分 类 号 学 号 M201472821

学校代码 10487 密 级



学位论文

基于共现图的兴趣点推荐算法研究

|  |  |
| --- | --- |
| 学位申请人： | 张 军 |
| 学科专业： | 计算机应用技术 |
| 指导教师： | 李玉华 副教授 |
| 答辩日期： | 2017年5月24日 |
|  |  |

**A Thesis Submitted in Full Fulfillment of the Requirements**

**for the Degree of the Master of Engineering**

Point of Interest Recommendation based on

Co-occurrence Graph

|  |  |
| --- | --- |
| **Candidate :** | Jun Zhang |
| **Major :** | **Computer Application Technology** |
| **Supervisor:** | Associate Professor Yuhua Li |
|  |  |

Huazhong University of Science & Technology

Wuhan 430074, P.R.China

May, 2017

**独创性声明**

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到，本声明的法律结果由本人承担。

学位论文作者签名：

  日期：     年   月   日

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密□，在\_\_\_\_\_年解密后适用本授权书。

本论文属于

不保密□。

（请在以上方框内打“√”）

 学位论文作者签名：                   指导教师签名：

日期：    年   月   日              日期：    年   月    日

# 摘 要

随着移动互联网和基于位置的社交网络（Location-based Social Network，LBSN）的发展，兴趣点(Point-of-Interest，POI)推荐成为工业界和学术界热门的研究方向。虽然个性化的兴趣点推荐系统能够极大的方便用户的出行，然而推荐的过程面临着许多的挑战，例如难以对用户的决策过程进行建模，以及数据稀疏性问题。现有研究基本都是从用户角度出发对兴趣点的推荐问题进行建模，然而数据的稀疏性影响该种方法建模的效果。

针对这些挑战和问题，本文提出了基于共现图的兴趣点推荐算法（Point Of Interest Recommendation based on co-occurrence Graph , PBG）。地理位置因素和兴趣点之间的相似性被同时融合到PBG模型中对兴趣点的推荐问题进行建模。在相似性方面，PBG模型基于用户访问的兴趣点之间存在着相似性这一观察，利用用户的签到信息构建兴趣点之间的共现图，并定义了共现图中节点间的两种相似性，进而推荐问题被转化为图中相似节点的搜索问题。兴趣点间的相似性被首次运用到POI推荐问题中，其使得推荐算法更强健，推荐结果更具解释性。针对地理位置因素，对每个用户的签到位置利用高斯核密度估计方法进行单独建模，这使得推荐更加个性化。

为了验证模型的有效性，我们在两个真实的位置社交网络数据集Foursquare和Gowalla中与已有相关的兴趣点推荐算法进行了大量对比，实验结果表明了PBG模型的有效性。

**关键词：**位置社交网络，兴趣点推荐，共现图，核密度估计，相似性

# Abstract

With the development of Mobile Internet and Location-based Social Network (LBSN), the recommendation of Point-of-Interest (POI) has become a popular research direction in industry and academia. Although a personalized POI recommender system can significantly facilitate users’ outdoor activities, it faces many challenging problems, such as the hardness to model user’s POI decision-making process and the difficulty to address data sparsity. Almost all of Existing works focus on the modeling of POI recommendation from the user’s point of view. However, the sparsity of data affects the effect of this method.

To address these challenges and problems, a POI Recommendation based on co-occurrence Graph, called PBG, is proposed in this paper. The similarity of POI and geographical influence is integrated into the PBG model for POI recommendation. In term of similarity, PBG uses the user's check-in information to construct the co-occurrence graph between the POI and define the two similarities between the nodes in the co-occurrence graph. Then, the recommendation problem is converted into the search of similar nodes in the graph. Specially, the similarity of POI is first applied into the recommendation, which make algorithm more robust and explainable. In term of geographical influence, we explore Gaussian kernel density estimation to model user’s location that makes recommendation more personalized.

To evaluate the prosed model PBG, We conduct extensive experiments with many related baseline methods on two real-world data sets, Foursquare and Gowalla. The experimental results show the effectiveness of our model.

**Key words:** Location-based Social Network，POI recommendation，co-occurrence graph，kernel density estimation，similarity

目录

[摘 要 1](#_Toc481169817)

[Abstract 2](#_Toc481169818)

[1 绪 论](#_Toc481169819)

[1.1 课题研究背景 5](#_Toc481169820)

[1.2 课题研究的目的和意义 6](#_Toc481169821)

[1.3 国内外研究现状 7](#_Toc481169822)

[1.4 论文的主要研究内容 12](#_Toc481169823)

[1.5 论文组织结构 12](#_Toc481169824)

[2 位置社交网络规律探索](#_Toc481169825)

[2.1 相关定义 14](#_Toc481169826)

[2.2 推荐问题定义 16](#_Toc481169827)

[2.3 LBSN网络规律探索 16](#_Toc481169828)

[2.4 本章小结 20](#_Toc481169829)

[3 基于共现图的兴趣点推荐算法（研究？）](#_Toc481169830)

[3.1 基于核密度的地理位置建模 22](#_Toc481169831)

[3.2 基于共现图的推荐模型 24](#_Toc481169832)

[3.3 融合的推荐模型 30](#_Toc481169833)

[3.4 基于共现图的兴趣点推荐算法 31](#_Toc481169834)

[3.5 与已有模型的区别 32](#_Toc481169835)

[3.6 本章小结 33](#_Toc481169836)

[4 实验与分析](#_Toc481169837)

[4.1 实验环境及数据情况 34](#_Toc481169838)

[4.2 实验评价度量 35](#_Toc481169839)

[4.3基准的方法 35](#_Toc481169840)

[4.4 实验结果及分析 36](#_Toc481169841)

[4.5 本章小结 45](#_Toc481169842)

[5 总结与展望](#_Toc481169843)

[5.1 论文总结 47](#_Toc481169844)

[5.2 论文展望 48](#_Toc481169845)

[致 谢 49](#_Toc481169846)

[参考文献 51](#_Toc481169847)

[附录1 攻读硕士学位期间投稿的论文 54](#_Toc481169848)

[附录2 攻读硕士学位期间参与的项目 55](#_Toc481169849)

# [1 绪 论](#_Toc283328270)

## 1.1 课题研究背景

21世纪移动互联网飞速发展，手机成为人们联通互联网与现实世界的重要大门。 智能手机和无线网络使得人们越来越容易实时获得位置信息。手机定位功能在人们的生活中占有重要的地位，其辅助于出行导航，位置记录等诸多功能。手机定位同时刺激了基于位置的社交网络LBSN(Location-based Social Network)的发展，比如新浪微博，Tweitter，[Foursquare](https://www.baidu.com/link?url=YZd5icPowOZBc1MH9eKTd4ZfD34tZZTHt3zUGTNW4iAptpTL2AJDIA0CCBodmADdzqNjcqn5CFIpnr01EAytg2xQ63txCzOwht8CQMC6FFK&wd=&eqid=cb99424000065c970000000358c509f0" \t "_blank)，Gowalla等社交网络。LBSN允许用户与他人建立联系，同时分享个人的感受，以及当前签到的POI位置信息。LBSN中，用户与POI之间的交互信息为研究个性化的POI推荐系统提供了千载难逢的机会。实际上，在LBSN网络中，个性化的POI推荐是一项非常必要的需求。用户考虑访问POI时，会面临着时间与信息量严重不对称的困境，而LBSN网络中的POI推荐功能能够化解此种困境，推荐给用户感兴趣的新位置；对于商家而言，LBSN网络中的POI推荐功能可以帮助商家发现潜在的客户以及竞争者，并据此有针对性的规划商业营销方案。

尽管目前常规推荐技术发展的如火如荼，但是对于LBSN网络中POI的推荐仍然是一个非常有挑战性的问题。考虑LBSN网络中用户的实际活动：用户在访问某一位置时，在LBSN网络中进行签到，签到信息是一个具有位置信息的博文，该博文映射了该用户此时此刻，此位置的活动以及此时的情感信息；用户全部的博文信息就可以一定的程度上反映出用户的日常生活规律以及兴趣爱好。用户不仅仅发表自我的状态博文，同时会通过LBSN网络中的关注或者朋友功能浏览身边或者朋友的签到信息。用户的活动中包含的用户，位置，时间，以及兴趣等多方面，多类别的因素给POI推荐的建模造成了巨大的挑战。另一方面，用户对POI的访问很大程度上受限于POI与用户当前位置的距离以及用户拥有的时间。用户的签到过程涉及到实际的地理位置的访问，距离的远近意味着不同的访问代价。当两个位置的性价比相当时，用户将更倾向于访问自己周边的位置。第三，LBSN网络中，用户的签到数据非常稀疏，平均每个用户不到一百个，部分用户的签到数据只有个位数。稀疏的签到数据主要由周到的到家服务以及快速的生活节奏使得用户很少到自己感兴趣的位置享受生活。用户在现实场景下的许多访问活动可能都未被用户记录。稀疏的签到信息使得一般的POI推荐算法很难够精确的表达用户的兴趣。第四，POI推荐具有冷启动的问 的历史签到信息，很难将新的POI推荐给一个用户或者给新用户推荐其感兴趣的POI。

针对POI推荐的特殊因素以及挑战，关键是如何最大化的利用已有的数据对POI推荐问题进行建模。本文将主要从POI的角度，利用POI之间的相似性对POI的推荐问题进行建模，以期获得更优的推荐效果。

