dfrac{d_1}{d_2 d}}, & i\in G_1\\ -\sqrt{\dfrac{d_2}{d_1 d}}, & i\in G_2\end{cases} \tag{13}$$

其中，d_1、d_2、d 分别表示图 G_1、G_2、G 的权值之和，$Cut(G_1,G_2)$的计算如式(13). 采用规范割集划分准则的聚类过程与上述过程相同，这里不再赘述.

由于损失函数 $Cut(G_1,G_2)$经推导可化为广义瑞利熵的形式[41]. 根据瑞利熵性质，即当 $\boldsymbol{q}$ 为 $\boldsymbol{L}$ 的最小特征值，次小特征值，…，最大特征值对应的特征向量时，可以分别取到 $\boldsymbol{q}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{q}$ 的最小值，次小值，…，最大值，由此可得到满足 $\min(\boldsymbol{q}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{q})$的最佳划分方案. 进而，当需要将兴趣点关系网络图划分为 k 个子图时，则可取前 m 个最小特征值对应的特征向量，组成一个 $m\times n$ 矩阵 $\boldsymbol{R}$，其中第 i 个列向量代表顶点 i，进而利用 k-means 聚类划分可最终得到 k 个兴趣点聚类.

5.2　兴趣点聚类划分例子

以图 3 中的数据样本为例，可以求得矩阵 $\boldsymbol{L}$ 的特征值和对应的特征向量(如表 3 所示). 根据表 3 所示的特征值与特征向量的对应关系，若只取最小特征值(即 0.0096)对应的特征向量作为划分依据，则顶点 1 对应特征向量的第一个元素 0.4493、顶点 2 对应特征向量的第二个元素 0.4524，…，顶点 7 对应特征向量的第 7 个元素 0.2578；若取前两个最小特征值，即(0.0096，0.0775)，对应的特征向量作为划分依据，则顶点 1 对应向量(0.4493，−0.2294)、顶点 2 对应向量(0.4524，−0.2430)，…，顶点 7 对应向量(0.2578，0.5381). 以此类推，可以用前 m 个最小特征值对应的特征向量元素组合来表示每个顶点坐标，进而可采用 k-means 算法进行聚类划分. 例如，若要将图 3 中七个顶点聚为两个聚类，假设取前两个最小特征值对应的特征向量作为划分依据，根据上述算法，则聚类 1 包含顶点{1，2，3，4}，聚类 2 包含顶点{5，6，7}.

表 3　矩阵 $\boldsymbol{L}$ 的特征值与特征向量

特征值	特征向量						
0.0096	0.4493	0.4524	0.4488	0.4397	0.2678	0.2461	0.2578
0.0775	−0.2294	−0.2430	−0.2274	−0.1896	0.4996	0.5114	0.5381
1.3800	−0.0053	−0.7257	0.1030	0.6718	0.0682	−0.0492	−0.0663
1.9046	−0.0318	0.0855	−0.0491	−0.0474	0.7667	−0.1069	−0.6227
2.1138	0.0145	−0.0276	0.0206	0.0065	−0.2884	0.8147	−0.5017
2.6766	0.7687	−0.1170	−0.6284	−0.0243	0.0025	−0.0009	−0.0005
2.7379	−0.3917	0.4335	−0.5819	0.5626	−0.0543	0.0202	0.0123

基于谱聚类的兴趣点聚类算法时间复杂度分析：由于在第 4 节已经预先计算出兴趣点之间的地理-社会关系相关度矩阵 $\boldsymbol{W}$，因此对角矩阵 $\boldsymbol{D}$ 和拉普拉斯矩阵 $\boldsymbol{L}$ 可以在 $\boldsymbol{W}$ 基础上计算得到，进而可以计算出 $\boldsymbol{L}$ 的特征值和特征向量；在此基础上，兴趣点的聚类实质上可以看成是在选取的 m 个最小特征值对应的特征向量矩阵上进行 k-means 聚类，因此算法的时间复杂度是 $O(kntm)$，其中 k 是聚类个数，n 是兴趣点个数，m 是特征值个数（也是每个兴趣点的特征维度），t 是迭代次数.

6 兴趣点多样性选取与个性化排序

利用第 5 节的兴趣点聚类方法，可得到多个具有差异性的兴趣点集合. 下面的任务是如何从每个集合中选出一个最贴近用户偏好的兴趣点.

推测用户偏好的基本思想是，如果用户对某个兴趣点的访问次数越多，则说明用户对该兴趣点的偏好程度越高，那么将该兴趣点推荐给用户的价值也越高. 对此，本文采用矩阵分解类算法，如奇异值分解（Singular Value Decomposition，SVD）[42]、非负矩阵分解（Non-negative Matrix Factorization，NMF）[32]、概率因子模型（Probabilistic Factor Model，PFM）[22]拟合用户-兴趣点访问次数矩阵. 通过对访问次数矩阵的拟合，可选出用户在每个聚类中偏好的兴趣点.

文献[22]将概率因子模型应用到商品推荐，该模型以泊松分布拟合用户对网站的访问次数，再结合矩阵分解进行概率分析，最终得到接近实际的用户-网站访问次数矩阵的拟合矩阵. 与用户访问网站类似，用户访问兴趣点的行为在很大程度上符合以固定平均瞬时速率随机且独立出现的特点，因此本文采用基于泊松分布的概率因子模型拟合用户在单位时间内访问兴趣点的次数矩阵.

如图 4 所示，本文用 $m\times n$ 矩阵 $\boldsymbol{F}$ 表示用户访问兴趣点的次数，其中 m 表示用户数，n 表示兴趣点个数. 例如，$\boldsymbol{F}$ 中的元素 f_{ij} 表示用户 i 访问兴趣点 j 的次数. 假设 f_{ij} 满足以 y_{ij} 为均值的泊松分布，则 y_{ij} 可构成一个与 $\boldsymbol{F}$ 具有相同行列数的 $m\times n$ 矩阵 $\boldsymbol{Y}$，$\boldsymbol{Y}$ 可被分解为一个 $m\times d$ 维的矩阵 $\boldsymbol{U}$ 和一个 $n\times d$ 维的矩阵 $\boldsymbol{V}$，其中 $\boldsymbol{U}$ 中的元素 $u_{ik}(k=1,2,\cdots,d)$ 表示用户 i 对兴趣点隐藏属性（latent feature）k 的偏好程度，$\boldsymbol{V}$ 中的元素 $v_{jk}(k=1,2,\cdots,d)$ 表示兴趣点 j 对隐藏属性 k 的贴近程度. 假设 u_{ik}、v_{jk} 服从 Gamma