## 1.2 课题研究的目的和意义

LBSN网络中的POI推荐服务是一个非常亟需的应用，高性能的POI推荐能够准确的识别用户的偏好，以及当前的实际场景，并返回给用户最优的POI列表。POI推荐不仅可以极大的节省用户的时间，提升用户的体验度；而且可以帮助商家发现潜在的顾客。因此，对于POI推荐问题的研究将有助于提升现实生活中用户的生活体验以及促进LBSN网络的发展。

在学术研究方面，POI推荐问题涉及到时间，空间，用户和POI等众多实体和因素，属于异构数据源，多影响因素的推荐场景。该问题在学术领域仍然是一个具有重大挑战性的问题。POI推荐不同于传统的推荐系统，其特殊的挑战在于：(1) 隐式的反馈数据。数据中只有用户的签到信息，而没有喜好信息。(2) 位置影响较大。 用户更倾向于访问较近的位置。（3）数据稀疏性问题。POI推荐中的数据稀疏性主要体现在：用户签到的POI数量稀少，对于签到活动的描述信息简短，同时缺乏POI本身的描述信息。（4）冷启动问题。难以对新的用户和新的POI进行精准的推荐。（5）时间特性。用户在不同的时间段访问的位置类型差异性很大。时间敏感的POI推荐算法也是目前学术研究的热点。（6）网络动态性。LBSN网络中，用户的签到信息每时每刻都在不断产生。增量式可扩展的POI推荐算法将更适用于生产实践。

本文将主要着眼于一般的POI推荐问题，特别是本地场景下的POI推荐，主要的原因是：异地场景下，作为一个旅游者，用户的访问行为更多的受大众的访问行为所影响，此时基于位置的流行度和类别标签的推荐完全可以满足用户需求。而在本地场景下，用户的访问行为更多的遵循于自我兴趣和日常生活需求，此时推荐系统需要挖掘出用户真正的兴趣。因此，相比异地推荐，对本地推荐进行研究将更有意义。目前对POI推荐问题的研究大多是从用户的角度出发，并考虑影响用户访问POI的各种因素例如：朋友关系，地理位置，用户兴趣等对该问题进行建模。另外有研究利用概率图模型的生成过程类比用户访问POI的决策过程，学习得到用户在各个因素影响下访问某一POI的概率。但是LBSN网络中稀疏的签到信息影响从用户角度对该问题进行建模的效果。

本文旨在从POI的角度出发，提出一个新的POI推荐算法，以解决数据稀疏性和POI推荐的性能与效率的问题。

## 1.3 国内外研究现状

POI推荐是数据挖掘领域的一个热门研究方向。因为POI连接了线上和线下，从而引出了许多不同于传统推荐问题的影响因素。目前针对POI推荐的研究主要从POI推荐的几个特点：位置特性，用户关系，用户兴趣，时间特性，以及冷启动问题进行入手。考虑多因素进行POI推荐的模型又可以分为融合模型和联合模型。融合模型对不同的因素进行分别建模，并融合分别得到的推荐结果；联合模型将多个因素统一到一个模型中进行建模，并给出唯一的推荐结果。本章节将分别按照研究的出发点对POI推荐的国内外相关研究现状进行简要介绍。

首先将介绍隐式反馈推荐相关的研究进展以及矩阵分解技术在POI推荐中的应用，之后按照POI推荐中的位置，用户关系，时间因素在POI推荐问题中如何应用的角度进行研究现状的介绍，最后对POI推荐问题的各种建模方案进行简单介绍。

### 1.3.1 隐式反馈的推荐

早期的推荐问题主要针对显示的评价信息，例如电影，商品的评分，并将推荐问题建模为评价的预测问题。 先前的研究[1-3]皆属于针对此种问题的研究。但是在现实的社交网络中，例如LBSN网络中，只存在用户的POI签到信息，而缺少用户对于POI的显示反馈信息。在缺少负样本的场景中，利用以往针对显示反馈的推荐方案对问题进行建模是不合理的。为了解决这个问题，一个非常流行的方法[4]将所有的缺失信息作为负样本进行建模学习，然而此种方法因为考虑了全部的观察数据和缺失数据，所以存在着学习效率的问题。针对隐式反馈的计数场景，Gopalan等[5]提出了基于泊松分解的推荐技术，泊松分解假设隐藏变量用户和物品服从Gamma分布，观察值服从泊松分布，该模型仅仅通过非零值进行参数优化，所有具有较高的学习效率。隐式反馈的难点在于难以度量用户对于物品的喜爱程度。研究[6]使用排序的方法解决该问题，排序过程给出用户和物品之间的相对关系，而不对评分值直接建模，因此推荐效果更加健壮。位置社交网络中的POI推荐属于特殊场景下的隐式反馈，其需要融合特殊的因素以优化推荐效果。POI推荐问题的位置特性和时间特性使得已有的矩阵分解技术很难运用到POI推荐场景种。Lian等[7] 提出了非负的加权矩阵分解的方法，增强用户已访问的POI的权重，同时弱化为访问的POI的权重，模型通过扩展用户和POI的隐含因子来囊括影响用户行为的地理位置因素。余等[8]提出了基于Ranking的泊松矩阵分解兴趣点推荐算法。该模型利用泊松分布模型替代高斯分布模型建模用户在兴趣点上签到行为；然后采用BPR(Bayesian personalized ranking)[6]标准优化泊松矩阵分解的损失函数，拟合用户在兴趣点对上的偏序关系；最后，利用地域影响力作为正则化因子约束泊松矩阵分解的过程。

### 1.3.2 位置因素

位置因素是POI推荐区别于其他传统推荐的主要因素，因为用户签到涉及到该POI的地理位置特征，不同的距离意味着不同的代价。已有许多研究[ 9-17]尝试利用位置因素去提升POI推荐的效果。位置因素通过距离的远近，以及地区热度对人们的访问行为产生影响。目前针对位置因素的建模主要分为三种方法：幂律分布，高斯分布和核密度估计模型。Ye等[11]使用幂律分布对位置的影响力进行建模，其假设两POI之间的距离与两个POI被同时访问的概率之间服从幂律分布，通过实验得出了用户的访问行为受地理位置因素的影响很大，同时用户的签到位置呈现聚集的现象；Cheng等[9]提出了多中心的高斯模型MGM对位置影响力进行建模，但是因为MGM模型会不均衡的将POI分配给多个活动中心，因此Zhao等[15]提出了基于遗传算法的高斯混合模型对POI推荐中的位置因素进行建模。Zhang等[14]结合地理位置因素，社交朋友关系，以及POI的类别关系提出了GeoSoCa模型，该模型使用自适应的核密度估计方法对用户签到的位置关系进行建模，同时使用幂律分布对朋友关系和用户对类别的喜好程度进行建模，通过相乘的方式融合三个因素得到用户对POI的最终评分。使用自适应的核密度函数对用户的位置影响力单独建模的GeoSoCa能够更准确的评估出用户在不同位置区域的访问热度，但是同时需要付出较大的计算代价。另一方面，Liu[10]从地理位置的角度利用位置因素，其分别从个体的层次和区域的层次对位置的邻居影响力进行出发建模。在个体的层面，用户对于一个位置的兴趣主要表现为用户对该位置的特殊兴趣，以及该位置附近邻居的兴趣组合；在区域层面，该研究使用L1范数作为正则项约束学习该位置的潜在向量。Wang等[18]提出的GeoSAGE模型同样考虑了区域性位置体现出的兴趣，但是其将区域性的兴趣分为本地人在该区域体现的兴趣，以及旅游者在该区域体现的兴趣。

### 1.3.3 朋友关系

基于LBSN网络中的线上朋友关系将拥有相近的喜好这一假设，已有研究[9-11; 19-23]利用朋友关系来提升POI的推荐效果。Ye等[22]提出了一个基于内存和朋友关系的协同过滤模型FCF，FCF模型从朋友中而不是所有用户中寻找与用户爱好相似的用户进行协同过滤。Cheng等[9]通过将社交关系作为正则项融入到概率矩阵分解模型PMF[24]，在该模型中，融入的社交关系使得具有朋友关系的用户具有相近的潜在特征空间。Zhang等[14]利用幂律分布对用户之间的朋友关系进行建模。尽管社交关系极大的改善了传统推荐算法的性能，但是在POI推荐中社交影响力却被证明具有极小的作用[9]。这可能是因为LBSN网络中的朋友关系受地理位置因素的约束，使得虽然用户具有相同的兴趣，但是却很难访问相同的POI，因为用户所处位置不同，其访问同一POI的代价也不同。Li等[21]对朋友关系进行了重新的定义，该方法将朋友关系分为了线上朋友，本地朋友，以及邻居朋友关系。其中线上朋友关系指通过LBSN网络连接起来的用户关系；本地朋友为访问过相同的POI的用户；邻居朋友为位于同一区域的用户，之后分别从三种朋友关系去挖掘用户可能感兴趣的商品，并以融合模型的方式融合得到的推荐结果。