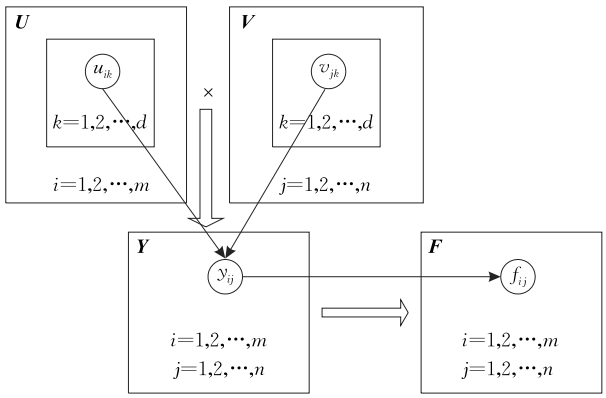


图 4 概率因子模型原理图

先验分布，则

$$p(\boldsymbol{U}|\alpha,\beta)=\prod_{i=1}^{m}\prod_{k=1}^{d}\frac{u_{ik}^{\alpha_k-1}\exp(-u_{ik}/\beta_k)}{\beta_k^{\alpha_k}\Gamma(\alpha_k)} \tag{14}$$

$$p(\boldsymbol{V}|\alpha,\beta)=\prod_{j=1}^{n}\prod_{k=1}^{d}\frac{v_{jk}^{\alpha_k-1}\exp(-v_{jk}/\beta_k)}{\beta_k^{\alpha_k}\Gamma(\alpha_k)} \tag{15}$$

其中，$\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_d\}$，$\beta=\{\beta_1,\beta_2,\cdots,\beta_d\}$，$u_{ik}>0$，$v_{jk}>0$，$\alpha_k>0$，$\beta_k>0$，$\Gamma(\cdot)$ 为 Gamma 函数. 在此基础上，$\boldsymbol{F}$ 的泊松概率分布可表示为

$$p(\boldsymbol{F}|\boldsymbol{Y})=\prod_{i=1}^{m}\prod_{j=1}^{n}\frac{y_{ij}^{f_{ij}}\exp(-y_{ij})}{f_{ij}} \tag{16}$$

其中，$y_{ij}=\sum_{k=1}^{d}u_{ik}v_{jk}$.

由于 $\boldsymbol{Y}=\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$，故在给定条件为 $\boldsymbol{F}$ 时，$\boldsymbol{U}$、$\boldsymbol{V}$ 的后验概率为

$$p(\boldsymbol{U},\boldsymbol{V}|\boldsymbol{F},\alpha,\beta)\propto p(\boldsymbol{F}|\boldsymbol{Y})p(\boldsymbol{U}|\alpha,\beta)p(\boldsymbol{V}|\alpha,\beta) \tag{17}$$

通过求式(17)的最大值，可得到最能拟合 $\boldsymbol{F}$ 的矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$. 为求式(17)的最大值，先取式(16)的对数：

$$\begin{aligned}\boldsymbol{L}(\boldsymbol{U},\boldsymbol{V}|\boldsymbol{F})=&\sum_{i=1}^{m}\sum_{k=1}^{d}((\alpha_k-1)\ln(u_{ik}/\beta_k)-u_{ik}/\beta_k)+\\&\sum_{j=1}^{n}\sum_{k=1}^{d}((\alpha_k-1)\ln(v_{jk}/\beta_k)-v_{jk}/\beta_k)+\\&\sum_{i=1}^{m}\sum_{j=1}^{n}(f_{ij}\ln y_{ij}-y_{ij})+const\end{aligned} \tag{18}$$

然后，对式(18)中的 u_{ik}、v_{jk} 分别求偏导，得：

$$\frac{\partial L}{\partial u_{ik}}=\sum_{j=1}^{n}(f_{ij}v_{jk}/y_{ij}-v_{jk})+(\alpha_k-1)/u_{ik}-1/\beta_k \tag{19}$$

$$\frac{\partial L}{\partial v_{jk}}=\sum_{i=1}^{m}(f_{ij}u_{ik}/y_{ij}-u_{ik})+(\alpha_k-1)/v_{jk}-1/\beta_k \tag{20}$$

接着，分别以 $\dfrac{u_{ik}}{\sum_{j=1}^{n}v_{jk}+1/\beta_k}$、$\dfrac{v_{jk}}{\sum_{i=1}^{m}u_{ik}+1/\beta_k}$ 为步长，采用随机梯度下降（stochastic gradient descent）

法，得到如下的迭代公式：

$$u_{ik} \leftarrow u_{ik} \frac{\sum_{j=1}^{n}(f_{ij}v_{jk}/y_{ij})+(\alpha_k-1)/u_{ik}}{\sum_{j=1}^{n}v_{jk}+1/\beta_k} \tag{21}$$

$$v_{jk} \leftarrow v_{jk} \frac{\sum_{i=1}^{m}(f_{ij}u_{ik}/y_{ij})+(\alpha_k-1)/v_{jk}}{\sum_{i=1}^{m}u_{ik}+1/\beta_k} \tag{22}$$

利用式(21)和(22)进行迭代，最终可得到拟合矩阵 $\boldsymbol{Y}=\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$，用来预测用户访问兴趣点的次数.

在此基础上，对于一个给定用户，可根据矩阵 $\boldsymbol{Y}$ 得到该用户的兴趣偏好，进而从第 5 节得到的每个兴趣点聚类中分别选取一个该用户访问次数最多的兴趣点，再将这些被选取的兴趣点按其被用户访问的次数进行降序排列，最终得到满足该用户的多样性与个性化需求的兴趣点推荐列表.

基于概率因子模型的兴趣点选取算法的时间复杂度分析：对于每个兴趣点聚类，需要从中选取一个最满足给定用户偏好的兴趣点(也就是给定用户访问次数最多的兴趣点)，假设每个聚类平均有 n 个兴趣点，有 m 个访问过这些兴趣点的用户，隐藏属性的个数为 d，对于兴趣点的选取实际上是要计算出每个聚类中用户访问兴趣点的次数矩阵 $\boldsymbol{Y}_{m\times n}$；由于采用的是随机梯度下降，假设迭代次数为 t，则求矩阵 $\boldsymbol{Y}$ 的时间复杂度为 $O(mnd+mdt+ndt)$，又因为有 k 个聚类，因此算法最终的时间复杂度为 $O(k(mnd+mdt+ndt))$.

7　效果与性能实验评价

7.1　实验数据

实验采用从 Stanford Large Network Dataset Collection 获得的 Gowalla 2009 年 2 月到 2010 年 10 月的用户签到数据，由用户社交关系网络图和用户签到记录的时空数据组成. 其中，社会关系网络图中的顶点代表用户，边代表用户之间的朋友关系，该数据集涵盖了 196591 个顶点以及 950327 条边；签到记录涵盖了 2009 年 2 月份到 2010 年 10 月份共计 6442890 条记录，主要由用户 id、签到时间、签到地点的经纬度以及签到地点 id 组成. 为了降低数据稀疏问题和发现社会关系紧密的兴趣点，本文的实验数据截取其中两部分数据集，分别是位于美国芝加哥市和日本东京市范围内，删除其中签到次数少于 5 次的用户以及被访问次数少于 5 次的兴趣点. 预处理后的实验数据集特征如表 4 所示.