### 1.3.4 时间因素

时间因素是POI推荐中的重要因素。一方面是因为，地理位置的限制使得用户的签到行为具有一定的规律，另外一方面，时间因素有助于时间敏感的POI推荐。POI推荐中的时间因素主要表现为两个方面：周期性，顺序性。用户签到的周期性主要表现为用户周期性的访问某一POI，例如用户每周末的超市购物行为。许多研究[12; 23-32]利用这一特性提升POI推荐的质量。顺序性主要是用户连续的签到，所形成的签到序列。顺序性主要体现在用户的行为习惯上，例如用户喜欢聚餐之后去KTV这一现象。Zhao等[29]利用连续签到的时间和空间属性对数据进行了分析，结果显示许多连续的签到信息具有高度的相关性：超过百分之四十和百分之六十的连续签到行为发生在4个小时之内，同时超过百分之九十的连续签到信息发生在32千米之内。针对POI 推荐领域中的时间特性，研究[23; 30-32]利用马尔可夫链对该序列模式进行建模。其中研究[30-31]假设一个用户短时间内签到的两个POI是高度相关的，并使用基于分解的个性化马尔可夫连FPMC[24]来推荐连续的POI。Zhang等[23]提出了一个加和的马尔可夫模型去学习两个连续POI的转移概率。Zhao等[29]利用潜在分解模型去获取连续POI之间的关系。

### 1.3.5 内容因素

LBSN网络中的内容信息主要包括了用户签到时所留下的出行建议信息，对POI的评价信息，用户所拍的POI照片，以及POI固有的描述信息。虽然用户的签到信息中没有显示的POI评分信息，但是签到信息中可能包含了用户的心情，以及对POI进行描述的词，比如：漂亮，壮观等，这些表达情感的词从一定层面上可以反映出用户对该POI的喜爱程度。另外一方面，从用户签到的内容信息中可以挖掘出体现用户喜好的信息，例如用户喜爱自然风光类的旅游景点，而不是所有的旅游景点。研究[33-37]都利用用户的评论信息来提升POI推荐的效果。Yang等[36]通过分析用户的评论信息得到用户对该POI的情感，进而调整用户对该POI的喜爱程度。在情感分析中，不同的词可以分为积极和消极两个层面，积极的词反映出用户对该POI的喜爱，消极的词反映出不满意。作者将积极和消极的词分别映射到不同的评分等级，进而将问题转化为显示反馈的推荐问题。

除了按照上述不同因素对POI推荐问题的研究方法进行分类外，下面按照不同的分析建模方法对普通的POI推荐问题进行分类，我们将其分为以下三类。第一类, 使用概率图模型的方式对用户的决策过程建模：Yin等[38]利用概率图模型对POI推荐场景下的时间因素，地理位置因素，以及内容因素进行建模，生成式的概率图模型对用户的POI访问场景进行模拟。Wang等[18]同样基于概率图模型提出了GeoSAGE模型对用户的访问行为进行建模。GeoSAGE模型考虑了用户的个体兴趣和目标区域的整体兴趣，同时将用户的身份细分为本地人，旅游者，不同的身份受不同的因素所影响。GeoSAGE通过使用金字塔模型，多层次求均值的方式削弱数据稀疏性对推荐性能的影响。任等[39]提出一种联合概率生成模型GTSCP模拟用户的决策过程。该模型融合了时间，地点，内容等多种因素，其属于联合模型。概率图模型能够以统一，直观的方式对用户的行为建模，但是在POI推荐这种数据稀疏的情况很难达到较好的建模效果。第二类方法使用排序而非预测的方式对该问题进行建模：Li等[16]提出了一个基于分解的排序方法Rank-GeoFM, 该模型尝试将推荐问题看作排序问题，而非评分的预测问题，并尝试拟合用户对于POI的喜爱程度，以缓解数据的稀疏性。 Li等[21]同样提出了排序模型 ARMF，但不同之处在于其将POI分为了用户已访问过的，感兴趣但目前未被访问的，以及其他的POI。作者通过实验证明了排序过程建模的效果优于矩阵分解建模的效果。排序的建模方法通过排序而非预测的方式减弱了数据稀疏性以及数据中噪音的影响，因此一般具有较好的推荐效果，然而排序过程的时间开销较大，且难以进行增量式的模型更新。第三类，使用目前流行的词嵌入与图嵌入的方式对POI推荐问题进行建模。Xie等[40]利用图嵌入的方式对POI推荐场景中的实体：用户，POI，类别等进行分布式表示的学习。Feng等[41]基于word2vec[42]的思想提出了POI2Vec模型，POI2Vec模型从POI的角度出发对POI进行分布式的向量表示学习。POI2Vec模型通过二分树的方式对位置因素进行建模。虽然实体的分布式表示在自然语言领域取得了巨大的成功，然而在POI推荐场景对实体进行分布式表示并不能带来更多的益处。

尽管上述的研究成果基本上考虑了POI推荐问题中的各种影响因素，但是它们基本都是对用户个人的兴趣进行直接建模或者从用户的角度出发解决问题。稀疏的数据使得该种推荐算法容易受噪音的影响，同时算法的效率差。在大规模的LBSN网络中，上述算法很难被应用到生产实践中去。鉴于已有研究所存在的缺陷，本课题将从POI的角度出发，通过在POI共现图中寻找相似的POI来进行POI推荐。本研究的主要目标是提出一个更健壮，性能更好的POI推荐模型。

## 1.4 论文的主要研究内容

本文以普通的POI推荐问题为研究背景，从POI角度出发利用POI之间的相似性对该问题进行建模，提出了一个基于共现图的兴趣点推荐模型。

本文主要的研究内容包括：

1. 基于LBSN社交网络中用户行为规律探索。将从兴趣点的类别，用户的朋友关系以及用户访问的地理位置三个不同的角度出发探索位置社交网络中的规律。

2. 基于共现图的兴趣点推荐算法。利用核密度估计的方法对影响用户访问的地理位置因素进行建模，基于共现图，利用节点之间的相似性对兴趣点进行推荐，通过融合模型的方式将位置因素和相似性的推荐进行结合。

3. 模型验证， 利用真实的LBSN网络数据集对所提出的模型在推荐的精度和召回率以及运行效率方面进行验证，以证明算法的有效性。

## 1.5 论文组织结构

第1章，绪论。对本课题的研究背景、目的及意义、国内外已有的研究现状进行总结，并且简要的介绍了本文所要研究的内容。

第2章，位置社交网络规律探索。本章针对LBSN网络中用户的签到数据进行多角度的分析，以发现LBSN网络中用户特殊的行为规律。

第3章，基于共现图的兴趣点推荐算法。在发现的规律基础之上，使用核密度的方式对用户的签到位置进行建模，基于共现图利用节点之间的相似性进行搜索推荐，并给出融合的推荐模型。

第4章，实验与分析。本章对实验的环境，数据，以及对比算法进行了说明，并且从推荐效果，以及性能方面与已有研究进行了细致深入的分析对比。

第5章，总结与展望。对本课题的工作进行概括总结，针对基于共现图的兴趣点推荐模型的缺点提出了未来可能进行改进的地方，并对后续的研究工作进行了介绍。

# 2．位置社交网络规律探索

推荐模型是推荐问题的核心部分，它的主要任务是对LBSN网络中的各种信息进行逻辑上的建模。LBSN网络中的推荐问题不同于一般场景下的推荐问题，其具有特殊性。本章首先对LBSN网络进行介绍和数学化的定义，并根据已有LBSN网络中的用户签到信息探索LBSN网络中用户的行为规律。

## 2.1 相关定义

下面，我们将给出本文涉及到的关键数据结构和数学符号的定义。

**定义2.1 POI。**本文使用符号 表示POI，POI指LBSN网络中对应的实体商家或者位置，具有实际的经纬度属性和类别信息。其中每个POI具有唯一的位置属性，而对于类别信息来说，每个*l*可能有多个类别信息。例如某一集成了吃饭，电影等功能的广场将会有食物，娱乐等类别标签。

**定义2.2 用户活动。**用户活动是一个的三元组, 其表示用户在时刻访问了位置。用户活动通过LBSN网络中用户的签到信息进行体现。用户的所有历史活动记录构成了访问矩阵R，其中表示用户访问位置的次数，是一个非负值。

**定义 2.3 朋友关系。**朋友关系指用户通过关注和互粉行为在LBSN网络中形成的联系。LBSN网络中，朋友之间可以浏览朋友的签到信息。本文使用符号表示朋友，使用符号表示用户的朋友集合。

**定义2.4 LBSN网络。**LBSN即基于位置的社交网络**。**POI推荐场景下的LBSN网络是一个用户和位置*l*组合而成的异质网络，该网络可以表示为*,* 其中代表网络中所有用户的集合；表示LBSN网络中所有POI的集合；*E*代表网络中的边，其包含了三种关系: 用户和用户之间的朋友关系，用户和位置之间的访问关系；以及位置与地理坐标和类别标签之间的所属关系；符号*C*代表的是网络中所有的类别标签集合。为了便于理解，图2-1给出了简单的LBSN网络场景图。

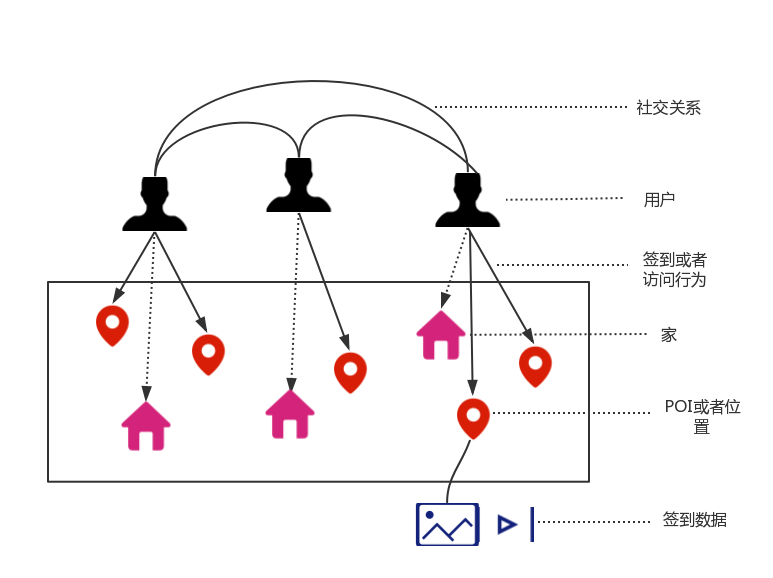
****

图2-1 LBSN网络场景图

本文模型所使用的数据集D包含了四个元素：用户，位置，时间，以及位置的类别属性。

为简明清晰，表2-1列出了本文涉及的数学符号的定义，以及相关含义。

表2.1 本文定义的数学符号以及相应含义

|  |  |
| --- | --- |
| 符号 | 含义 |
| *U* | 用户的集合 |
| *L* | LBSN网络中POI的集合 |
| *E* | LBSN网络中的边的集合 |
|  | 用户之间的朋友关系矩阵 |
|  | 用户与位置之间的访问矩阵 |
|  | 用户*u*访问过的位置集合 |

为了理解和叙述方便，本文做如下约定：

1. 我们将混合使用POI三种术语来表示兴趣点。

2. 我们使用表示用户访问过的位置。对于其他矩阵，我们使用统一的表达含义。矩阵带单个下标表示：矩阵第行所有值不为0的元素对应的列所组成的集合，双下标*ij*表示矩阵行*j*列的值。例如表示用户*u*的所有朋友，表示用户*u*和*v*是否是朋友。

3. 我们使用取模符号表示集合的大小。例如：表示集合的大小，即用户已访问位置的数量。

## 2.2 推荐问题定义

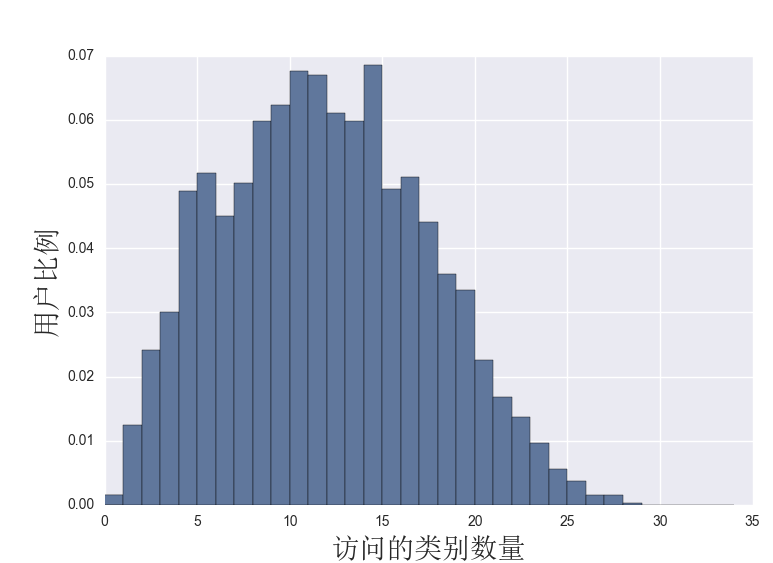
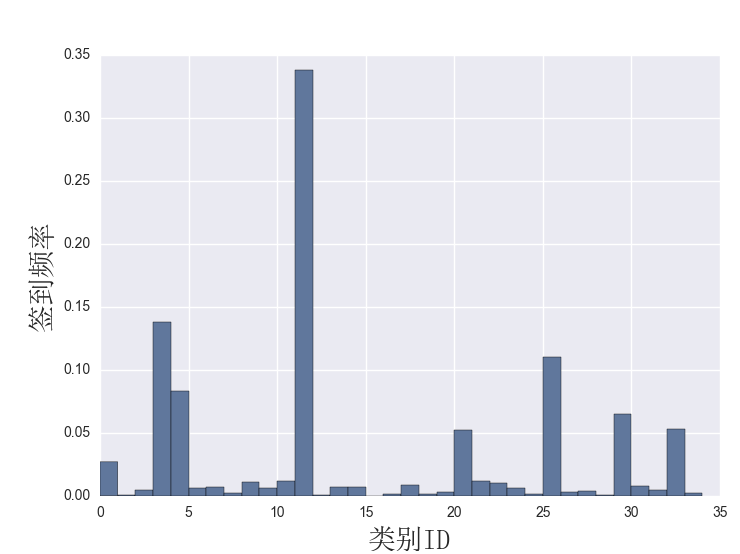
**定义2.5 POI推荐问题。**POI推荐问题定义为：给定LBSN网络中用户的历史签到信息（包含经纬度以及类别信息），POI推荐的目标是：预测用户*u*对于未访问过的兴趣点*l*的评分，并返回评分最高的top-K个兴趣点推荐给用户*u*。

## 2.3 LBSN网络规律探索

为了更好的理解LBSN网络中用户的签到习惯，本小节我们将对Foursquare和Gowalla中的签到数据进行分析。数据的来源和具体描述可参见章节4.1。数据分析将分别从POI的类别标签，朋友关系，以及签到的地理位置三个方面进行。

### 2.3.1 类别标签

在LBSN网络中，每个POI都被标注了几个类别标签，比如食物，咖啡厅，游乐场等。POI的类别标签能够很大的程度上表明该POI提供的服务类型，例如”shop and food service”的标签表明该POI能够提供购物和用餐的服务。用户在LBSN网络中可以通过类别标签进行搜索自己需要的POI。通过分析用户访问过的POI的类别标签可以得到用户对于类别的偏好。因为Gowalla数据集中类别标签信息缺失，所以类别标签的分析实验仅仅在Foursquare数据集上进行。数据分析的主要目的是探索用户对于类别的喜好规律，以及用户的类别喜好随时间的变化情况。图2-2是对于Foursquare数据集签到信息的POI类别标签进行分析的结果，其中图2-2(a )是对所有签到信息的POI类别进行统计得到的结果，从中可以看出不同的标签具有不同的流行度，其中具有较高流行度的标签类别3，4，11，25分别为：“Shop & Service”，“Arts & Entertainment”，“Food”，“Nightlift Spot”。图2-2(b)是对用户所访问的类别数进行统计得到的结果，从图中可以得出每个用户平均访问的POI个数大概为10个，这说明用户的兴趣是广泛而非单一的。



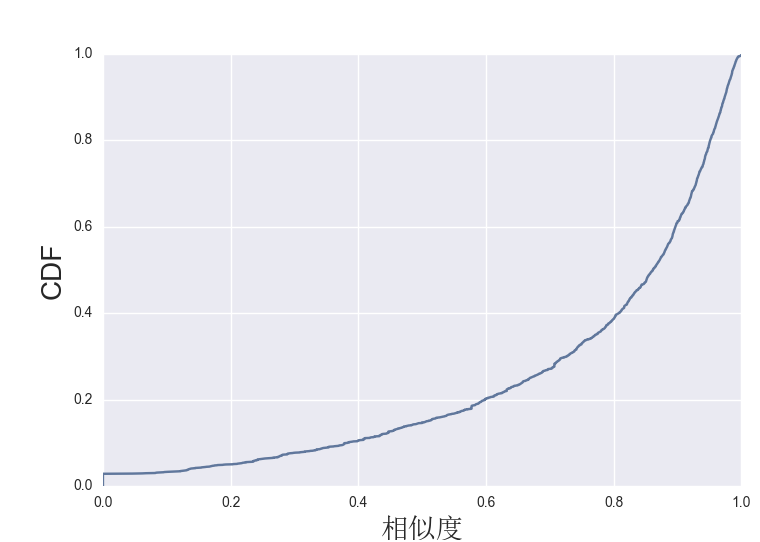
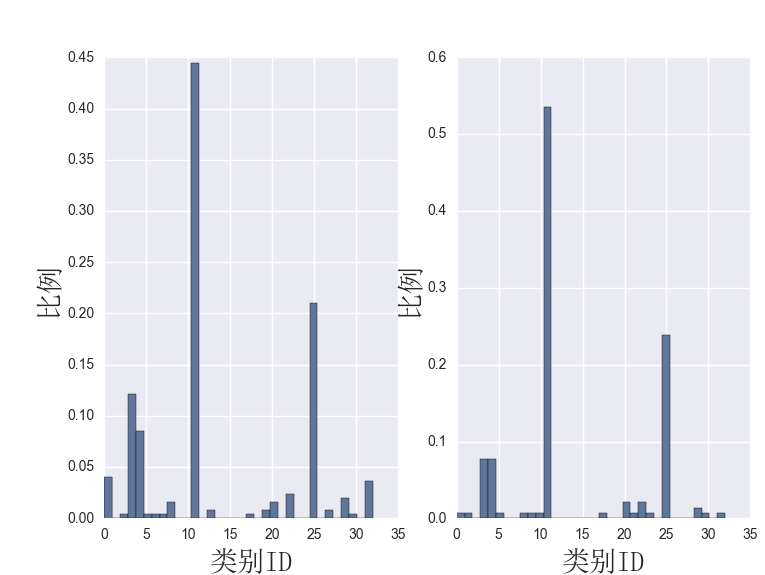
(a)类别签到频率 (b)访问类别数所对应的用户比例