表 4　实验数据集特征

城市	经度	纬度	签到记录数	兴趣点数	用户数
芝加哥	[−88.04, −87.50]	[41.68, 41.98]	41742	1078	739
东京	[139.43, 139.94]	[35.63, 35.74]	100952	3922	1405

7.2　兴趣点的地理-社会关系评估模型效果实验

7.2.1　聚类效果对比

该实验将本文提出的基于地理-社会关系模型的聚类算法(简称 DCPGS-G-New)与文献[16]提出的基于地理-社会关系模型的聚类算法 DCPGS-G 进行聚类效果对比. 二者区别在于兴趣点之间的社会距离计算方法不同. 给定一对兴趣点 p_i 与 p_j，DCPGS-G 计算社会距离 $D_S(p_i,p_j)$ 的方法为

$$D_S(p_i,p_j)=1-\frac{|CU_{ij}|}{|U_{p_i}\cup U_{p_j}|} \tag{23}$$

其中，

$$CU_{ij}=\{u_a\in U_{p_i}\mid u_a\in U_{p_j}\ \text{or}\ u_b\in U_{p_j},(u_a,u_b)\in E\}\cup\{u_a\in U_{p_j}\mid u_a\in U_{p_i}\ \text{or}\ u_b\in U_{p_i},(u_a,u_b)\in E\}.$$

U_{p_i} 和 U_{p_j} 分别表示访问过 p_i 和 p_j 的用户集合.

为了便于聚类效果对比，本文也使用了文献[16]提出的基于密度的聚类方法，该方法需要计算给定兴趣点 p_i 的地理-社会邻域，即 $N_\varepsilon(p_i)$，表示 p_i 的地理-社会距离的 ε-邻域，该邻域包括了所有满足条件的地点 p_j，使得 $D_{gs}(p_i,p_j)\le\varepsilon$，$D_S(p_i,p_j)\le\tau$，$E(p_i,p_j)\le maxD$，其中 ε 是兴趣点之间的地理-社会距离阈值、τ 是社会距离阈值、$maxD$ 是地理距离阈值，三者均为可调参数. 在聚类过程中，先将要进行聚类的地点放入每小格边长为 $maxD/\sqrt{2}$ 的网格中(如图 5)，这样每个方格的对角线长度为 $maxD$，

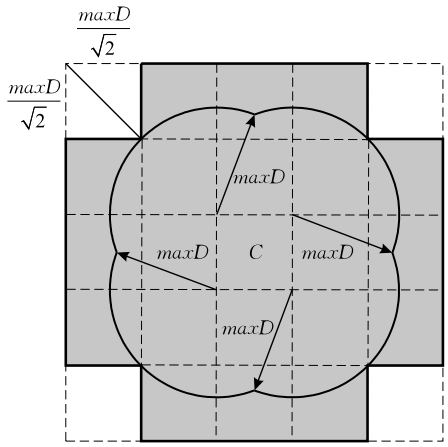
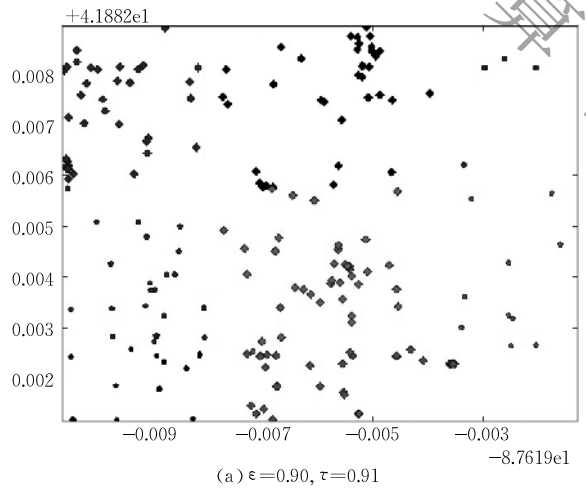


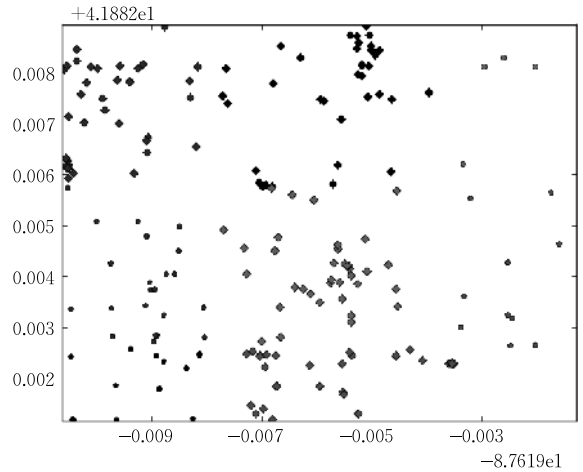
图 5　兴趣点网格划分图

这也是落入同一方格内两个地点的最大距离;然后,求出该小格及该小格的临近格中每个地点的ε-邻域内的点;最后,将每个ε-邻域内的地点数目大于阈值 $MinPts$ 的地点聚成一类.

该实验使用芝加哥数据集,实验参数 $maxD$ 的取值范围为[0,500 m],式(1)中的 ω(地理-社会距离权重)、ε、τ 的取值范围为[0,1],$MinPts$ 的取值范围为[0,20].确定各参数最佳取值的策略是:先以各参数取值范围的十分之一为步长设定各参数值,然后进行实验,通过比较实验结果确定各参数的最佳取值区间,接着再进行细化实验.例如,对于参数 ω,先以步长 0.1 确定取值,即{0,0.1,…,0.9,1},通过对比不同取值下的实验结果,得知 ω 的值在区间[0.6,0.8]时的实验效果最好,最后将该区间进行十等分,进一步比较实验结果从而确定 ω 的最佳取值.图 6 和图 7 分别给出了 DCPGS-G 和 DCPGS-G-NEW 算法在 $maxD=350$,$\omega=0.7$,$\varepsilon=0.90$,$\tau=0.91$,$MinPts=5$ 情况下,参数 ε 和 τ 以 0.01 为步长逐渐递增所得到的聚类结果变化情况.