图2-2 用户签到类别分析图

在传统的推荐场景中，一个常见的问题是用户的兴趣会随着时间发生漂移。为了观察在LBSN网络中用户的兴趣变化情况，我们将签到数据按照发生时间排序，并对其两等分；之后对每份中的数据进行用户兴趣分析，计算用户对每个类别物品的喜好程度。因为数据稀疏性导致的样本数量小，以及用户签到不全等噪音都使得利用数据刻画用户每个时期的真实兴趣是相当困难的。这里我们仅仅给出一个示例用户9两个时间段的兴趣占比图，如图2-3(a )所示，以及统计得到每个用户前后两个时间段的兴趣相似度，并给出了累计分布函数（cumulative distribution function，CDF）[[1]](#footnote-1)值，如图2-3(b)所示。我们使用了余弦相似度度量用户前后兴趣的相似性，并且在实验中剔除了任一时间段没有签到信息的用户，因为这些用户相似度是未定义的。

从图2-3(a) 中可以发现用户9前后两个时间段的兴趣大致相同。从图2-3(b)中可以发现前后两个时间段内兴趣相似性在0.6以上的用户占比达到了80%以上。据此，我们可以得出用户的兴趣随时间的变化幅度相对较小。

总的来说，LBSN网络中用户对于类别的喜好有所偏重，且用户的兴趣呈多元化发展，同时个人兴趣随时间的变化幅度相对较小。据此，我们可以得出用户前后两个时间段所访问的POI在类别上具有一定的自相关性[[2]](#footnote-2)。

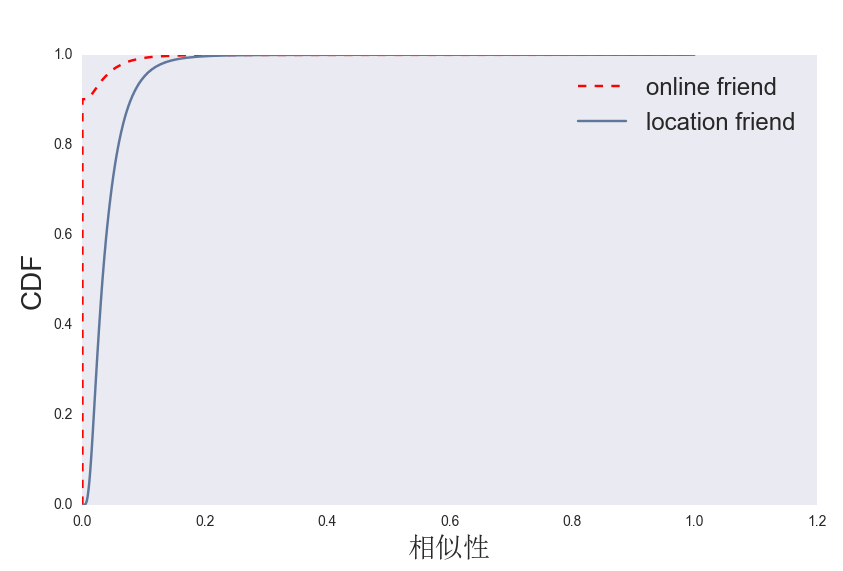
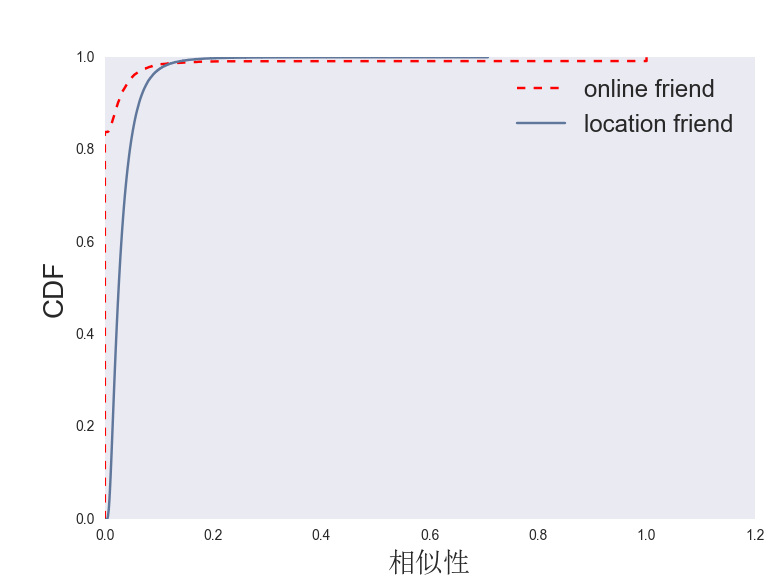


(a)用户9先后两个时间段访问的类别比占 (b)所有用户前后两个时间段兴趣的相似度

图2-3 用户兴趣随时间变化的分析

### 2.3.1 朋友关系

在LBSN网络中，朋友关系是一般指通过线上关注所形成的签到信息相互分享的关系。现实中相互认识的用户，以及具有相同兴趣的用户都有可能结成线上朋友。 Li等[21]将朋友关系进一步的划分为了线上社交关系形成的朋友(online friend)，访问过相同位置形成的朋友关系(location friend)，和家庭位置距离很近的用户之间所形成的朋友关系。最后一种朋友关系本质上仍然是由于距离和兴趣因素所促成的朋友关系。这里，我们分析数据集中前两种朋友关系。图2-4为两种 LBSN网络中online friend 和location friend之间的相似性相对应的CDF值，实验中我们使用余弦相似度计算朋友之间的相似性。



1. Foursquare数据集 (b)Gowalla数据集

图2-4 LBSN网络中朋友之间的相似性

从图2-4中可以看出，线上朋友之间具有非常低的相似度，百分之八十的朋友之间的相似度基本为0；访问过相同位置形成的朋友之间的相似度也非常低，所有的朋友之间的相似度低于0.2；同时可以发现Gowalla数据集中访问过相同位置所形成的朋友之间的相似性较Foursquare数据集中朋友之间的相似高。在数据分析实验中发现，朋友之间较低的相似性是由LBSN网络中稀疏的数据，及用户签到数量不均衡造成的。

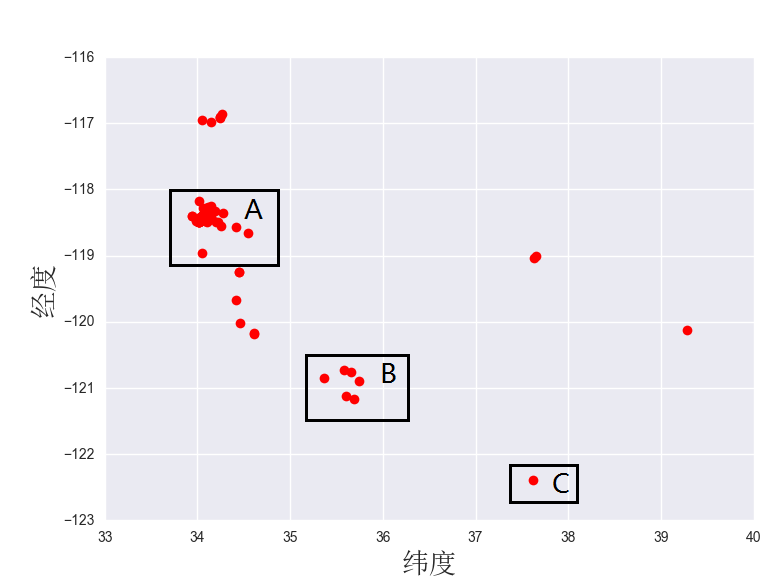
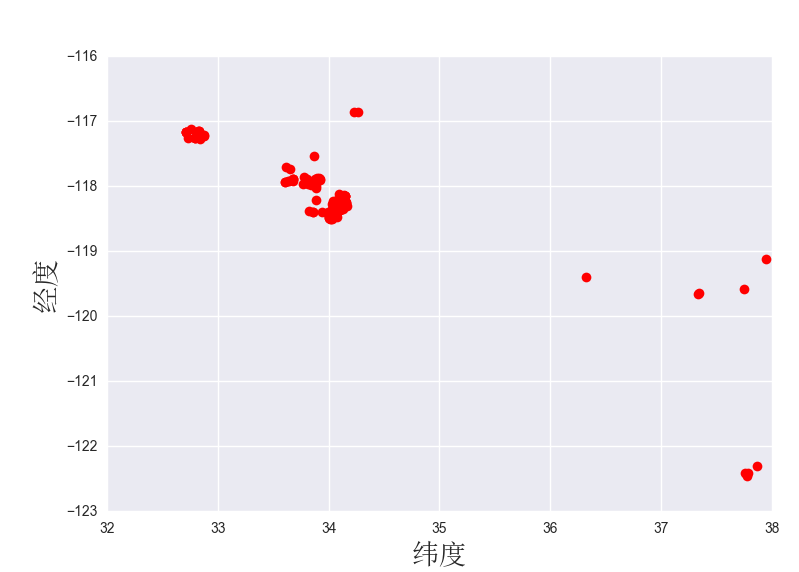
总的来说，LBSN网络中用户朋友之间的相似性较低，线上朋友之间的相似度基本可以忽略不计，同时数据稀疏以及用户签到数量不均衡是造成相似度低的主要原因。

### 2.3.3地理位置

不同于传统推荐中的商品，例如：图书，音乐，电影等，LBSN网络中对应的物品POI具有经纬度属性。当用户访问某一POI时，必须从家或者当前位置到达POI的位置进行消费。用户移动到POI所在地的代价是用户在访问该POI时需要付出的代价，该代价包括时间，金钱和体力等。一般来说，距离越短，代价越低，因此当两个POI提供的服务性价比相当时，用户将更偏向于访问距离较近的POI。

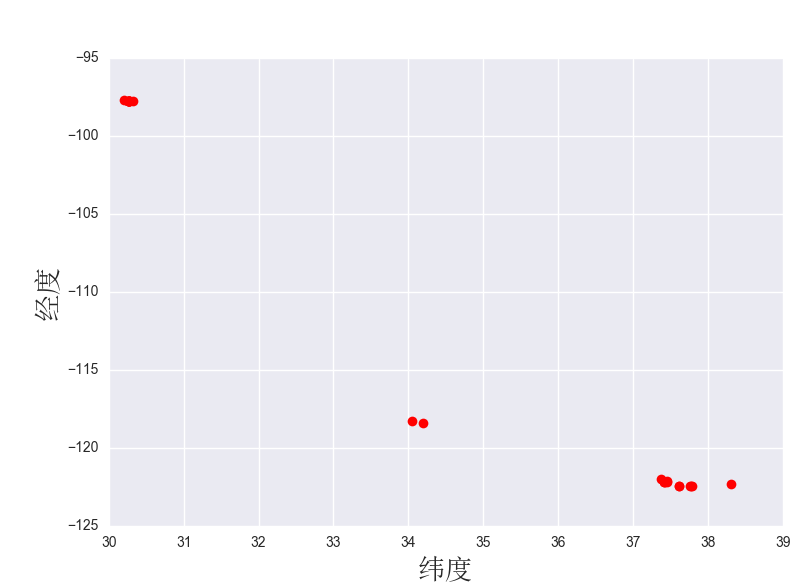
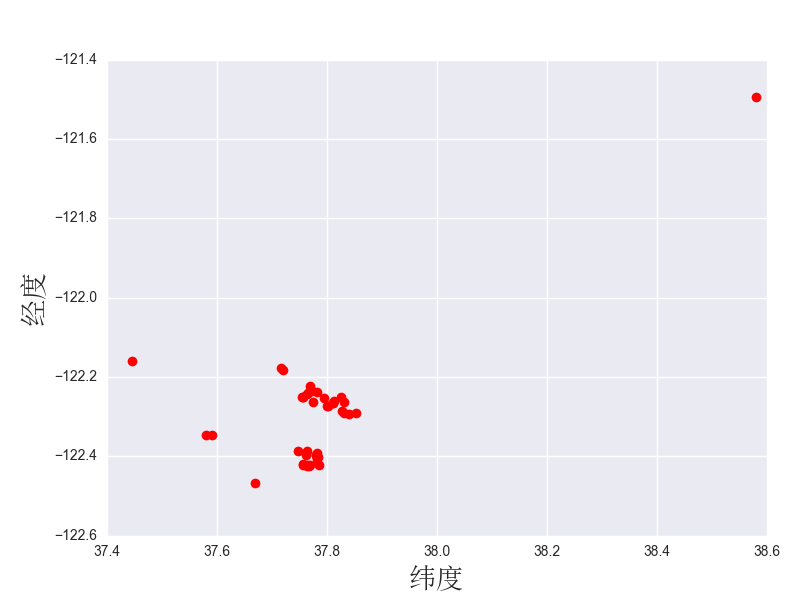
为了观察用户在位置上的签到规律，我们分别在两个数据集中随机选取了两个用户，并将其签到的位置信息映射到地图中，得到的结果如图2-5，2-6所示。从图中可以看出，用户在地理位置上的签到行为呈多中心聚簇的现象，一个可能的解释是：聚簇中心分别对应着办公地点，用户的家或者用户亲人的住地。因为距离聚簇中心越近的位置，被用户访问的可能性越大，距离越远的位置被访问的可能性越小，所以考虑某一位置被用户访问的可能性时，应该由距离该位置最近的聚簇中心所决定。

从图2-5中，另外一个可以发现的规律是：用户在不同区域表现出不同的访问概率密度分布。如图中用户5签到的两个聚簇中心A和B，A区域的密度较B区域的密度大，所以我们有理由可以相信用户5未来访问区域A附近的POI的可能性更大。区域访问密度的不同可能是由于用户的兴趣和该区域的发展状况有关。同时用户的签到图中存在着一些离聚簇中心较远的异常点，这些异常点可能是一些旅游景点，或者朋友的家庭等，例如，用户5访问的C点对应于美国峡谷国家公园。

(a)用户5的签到位置 (b) 用户9的签到位置

图2-5 Foursquare中两个用户的签到位置



(a)用户14的签到位置 (b) 用户34的签到位置

图2-6 Gowalla中两个用户的签到位置

图2.6是Gowalla数据集中两个随机用户的签到位置信息，从图中可以发现用户的位置访问规律与Foursquare数据集中的规律类似。

总的来说，用户的签到活动有以下几个特点：每位用户的活动区域不同，但是活动区域都呈现多中心，聚簇的特点；活动中存在着少许偏离聚簇中心的异常活动；用户在每个聚簇中心的访问密度具有差异性。

## 2.4 本章小结

本章给出了论文中将会使用到的数学符号的定义，并对LBSN网络进行了数学化的描述和介绍，之后针对真实的LBSN网络签到数据从POI的类型，朋友关系，以及地理位置因素三个方面进行了分析。下一章将根据本章所分析出的LBSN网络用户行为规律设计高性能的POI推荐模型。

# 3 基于共现图的兴趣点推荐算法

LBSN网络中用户访问POI的行为在POI的类别，地理位置，及用户的社交关系中都呈现出不同的规律。本章将基于这些潜在的规律分别对POI推荐问题中的位置因素，以及用户喜好进行建模。

## 3.1 基于核密度的地理位置因素建模

通过第二章中对于地理位置因素的分析，我们可以得到用户签到的位置呈现多签到中心的现象，距离签到中心越近的位置，用户访问的可能性越大；同时，在不同的区域，用户呈现出不同的签到密度的结论。据此，我们使用无参的密度估计方法——核密度估计，来对影响用户签到的地理位置因素进行建模。本节首先对核密度的知识进行简单介绍，之后给出使用高斯核函数对地理位置因素进行建模的方案。

### 3.1.1 核密度估计

核密度估计[43]是一种非参数化概率密度估计方法，相比参数化的概率密度建模方法，核密度估计不对数据真实所服从的概率分布做任何的假设，而是直接从数据中估计潜在的概率密度分布。因而核密度估计具有更高的准确性和适用性。

假设观测到D维空间某一未知概率分布*p(x)*的数据集，概率密度估计的目标就是估计*p(x)*的值。传统概率密度估计方法直方图将D维空间划分为较小的空间, 通过统计中数据点的个数估计该区域的密度分布。但是使用直方图方法造成估计的概率密度具有不连续性。

核密度估计和直方图具有相似的思想。假设区域R为以*x*为中心的小超立方体， 则可以使用公式(3.1)和公式(3.2)统计位于区域R内数据点的数量K:

(3.1)

(3.2)

上式中如果数据点位于以*x*为中心的立方体中，那么的值为1，否则值为0。其中参数*h*被称为带宽，其用于控制概率密度函数的光滑度。被称为核函数。上式中是一个间断函数，也会造成概率密度估计的不连续性。更多可选的平滑核函数包括高斯核函数，线性核函数，抛物线核函数等。

### 3.1.2 基于高斯核函数的地理位置因素建模

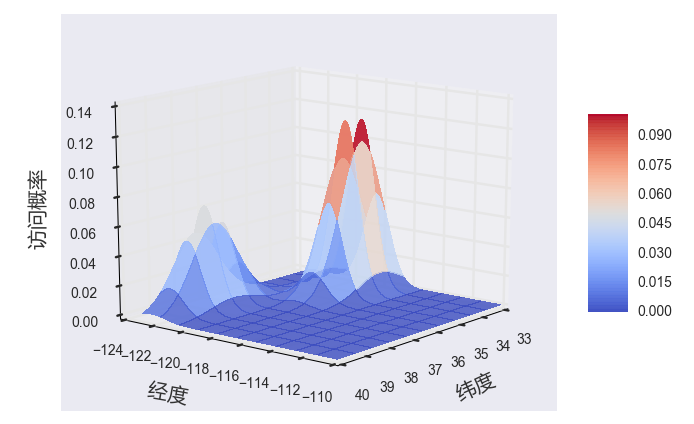
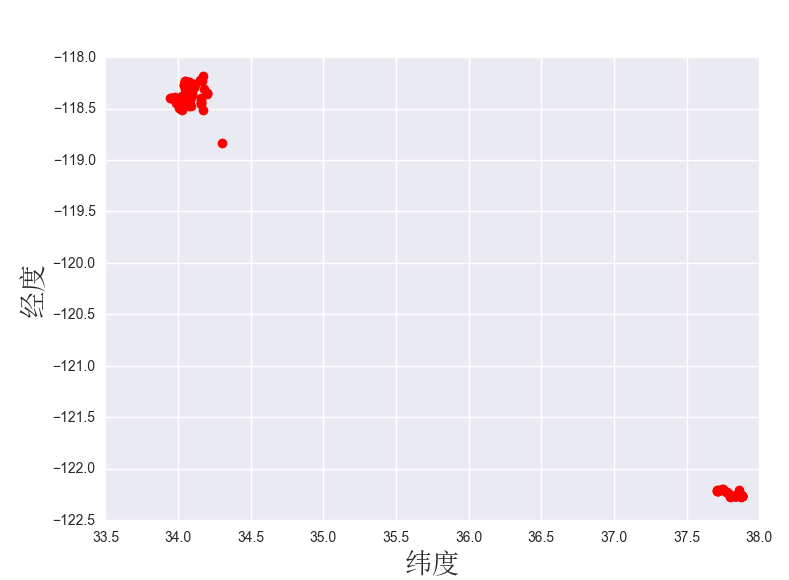
在我们的问题中，我们选取最常用的高斯核函数对用户的签到位置进行密度估计。假设已知用户已访问的位置集合，共*n*个位置。则对于一个未被访问的地点*l*，*u*访问*l*的概率可以用公式(3.3)进行计算。