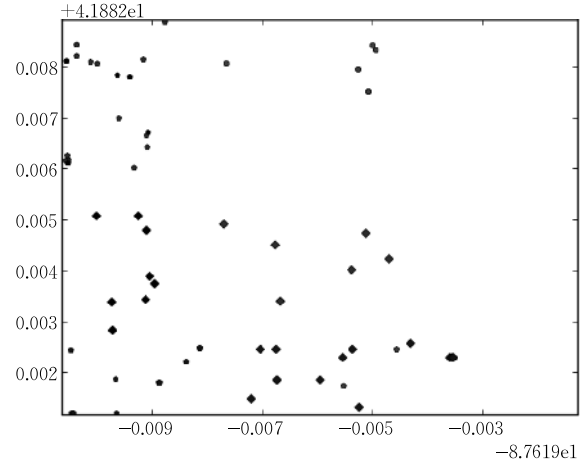


(a) ε=0.90, τ=0.91

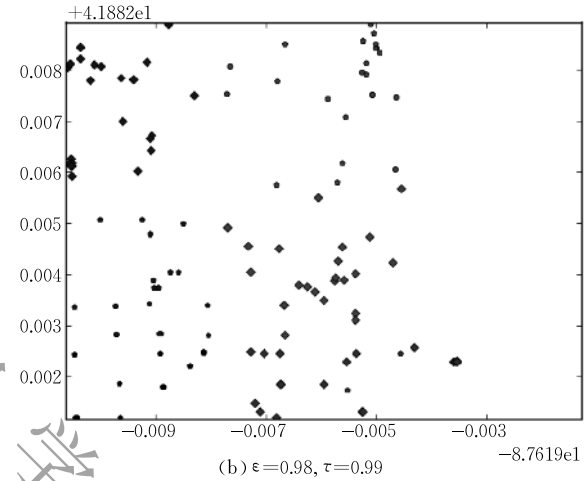


(b) ε=0.98, τ=0.99

图 6 DCPGS-G 聚类结果随参数 ε 和 τ 的变化图



(a) ε=0.90, τ=0.91



(b) ε=0.98, τ=0.99

图 7 DCPGS-G-NEW 聚类结果随参数 ε 和 τ 的变化图

从图 6 和图 7 可以看出,DCPGS-G 随着参数变化,其聚类结果几乎不变,而 DCPGS-G-NEW 的聚类结果则随着参数改变而有着明显改变(用上述实验参数,在东京数据集上也得到了类似结果).原因是,DCPGS-G 的地理-社会关系模型在计算兴趣点社会距离时,仅统计了访问兴趣点的用户群体重叠度,没有进一步评估用户之间社会关系紧密度;而 DCPGS-G-NEW 的地理-社会关系模型通过进一步评估访问兴趣点的用户之间社会关系紧密度,使得聚类结果对参数变化较为敏感,也使聚类结果更加细致可控,从而也体现出用户社会关系紧密度对于兴趣点聚类具有较为重要的影响.

7.2.2 聚类结果的社会性评估

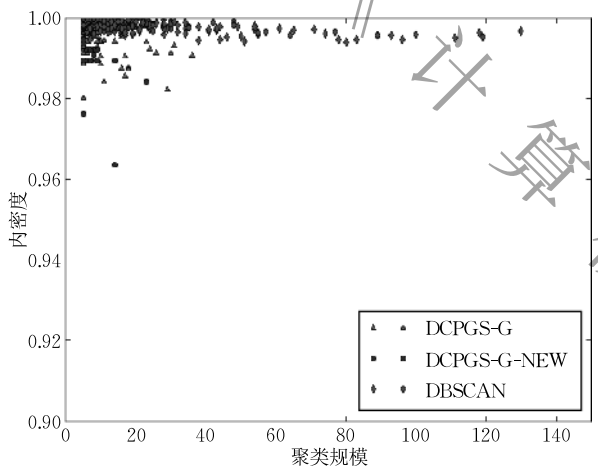
本文采用文献[43]提出的内密度(Internal density)和传导率(Conductance)对聚类内部地点的社会性(Social quality)进行衡量,从而进一步验证本文的地理-社会关系模型的有效性.其中,内密度定义如下:

$$I=1-m_{U_{pc}}/((|U_{pc}|(|U_{pc}|-1)/2) \quad (24)$$

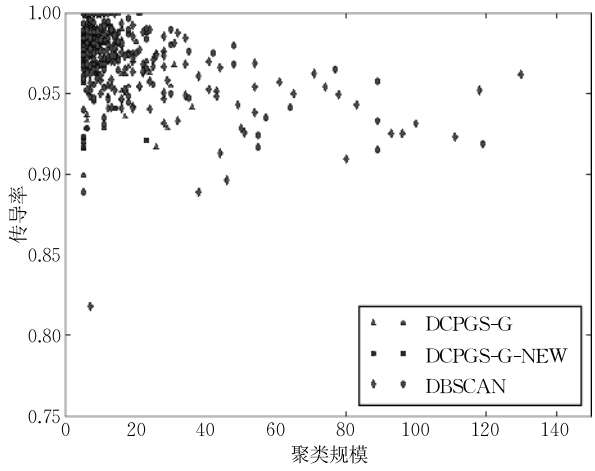
其中，U_{pc} 为访问过聚类 pc 中兴趣点的用户集合，$m_{U_{pc}}=|\{(u,v)|u\in U_{pc},v\in U_{pc}\}|$，$(u,v)$ 表示用户 u、v 之间具有直接朋友关系，访问过聚类 pc 中兴趣点的用户集合之间联系越紧密 $m_{U_{pc}}$ 的值越大. 传导率定义如下：

$$T=OU_{pc}/(2m_{U_{pc}}+OU_{pc}) \tag{25}$$

其中，$OU_{pc}=|\{(u,v)|u\in U_{pc},v\notin U_{pc}\}|$，访问过聚类 pc 中兴趣点的用户集合之间的联系越不紧密则 OU_{pc} 的值越大. 根据式(24)和式(25)可知，聚类结果的内密度和传导率越低，属于同一聚类的地点之间的社会关系越紧密，说明聚类结果的社会性越好. 图 8 给出了基于地理距离的 DBSCAN 聚类(作为对比的基准)、文献[16]的 DCPGS-G 聚类以及本文的 DCPGS-G-NEW 聚类的内密度和传导率对比.



(a) 内密度



(b) 传导率

图 8　不同聚类方法得到的聚类结果的社会性对比图

图 8 给出了在 $maxD=350$，$\omega=0.7$，$\varepsilon=0.90$，$\tau=0.91$，$MinPts=5$ 情况下，各算法得到的聚类结果随聚类规模增大而导致的内密度和传导率的变化情况. 从图中可以看出，DBSCAN 聚类结果的内密度和传导率最高，DCPGS-G 次之，DCPGS-G-NEW 最低；也就是说，DCPGS-G-NEW 聚类结果的社会性最好，DCPGS-G 稍弱，DBSCAN 最弱，这说明本文提出的地理-社会关系模型用于兴趣点的聚类更为有效.

7.3　推荐效果实验

该实验目的是将本文提出的推荐方法与各类主流推荐方法的推荐效果进行对比，采取的对比策略如下：分别采用推荐系统中常用的概率因子模型(PFM)、奇异值分解算法(SVD)、非负矩阵分解算法(NMF)实现兴趣点的推荐，并进一步在 PFM、SVD、NMF 模型上分别引入本文提出的多样性分析方法得到兴趣点多样性推荐模型 DPFM(本文提出的推荐模型)、DSVD、DNMF，即先利用谱聚类方法在兴趣点相关度矩阵上进行兴趣点聚类划分，然后再分别利用 PFM、SVD、NMF 模型拟合用户-兴趣点访问矩阵，最终从每个聚类中选出一个最贴近用户偏好的兴趣点，得到推荐列表. 在此基础上，对 PFM、SVD、NMF、DPFM，DSVD、DNMF 推荐模型进行效果与性能方面的对比.

效果评价标准：分别采用多样性和准确率指标对算法的效果进行评价. 多样性指标的度量方法：

$$Div_{L_{\text{rec}}}=\frac{\sum_{p_i\in L_{\text{rec}}}\sum_{p_j\in L_{\text{rec}}}D_{gs}(p_i,p_j)}{|L_{\text{rec}}|^2} \tag{26}$$

其中，p_i、p_j 代表兴趣点，L_{rec} 表示由推荐算法得到的兴趣点推荐列表，$D_{gs}(p_i,p_j)$ 表示 p_i 和 p_j 的地理-社会关系距离(可由式(1)计算得到). $Div_{L_{\text{rec}}}$ 的值越高，表示推荐列表 L_{rec} 的多样性程度越高.

准确率 $Precision@k$ 的计算方法：

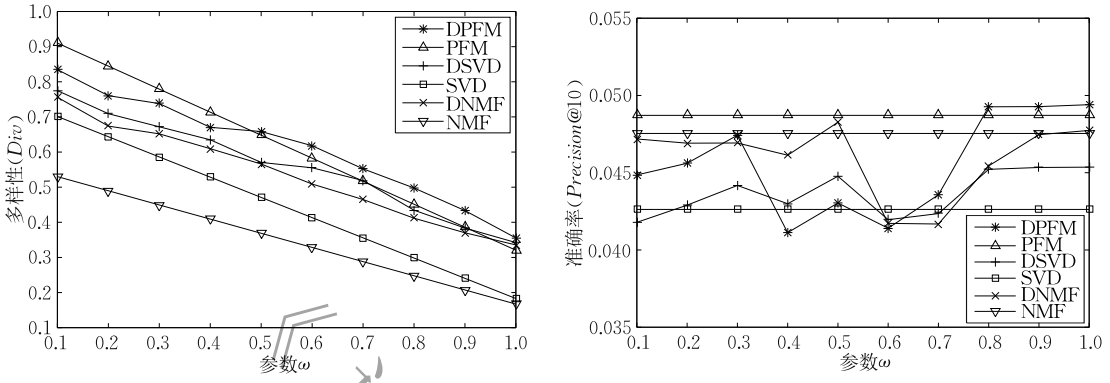
$$Precision@k=\frac{|L_{\text{rec}}\cap L_{\text{test}}|}{k} \tag{27}$$

其中，L_{test} 表示准确的兴趣点列表，是从测试集中选出的用户访问次数最多的前 k 个兴趣点组成. 本实验从数据集中选取 10 个用户进行测试，得到的多样性和准确率是这 10 个用户的平均值.

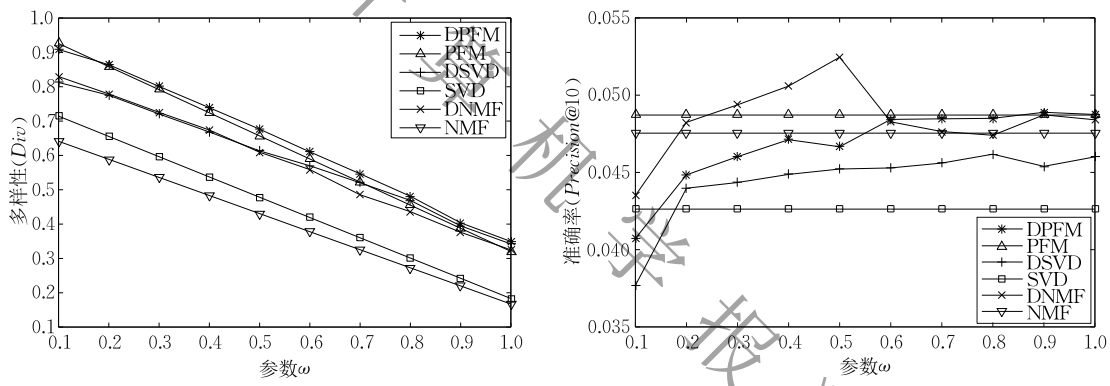
7.3.1　参数变化对各推荐算法效果的影响分析

以芝加哥数据集为例，首先计算出所有兴趣点之间的地理距离和社会关系距离，并将这些距离值存储在数据文件中. 本文推荐算法主要有两个参数，一个是式(1)中的 ω；另一个是第 5 节中的 m，指在谱聚类选取式(11)中矩阵 $\boldsymbol{L}$ 的最小特征值个数，也就是图 3 中每个顶点坐标的分量数. 此外，本实验设

定推荐结果列表中包含的兴趣点个数为 10，相应的谱聚类划分的子图个数也为 10；PFM 模型中的参数根据文献[32]中的最优参数设定，即 $\alpha_k=20$，$\beta_k=0.2$，该参数经测试，同样适用于本文所用的数据集. 由于谱聚类过程使用了 k-means 聚类，其初始聚类点的随机选取可能导致聚类结果不一致，为克服该问题，本文实验结果都为重复 10 次实验所得结果的平均值. 接下来讨论 ω、m 的选取对推荐效果的影响. 固定 $m=2$，从整个数据中随机选取 50% 数据集作为矩阵分解算法的训练集，剩下的作为测试集. 图 9(a)和(b)分别给出了地理-社会关系模型中未考虑(和考虑)朋友之间社会关系紧密度情况下各推荐算法的推荐结果多样性和准确率随参数 ω 的变化情况.



(a) 未考虑朋友之间社会关系紧密度情况下推荐结果的多样性和准确率随参数ω的变化情况



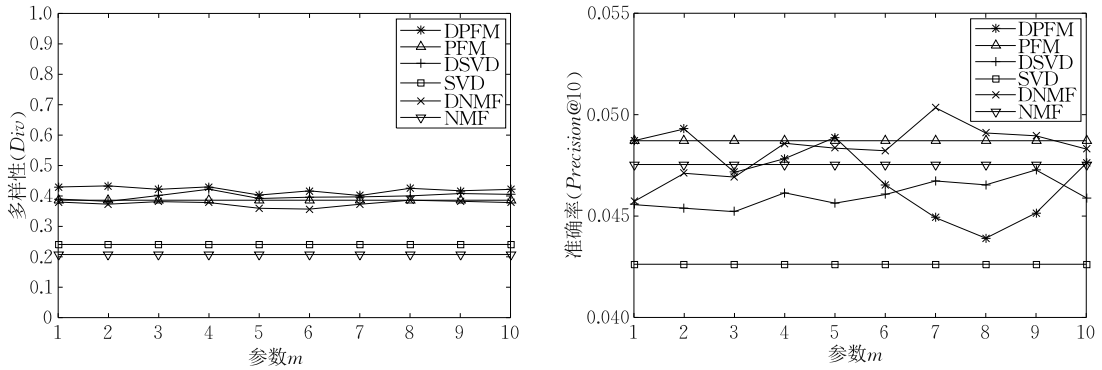
(b) 考虑朋友之间社会关系紧密度情况下推荐结果的多样性和准确率随参数ω的变化情况