(3.3)

其中高斯核函数的计算公式如(3.4)所示：

(3.4)

在核密度估计中，带宽的选择是问题的关键。太大或太小的带宽都会降低模型的估计效果。我们没有使用众所周知的数学规则[43]，而是采用了数据驱动的带宽选择方法，即使用交叉验证的方式选择最大化释然函数的带宽。

为了形象化理解核密度估计方法建模的有效性，我们给出了一个示例用户的建模效果。如图3-1是对示例用户签到位置建模的效果图。图3-1(a)是用户的签到位置，3-1(b)是通过核密度估计得到的概率密度图形，图中不同的颜色表示不同的取值。

(a) Gowalla用户1的签到位置 (b) 核密度建模效果

图3-1 核密度对位置建模的效果示例

从图3-1(a)可以看出，用户的签到位置呈现两个主要的聚簇中心，两个聚簇中心的访问密度也不相同。图3-1(b)中的高斯核密度估计很好的捕获了两个聚簇中心，且两个聚簇中心得到的访问密度有所区别，而对于坐标在(34, -118)区域的聚簇中心又分为了多个访问热点区域。

### 3.1.3计算复杂度

值得注意的是使用高斯核函数对用户活动的地理位置进行密度估计时，计算代价会随着用户签到数量的增长而增大。从计算公式(3.3)和公式(3.4)中可以得出，计算用户*u*访问位置*l*的概率时间复杂度为。当对用户进行POI推荐时，需要计算用户*u*对所有个位置的访问概率，以返回概率最高的top-K个POI推荐给用户，其时间复杂度为*。*当用户访问的位置数量增倍时，使用核密度进行概率密度估计的代价也将会翻倍。

在本文中，使用高斯核函数对用户活动的地理位置进行概率密度估计是可取的，其原因有以下几点：第一，LBSN网络中固有的稀疏性即用户签到的位置仅仅是整个位置集合的一小部分即*。*第二，LBSN网络是一个具有时效性的位置网络，例如当用户工作的地点发生变化时，之前用户工作地点附近的签到信息将无助于地理位置因素的建模。因此，仅仅选取过去一段时间或者数量的签到信息将是合理和明智的。此时的值将处于可控的范围之内。第三，在寻找用户可能喜欢的位置时，我们仅仅考虑了全部位置集中的一部分位置，即且。换句话说，推荐给用户的位置*l* 将从集合中选取而非集合中选取。集合的选取规则参见章节3.2。此时使用核密度估计的时间复杂度为。

## 3.2 基于共现图的推荐模型

基于用户访问类别的自相关性，我们可以假设同一用户所访问的位置之间具有相关性。两位置之间的相关程度由同时访问这两个位置的用户数所决定，同时访问的用户数越多，相关性越强。位置之间的相关性体现了两位置在距离，服务上的相近程度，或者两位置在现实中作为互补品的强弱关系。我们统一使用相似性来描述两位置之间的相近或者互补的关系。假如我们用边将相似节点连接起来，则边的权重将体现节点之间的相似程度，越相似的节点将越有可能被同时访问。另一方面，用户所访问的节点体现了用户的兴趣，因此与用户访问过的节点越相似的节点将越有可能被用户在将来所访问。因此对于用户的推荐过程可以转化为图中搜索相似节点的问题。本节中，我们首先给出共现图的定义以及构建方法，之后提出共现图中节点间的两种相似性，最后给出了基于共现图的推荐模型。

### 3.2.1 共现图

**定义3.1** **POI共现图。**POI共现图图是一个无向带权图，其表示POI之间的共现关系。该网络可以表示为*，*其中图中节点集合, 即 LBSN网络中所有POI的集合；*E*代表网络中的边，边的权重表示POI共现次数， 例如：边表示位置和位置 共现次数为*w*。

POI共现图的构建方法如下： 通过用户访问矩阵*R*, 得到每个用户*u*所访问过的POI列表，对于中的任意的两个POI 和*,* 通过一条无向边相连, 以表示共同出现在一个用户的访问列表中。因此可以得到边中权重*w*的计算公式如(3.5)所示。

(3.5)

其中为指示函数，其当且仅当用户*u*同时访问POI 和时为1，否则为0。

下面给出一个简单的共现图构建示例，用户的访问示例如表3-1所示。

表3-1 简单的访问示例

|  |  |
| --- | --- |
| 用户 | 访问的POI列表 |
|  |  |
|  |  |
|  |  |

由表3-1的用户访问列表构建得到的POI共现图，如图3-2所示。 为了清晰，图中我们使用节点的下标表示该节点，例如1代表节点。

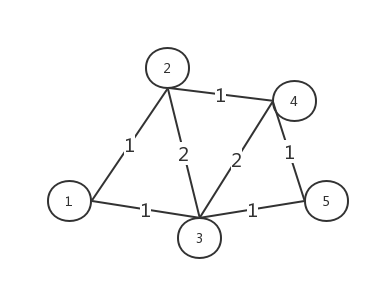


图3-2 POI共现图

如图3-2中所示，节点和之间的边权重为2，其表示共有2个用户同时访问过和。节点和之间的边权重为1，其表示只有1个用户同时访问过节点和。共现图中边的权重是由所有用户的决策所形成的，因为集体的智慧具有更高的可信度，所以我们可以认为相对于节点和，节点和之间具有更强的关联性，因而未来更可能被用户同时访问。

### 3.2.2 共现图中的相似性

基于已构建的POI共现图，我们将定义图中POI节点之间的相似性。POI之间的相似性体现了实际的场景中两POI在服务，位置，价格等方面的相近程度。

**定义3.2 共现概率。**共现概率定义了共现图中，两个POI同时出现的概率。我们使用联合概率表示节点和间的共现概率，计算公式如(3.6)所示。

(3.6)

其中为共现图中节点和间边的权重。其中*W*为POI共现图中所有边的权重和，在此其作为归一化因子。当两节点间无直接边相连时，共现概率为0。

**定义3.3** **一阶相似性**。一阶相似性定义了共现图中，两个直接相连的节点之间的相对邻近程度。我们使用条件概率表示节点相对于节点的一阶相似性。由节点和的共现概率以及贝叶斯公式，我们可以得出一阶相似性即条件概率的计算公式如(3.7)所示。

(3.7)

条件概率定义了已知节点的条件下，到达节点的概率。值得注意的是：条件概率是非对称的，即一般不等于，也就是说节点相对于的邻近程度不等于节点相对于的邻近程度。一阶相似性的不对称主要是由于节点的流行度不同以及所处地理位置的差异所造成的。这种现象在一般的社交网络中同样存在，例如某粉丝关注了一位明星，明星阅读粉丝微博的概率要远远低于粉丝阅读明星微博的概率。

一阶相似性度量了共现图中两个POI节点的直接相似性。该种相似性是由用户的签到数据直接体现而来的。共现图是通过已有的用户签到数据构造而来，当出现新的签到数据时，共现图的权重会发生变化或者出现新的链接。更新的权重将使得已得到的一阶相似性更加精确；新的链接则表示潜在的相似关系被发现。因此，可以认为共现图仅仅是已观察到的关系，大部分节点之间的关系还未被用户挖掘出来。为了得到更好的推荐效果，发现节点和未直接相连的节点之间的相似性是非常有必要的。在共现图中，两个非常相关的节点和可能由于签到信息遗失而并未直接相连，此时因为，所以其一阶相似性为0。因此，仅仅使用一阶相似性很难捕捉这种情形下节点之间的相似性。寻找另外一种相似性方法去度量这种隐含的相似性是非常有必要的。

在网络图中，一个非常自然的直觉是：两个节点间的共同邻居越多，则两个节点越相似。例如在单词共现图中，经常和相同的单词集合一起出现的单词之间具有相似的语义(这个假设被用于自然语言处理领域, 学习得到词的分布式表示，取得了巨大的成功)；在社交网络中，拥有相同朋友的用户之间倾向于具有相同的兴趣爱好，因此未来他们很可能成为朋友；在POI的共现图中，这个直觉同样成立。如图3-2中的POI共现图图, 节点1和节点4拥有共同的邻居2和3，假设用户的访问的位置具有自相关性时，可以得出节点1和节点4之间应该存在着一条边。因此，作为一阶相似性的补充，我们定义了节点间的二阶相似性，以挖掘潜在的相似关系。

**定义3.4 二阶相似性。** 节点对和间的二阶相似性定义为和的邻居结构的相似性。数学上，定义为节点*l*的邻居节点，则节点对和之间的二阶相似性即为和之间的相似性。如果节点和之间没有相同的直接邻居，则节点和之间的二阶相似性为0。

经典的杰卡德(Jaccard )相似度通过计算节点之间共享的节点数除以并集的数量，计算公式如(3.8)所示：

（3.8）

从Jaccard 相似度可以发现，当两节点共享的邻居占比越大，节点之间越相似。而当节点和之间共享的邻居占比越大，从出发通过两步到达的概率越高。因此，我们仍然可以使用条件概率表示不直接相连但是有相同邻居的两节点和之间的相对相似性。二阶相似性的公式定义如(3.9)所示。