图 9 不同情况下推荐结果的多样性和准确率随参数 ω 的变化情况

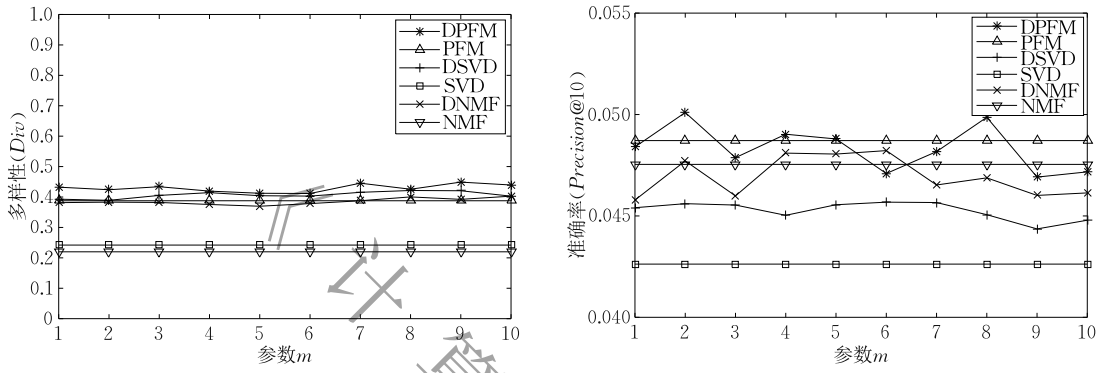
从图 9 可知，在多样性方面，随着 ω 的增大，各推荐算法得到的推荐列表的多样性均呈下降趋势. 在准确率方面，由于 PFM、SVD、NMF 不具备多样性推荐功能，因此这些算法的准确率不受 ω 的影响；具有多样性推荐功能的算法(DPFM、DSVD、DNMF)，地理-社会关系模型中在考虑朋友社会关系紧密度情况下(如图 9(b))，推荐列表的准确率大体上随着 ω 增加而呈上升趋势，并且 DPFM 算法在 $\omega=0.9$ 时取得最高准确率.

接下来，固定 $\omega=0.9$，测试参数 m 的变化对各算法推荐效果的影响，实验结果如图 10 所示.

图 10 给出了当 $\omega=0.9$ 时参数 m 的变化对各推荐算法得到的推荐结果的多样性和准确率的影响. 从图 10 可以看出，不具备多样性推荐功能的算法(PFM、SVD、NMF)，其推荐结果的多样性和准确率不受参数 m 的影响；具备多样性推荐功能的算法(DPFM、DSVD、DNMF)，其推荐结果的多样性和准确率随着参数 m 的变化而起伏不定. 需要指出的是，地理-社会关系模型中在考虑朋友之间社会关系紧密度情况下，DPFM 的推荐结果准确率要高于未考虑朋友之间社会关系紧密度情况下的准确率. 从图 10(b)可知，当 $m=9$ 时 DPFM 推荐结果的多样性最高，当 $m=6$ 时 DPFM 推荐结果的多样性最差；当 $m=2$ 时，DPFM 推荐结果的准确率最高，而当 $m=9$ 时 DPFM 推荐结果的准确率最低；综合来讲，当 $m=2$ 时，DPFM 的推荐结果都达到了较高的多样性和准确率，这可作为实际应用中参数选取的一个参考.



(a) 未考虑朋友之间社会关系紧密度情况下推荐结果的多样性和准确率随参数m的变化情况



(b) 考虑朋友之间社会关系紧密度情况下推荐结果的多样性和准确率随参数m的变化情况

图 10　不同情况下推荐结果的多样性和准确率随参数 m 的变化情况

7.3.2　DPFM 算法在不同训练集下的推荐效果评价

为更好验证本文提出的 DPFM 算法的推荐效果，设定 $\omega=0.9$ 和 $m=2$(实验 7.3.1 表明 DPFM 算法在该参数下具有最好效果)，设定推荐结果列表中包含的兴趣点个数为 10(即 $k=10$)，并分别以 10%，20%，…，90% 的数据集作为 PFM、SVD、NMF 算法的训练集，剩下的 90%，80%，…，10%的数据作为测试集，在此基础上对 DPFM、PFM、DSVD、SVD、DNMF、NMF 所得到的推荐列表的多样性和准确率进行了对比实验.

表 5 给出了 DPFM、PFM、DSVD、SVD、DNMF、NMF 在不同数据集和不同比例的训练集、测试集下的推荐结果的多样性评估结果(每列结果的前者为芝加哥数据集、后者为东京数据集). 从表 5 数据可知 DPFM 推荐结果的平均多样性程度最高，在芝加哥和东京数据集上的平均多样性分别为 42.41% 和 75.76%，且具有多样性推荐功能算法的多样性普遍优于不具有多样性推荐功能的算法的多样性.

表 5　各算法在不同数据集上的推荐结果的多样性评估

训练集占比/%	DPFM	PFM	DSVD	SVD	DNMF	NMF
10	**0.4234/0.7569**	0.3922/0.7421	0.2546/0.6983	0.1655/0.6584	0.2535/0.7218	0.1395/0.6661
20	**0.4347/0.7563**	0.3920/0.7428	0.3592/0.6818	0.2239/0.6368	0.3384/0.7197	0.2050/0.6710
30	**0.4250/0.7575**	0.3929/0.7431	0.3607/0.6938	0.3081/0.6491	0.3623/0.6889	0.2477/0.6493
40	**0.4220/0.7596**	0.3915/0.7439	0.3575/0.6694	0.3216/0.6212	0.3698/0.7006	0.1889/0.6415
50	**0.4240/0.7555**	0.3878/0.7447	0.3891/0.6538	0.2421/0.6184	0.3832/0.6965	0.2198/0.6323
60	**0.4197/0.7578**	0.3889/0.7439	0.3654/0.6424	0.2521/0.6010	0.3888/0.6740	0.2159/0.6125
70	**0.4212/0.7577**	0.3913/0.7457	0.3555/0.6243	0.3359/0.6054	0.3759/0.6113	0.2267/0.5634
80	**0.4251/0.7521**	0.3920/0.7367	0.3511/0.5767	0.2239/0.5527	0.3343/0.6116	0.2050/0.5517
90	**0.4220/0.7654**	0.3922/0.7329	0.2444/0.4931	0.1655/0.4650	0.2521/0.4508	0.1395/0.4104
平均值	**0.4241/0.7576**	0.3912/0.7417	0.3375/0.6371	0.2487/0.6009	0.3398/0.6528	0.1987/0.5998

表 6 给出了 DPFM、PFM、DSVD、SVD、DNMF、NMF 在不同数据集和不同比例的训练集、测试集下推荐结果的准确率评估结果(每列结果的前者为芝加哥数据集、后者为东京数据集). 从实验结果可以发现，具有多样性推荐功能算法的准确率在绝大多数情况下优于不具有多样性推荐功能的算法的准

确率,并且本文提出的 DPFM 推荐结果的平均准确率最高,在芝加哥和东京数据集上的平均准确率分别为 5.06%和 2.8%.但需要指出的是,各算法在两个数据集上的推荐结果准确率都普遍偏低,主要原因是数据稀疏问题导致(比如有些用户访问的兴趣点个数还不足 10 个),本文下一步研究工作将结合用户和兴趣点的多种关联信息,采用 Embedding 和深度学习技术,解决数据稀疏问题.

表 6 各算法推荐结果的准确率评估(*Precision*@10)

训练集占比/%	DPFM	PFM	DSVD	SVD	DNMF	NMF
10	**0.0537/0.0404**	0.0502/0.0314	0.0375/0.0234	0.0342/0.0234	0.0438/0.0344	0.0421/0.0295
20	**0.0528/0.0397**	0.0502/0.0302	0.0444/0.0225	0.0401/0.0227	0.0481/0.0352	0.04899/0.0264
30	**0.0533/0.0352**	0.0505/0.0226	0.0412/0.0231	0.0440/0.0229	0.0501/0.0312	0.0481/0.0241
40	**0.0500/0.0324**	0.0484/0.0248	0.0474/0.0326	0.0445/0.0223	0.0478/0.0305	0.0475/0.0275
50	**0.0501/0.0296**	0.0487/0.0235	0.0456/0.0155	0.0426/0.0162	0.0477/0.0263	0.0475/0.0199
60	**0.0445/0.0246**	0.0451/0.0181	0.0426/0.0223	0.0394/0.0220	0.0461/0.0244	0.0455/0.0224
70	**0.0446/0.0229**	0.0447/0.0165	0.0407/0.0131	0.0383/0.0125	0.0431/0.0229	0.0441/0.0174
80	**0.0532/0.0159**	0.0502/0.0115	0.0441/0.0254	0.0401/0.0243	0.0481/0.0234	0.0490/0.0213
90	**0.0536/0.0111**	0.0502/0.0086	0.0381/0.0122	0.0342/0.0112	0.0436/0.0118	0.0421/0.0091
平均值	**0.0506/0.0280**	0.0487/0.0212	0.0424/0.0211	0.0397/0.0197	0.0465/0.0267	0.0461/0.0220

总体来讲,通过表 5 和表 6 的对比结果可知:(1)本文提出的 DPFM 算法得到的推荐列表的多样性普遍优于 PFM、DSVD、SVD、DNMF、NMF 算法;(2)具备多样性分析功能的 DPFM、DSVD、DNMF 算法的推荐结果多样性普遍高于相应的经典 PFM、SVD、NMF 算法,这反映出本文提出的多样性分析解决方案在提高推荐结果的多样性方面具有优势;(3)从表 6 可知,DPFM 的推荐准确率普遍高于 PFM、DSVD、SVD、DNMF、NMF 算法.此外,在绝大部分情况下,在本文框架下拓展的 DSVD 和 DNMF 推荐结果的准确率都高于相应的 SVD 和 NMF 算法,说明了本文提出的推荐系统框架在提高推荐结果的准确性方面也具备一定的优越性.

7.3.3 DPFM 算法的执行时间测试

该实验测试了在芝加哥和东京数据集上,DPFM 算法在不同推荐结果数(即 k 值)下的在线执行效率(如图 11 所示).从图 11 可以看出,DPFM 算法随着数据量和 k 值的增加,执行时间大致呈线性增长趋势,这也与推荐算法的计算复杂度分析相一致.

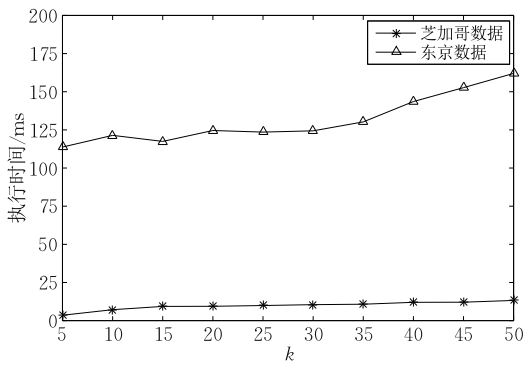


图 11 DPFM 算法在不同数据集和 k 值下的执行效率

8 结束语

本文通过综合考虑空间兴趣点之间的位置关系和社会关系,构建了兴趣点的地理-社会关系模型,提出了基于谱聚类的兴趣点集合多样性划分方法.在现有矩阵分解模型基础上,构建了一种新型的空间兴趣点多样性与个性化推荐解决方案并提出了具体实现方法.

实验结果和分析表明,本文方法提出的地理-社会关系评估模型能够更好地量化兴趣点之间的地理-社会关系相关度,得到的聚类结果更合理.在此基础上,推荐结果不仅比现有经典的矩阵分解算法具有更高的多样性程度,也具备了更高的准确率,能够同时较好地满足用户的多样性与个性化需求.此外,本文提出的算法也具备良好的可扩展性.在接下来的工作中,一方面将尝试综合更多的兴趣点属性信息进行多样性分析,如兴趣点的用户评论、兴趣点的消费水平、兴趣点的周边路况等,从而使推荐列表在更高维度上具有多样性;另一方面还要根据用户和兴趣点的多方面关联信息,利用 Embedding 和深度学习技术挖掘用户和兴趣点之间的深层关联信息,解决数据稀疏和冷启动问题,进一步提升推荐的准确性.

参 考 文 献

[1] Peng Hong-Wei, Jin Yuan-Yuan, Lv Xiao-Qiang, et al. Context-aware POI recommendation based on matrix factorization. Chinese Journal of Computers, 2019, 42(8): 1797-1811(in Chinese)
(彭宏伟, 靳远远, 吕晓强等. 一种基于矩阵分解的上下文感知 POI 推荐算法. 计算机学报, 2019, 42(8): 1797-1811)

[2] Liao Guo-Qiong, Jiang Shan, Zhou Zhi-Heng, et al. Dual fine-granularity POI recommendation on location-based social networks. Chinese Journal of Computers, 2017, 54(11): 2600-2610(in Chinese)
(廖国琼，姜珊，周志恒等. 基于位置社会网络的双重细粒度兴趣点推荐. 计算机研究与发展，2017，54(11)：2600-2610)

[3] Zhao S L, Zhao T, Yang H Q. STELLAR: Spatial-temporal latent ranking for successive point-of-interest recommendation //Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 315-322

[4] Jiang D, Leung W T, Vosecky J. Personalized query suggestion with diversity awareness//Proceedings of the 30th International Conference on Data Engineering. Chicago, USA, 2014: 400-411

[5] Meng X F, Cao L B, Zhang X Y, et al. Top-*k* coupled keyword recommendation for relational keyword queries. Knowledge and Information Systems, 2017, 50(3): 883-916

[6] Meng X F, Tang Y H, Zhang X Y. DP-POIRS: A diversified and personalized point-of-interest recommendation system// Proceedings of the IEEE International Conference on Data Science and Advanced Analytics. Tokyo, Japan, 2017: 332-333

[7] Cong G, Jensen C S, Wu D M. Efficient retrieval of the top-*k* most relevant spatial web objects. PVLDB, 2009, 2(1): 337-348

[8] Qi J H, Zhang R, Jensen C S, et al. Continuous spatial query processing: A survey of safe region based techniques. ACM Computing Surveys, 2018, 51(3): 1-39

[9] Cong G, Jensen C S. Querying geo-textual data: Spatial keyword queries and beyond//Proceedings of the ACM SIGMOD International Conference on Management of Data. San Francisco, USA, 2016: 2207-2212

[10] Hua W, Wang Z, Wang H, et al. Short text understanding through lexical-semantic analysis//Proceedings of the 31st International Conference on Data Engineering. Seoul, South Korea, 2015: 495-506

[11] Wang Zhong-Yuan, Cheng Jian-Peng, Wang Hai-Xun, et al. Short text understanding: A survey. Journal of Computer Research and Development, 2016, 53(2): 262-269(in Chinese)
(王仲远，程健鹏，王海勋等. 短文本理解研究. 计算机研究与发展，2016，53(2)：262-269)

[12] Ren Xing-Yi, Song Mei-Na, Song Jun-De. Point-of-interest recommendation based on the user check-in behavior. Chinese Journal of Computers, 2017, 40(1): 28-50(in Chinese)
(任星怡，宋美娜，宋俊德. 基于用户签到行为的兴趣点推荐. 计算机学报，2017，40(1)：28-50)

[13] Zhu Jing-Hua, Ming Qian. POI recommendation by incorporating trust-distrust relationship in LBSN. Journal of Communications, 2017, 39(7): 157-165(in Chinese)
(朱敬华，明骞. LBSN 中融合信任与不信任关系的兴趣点推荐. 通信学报，2017，39(7)：157-165)

[14] Wang D, Pedreschi D, Song C. Human mobility, social ties, and link prediction//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 1100-1108

[15] Wang H, Terrovitis M, Mamoulis N. Location recommendation in location-based social networks using user check-in data// Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Orlando, USA, 2013: 364-373

[16] Shi J M, Mamoulis N, Wu D M. Density-based place clustering in geo-social networks//Proceedings of the ACM SIGMOD Conference on Management of Data. Snowbird, USA, 2014: 99-110

[17] Ye M, Yin P, Lee W C. Exploiting geographical influence for collaborative point-of-interest recommendation//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 325-334

[18] Yang B, Lei Y, Liu D Y. Social collaborative filtering by trust//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2747-2753

[19] Li H Y, Ge Y, Hong R C, et al. Point-of-interest recommendations: Learning potential check-ins from friends// Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 975-984

[20] Cheng C, Yang H, King I, et al. Fused matrix factorization with geographical and social influence in location-based social networks//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 48-48

[21] Salakhutdinov R, Mnih A. Probabilistic matrix factorization //Proceedings of the International Conference on Neural Information Processing Systems. Kitakyushu, Japan, 2007: 1257-1264

[22] Ma H, Liu C, King I. Probabilistic factor models for Web site recommendation//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 265-274

[23] Luo X, Zhou M C, Xia Y N. An efficient non-negative matrix-factorization-based approach to collaborative-filtering for recommender systems. IEEE Transactions on Industrial Informatics, 2014, 10(2): 1273-1284

[24] Luo X, Zhou M C, Leung H, et al. An incremental-and-static-combined scheme for matrix-factorization-based collaborative filtering. IEEE Transactions on Automation Science and Engineering, 2016, 13(1): 333-343

[25] Luo X, Zhou M C, Li S, et al. An efficient second-order approach to factorizing sparse matrices in recommender systems. IEEE Transactions on Industrial Informatics, 2015, 11(4): 946-956

[26] Ng R T, Han J. Efficient and effective clustering methods for spatial data mining//Proceedings of the International Conference on Very Large Data Bases. Santiago de Chile, Chile, 1994: 144-155

[27] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases//Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. Quebec, Canada, 1996: 103-114

[28] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise//Proceedings of the International Conference on Knowledge Discovery and Data Mining. Portland, USA, 1996: 226-231

[29] Karypis G, Han E H, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. Computer, 1999, 32(8): 68-75

[30] Zhao Y L, Nie L, Wang X, et al. Personalized recommendations of locally interesting venues to tourists via cross-region community matching. ACM Transactions on Intelligent Systems & Technology, 2014, 5(3): 50

[31] Yin H Z, Cui B, Zhou X F, et al. Joint modeling of user check-in behaviors for real-time point-of-interest recommendation. ACM Transactions on Information Systems, 2016, 35(2): 1-44

[32] Hernando A, Bobadilla J, Ortega F. A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. Knowledge-Based Systems, 2016, 97: 188-202

[33] Ye M, Yin P, Lee W C, et al. Exploiting geographical influence for collaborative point-of-interest recommendation// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 325-334

[34] He J, Li X, Liao L, et al. Category-aware next point-of-interest recommendation via listwise Bayesian personalized ranking//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 1837-1843

[35] Li H, Ge Y, Lian D F, et al. Learning user's intrinsic and extrinsic interests for point-of-interest recommendation: A unified approach//Proceedings of the International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 2117-2123

[36] Liu X, Liu Y, Aberer K. Personalized point-of-interest recommendation by mining users' preference transition// Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. San Francisco, USA, 2013: 733-738

[37] Li X, Jiang M M, Hong H T, et al. A time-aware personalized point-of-interest recommendation via high-order tensor factorization. ACM Transactions on Information Systems, 2017, 35(4): 1-23

[38] Liu Shu-Dong, Meng Xiang-Wu. Recommender systems in location-based social networks. Chinese Journal of Computers, 2015, 38(2): 322-336(in Chinese)
(刘树栋，孟祥武. 基于位置的社会化网络推荐系统. 计算机学报，2015，38(2)：322-336)

[39] Chen X F, Zeng Y F, Cong G. On information coverage for location category based point-of-interest recommendation// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 37-43

[40] Chen X J, Hong W J, Nie F P, et al. Spectral clustering of large-scale data by directly solving normalized cut//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. London, UK, 2018: 1206-1215

[41] Shi J B, Malik J. Normalized cuts and image segmentation// Proceedings of the International Conference on Computer Vision and Pattern Recognition. San Juan, USA, 1997: 731-737

[42] Guan X, Li C T, Guan Y. Matrix factorization with rating completion: An enhanced SVD model for collaborative filtering recommender systems. IEEE Access, 2017, 5: 27668-27678

[43] Leskovec J, Lang K J, Mahoney M W. Empirical comparison of algorithms for network community detection//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 631-640

MENG Xiang-Fu, Ph. D., professor, Ph. D. supervisor. His research interests include spatial data management, recommender system and Web database query.

ZHANG Xiao-Yan, master, engineer. Her research interests include spatial data management, city computing, and data mining.

TANG Yan-Huan, M. S. His research interest is recommender system.

JIA Di, Ph. D., associate professor, Ph. D. supervisor. His research interests include image processing, machine learning and recommender system.

QI Xue-Yue, M. S. candidate. Her research interest is LBSN-based recommender system.

MAO Yue, M. S. candidate. Her research interests include short text analysis and recommender system.

Background

With the popularity of mobile Internet, Point-of-Interest (POI) recommendation system has become a new research hot spot. However, most POI recommendation algorithms only focus on meeting the user's personalized needs, and strive to provide the most close to the user preference list of POIs for the user, but ignore the diversity of the list of contents, lead to the single content of the recommendation list, so as not to broaden the user's view. To deal with this problem, we propose a diversified and personalized POI recommendation approach based on the geo-social relations between POIs, by fusing spectral clustering and matrix factorization algorithms. The experimental results demonstrate that our method cannot only improve the diversity of the recommendation list but also achieve the higher ranking precision, which realizes the integration of the diversification and personalization of POI recommendation.