（3.9）

其中为节点的直接邻居节点。当和直接相连时，二阶相似性为0。

对于节点间的更高阶相似性可以按照相同的原理计算节点到达目标节点的概率值。然而对于POI推荐问题，寻找与节点更高阶相似的节点是没有必要的，其主要有三个方面的原因：第一，用户的访问类别具有自相关性，具有相关性的节点在共现图中使用直接的边进行连接，可能相关的节点通过相同邻居体现出来，而对于其他情况，说明节点间的相似性低，被用户访问的可能性也低；第二，用户更倾向于访问与其具有相同喜好的用户所访问过的POI。假设由于共同访问过节点，用户和体现出共同爱好，则按照POI共现图的构建原则，从节点出发可以一步到达所有和访问过的节点，因此计算一阶相似度包含了这部分的所有节点；第三，POI推荐场景下，POI具有地理位置属性。当两个节点之间既没有直接连接又没有直接的邻居时，则两节点代表的POI之间的距离可能非常远，因此被用户同时访问的可能性也较低。

### 3.2.3 基于共现图的推荐模型

推荐的目的在于寻找POI集合中，用户感兴趣的POI。而用户所访问过的POI最能反映出用户的兴趣爱好，所以我们将推荐问题转化为在共现图中，寻找与用户已经访问过的POI最为相似的POI。

基于共现图中的相似性度量方法，POI推荐方案如下：

给定用户已访问过的位置集合，用户对于位置的喜好(preference)程度为，可由公式(3.10)计算得到。

（3.10）

其中为相应的权重信息。权重可以解释为用户对于已访问位置的喜爱程度，当用户对于位置越喜爱，那么根据人们爱屋及乌的心里，用户就越有可能访问与位置相近的位置。根据现实中的经验，权重可选的赋值方案有以下几种方式：

（1）按照用户对POI的喜好程度进行赋值，其中喜好程度使用用户对POI的访问频率进行近似。

（2）平等化对待，即对于所有的，令。

（3）按照用户对类别的喜好程度进行赋值，其中类别喜好程度使用用户对于类别的访问频率进行近似。

（4）用户对于该位置的兴趣随时间进行衰减。一个直观的做法是：赋予最近被访问的POI较大的权重，因为其更可能代表用户目前的喜好。

在公式(3.10)中，表示节点与节点之间的相似性。候选节点仅仅选取用户已访问的节点的直接邻居节点，以及从节点出发两步能够到达的节点。对于用户，所有的候选节点构成候选集。当节点与节点之间直接相连时，为一阶相似性，当节点与节点不直接相连但是有相同的邻居节点时，为二阶相似性。

推荐模型最后将候选集中值最高的top-K个节点组成推荐列表返回给用户。

### 3.2.4 相似度的计算与分析

由公式(3.7)可知，节点相对于节点的一阶相似性定义为条件概率。假设使用共现矩阵H存储存储共现图，其中表示节点*i*和节点*j*之间的共现次数; 使用矩阵A表示节点间的一阶相似性, 其中表示节点相对于节点的一阶相似性即；矩阵B表示节点间的二阶相似性，其中表示节点相对于节点的二阶相似性。则一阶相似性矩阵A可以由矩阵H通过行归一化的方式计算得到，即有公式(3.11)。

(3.11)

假设H矩阵每行不为0的元素平均为k个，则利用公式(3.11)计算得到矩阵A的时间复杂度为o(k |L|)。因为矩阵H的稀疏性，使得k<<N，因此计算矩阵A的时间复杂度约为o(|L|)。

根据二阶相似性的计算公式(3.9)可以得到，二阶相似性矩阵B可由一阶相似性矩阵A相乘得到，即有二阶相似性的计算公式(3.12)。

(3.12)

一般来说，两个N阶的方乘法的时间复杂度为o(N3)。但对于B 的计算而言，因为A是稀疏矩阵，所以计算元素的时间复杂度为o(k)，进而计算矩阵B的时间复杂度为o(k|L|2)。

由上面的讨论可知，计算一阶相似性的时间复杂度为线性复杂度，因此当使用一阶相似性的PBG模型将具有较高的时间效率。而当使用二阶相似性，o(k|L|2)的时间复杂度较高，然而以下几个技巧可以化解该方法带来的效率问题：（1）不同区域的POI之间可以认为没有相关性，所以可以将POI按照区域进行划分，每个区域单独进行训练学习POI之间的相似性，这个过程可以很容易的并行化，（2）矩阵的乘法本身可以通过分块的方式并行化的计算。

## 3.3 融合的推荐算法

经过上文的讨论，POI的推荐可以通过在共现图中寻找相似节点以及对位置因素建模两种方法实现。一个直观的做法是分别利用二者构建不同的推荐系统，之后将两个因素推荐得到的结果进行融合。但是我们这里使用相乘的方式得到一个统一的喜好分数，用户对于位置的总体评分可由公式(3.13)计算得到。

（3.13）

其中表示在地理位置因素的影响下，用户访问位置所处区域的概率。表示因为用户喜好，位置被访问的概率。

## 3.4 基于共现图的兴趣点推荐算法

基于共现图的兴趣点推荐算法同时使用位置因素和喜好信息进行建模，算法3.1给出了PBG推荐算法的整体伪代码描述。

|  |
| --- |
| 算法**3.1** 推荐算法 |
| **输入：**网络用户集合，POI集合, 用户访问矩阵*R*， 推荐个数top-K  **输出：**用户的推荐列表S |
| 1. 初始化用户推荐列表S 2. 初始化共现图G=（L, E） //初始化共现图 3. **For** user t in U { //构建共现图 4. //初始化用户历史访问的兴趣点列表 5. **For**in {   6. //将节点和之间的边权重增1  7. }//**End\_For**  8. }//**End\_For**  9. 初始化ERS为矩阵E按行求和得到的矩阵  10. ERS  **/**/ 计算一阶相似性矩阵  11. //计算二阶相似性矩阵B  12.  13. //矩阵C同时保存一阶与二阶相似性  14. **For** user t in U { //对用户进行推荐  15. //初始化用户历史访问的兴趣点列表  16. **If** 一阶 推荐 then  17.  //将赋值为A中节点对应行的和  **18. ELSE**  19. //将赋值为A中节点对应行的和  17. //加入位置因素  17. 赋值为 中得分最高的top-K个元素对应的下标  18. } //**End\_For**  **19. return** S |

在算法3.1中，步骤2到8为共现图的构建过程，步骤9到13为共现图中相似度的计算过程，步骤14到18为用户推荐过程，步骤16到19为利用相似性对用户访问位置的概率进行估计，步骤17为使用核密度的对用户访问位置的概率进行估计，步骤19最后返回用户的推荐结果S。

在推荐算法中，因为用户访问稀疏，所以构建共现图的时间复杂度为, 基于共现图，计算一阶相似性的时间复杂度为，计算二阶相似性的时间复杂度为。因此当使用一阶相似性算法整体的时间复杂度为，使用二阶相似性时的时间复杂度为。

## 3.5 与已有模型的区别

我们提出的PBG模型针对POI问题中的位置，以及用户喜好因素进行建模。已有众多研究学者利用位置和喜好信息进行对该问题建模，为了清晰的区别我们提出的模型与已有模型的不同，我们在本小节列出我们的模型区别于已有模型之处。总的来说，区别有以下几点：

1. 本文不同于已有从用户的角度出发利用概率图模型对用户的行为进行建模的方法，以及考虑用户之间的相似性进行推荐的方法，本文从POI的角度进行出发，构建POI-POI之间的共现图。因为共现图中边的权重是由用户的集体决策形成的，因此相比之下，该方案更加健壮和有效，能够在一定的程度上解决数据的稀疏性问题。
2. 本文不同于ItemCF的方法在于：本文利用LBSN网络中的签到规律，构建了POI-POI之间的共现图图，提出了一种新的相似度度量方法。新的相似性度量方法是从网络图的角度出发考虑两个节点之间的相对相近性。同时PBG推荐模型将用户访问过的POI的一阶邻居节点和二阶邻居节点中作为候选节点集，相比ItemCF推荐算法而言，能够搜寻更多的候选节点。而PBG模型与ItemCF方法的相同点在于：PBG模型同样从节点之间的相似性出发进行推荐，与ItemC算法一样，具有较高的伸缩性和效率。
3. 本文与已有基于图的推荐算法的不同之处在于：已有基于图的推荐方法利用二分图的方式对推荐问题建模，并通过随机游走的PageRank算法进行推荐，该算法的运行时间复杂度高。相对而言，PBG模型基于POI推荐场景进行了特殊设计，构建了共现图，并基于相似性进行推荐。

## 3.6 本章小结

本章中，我们基于LBSN网络中特殊的规律对POI的推荐问题进行了建模。建模的过程主要分为了两个部分：使用核密度估计的方式对位置因素进行了建模，将POI推荐的过程转化为在POI共现图中进行相似性搜索的过程。之后，我们使用融合的方式将两个因素的建模统一到单一的推荐框架中， 最后给出了整个的推荐算法以及复杂度的分析。在下一章中，我们将针对提出的算法，设计并实施大量的实验以验证算法的有效性。

# **4** [**实验与分析**](#_Toc283328287)

在本章中，我们设计并实施了大量的实验去对比我们提出的PBG模型与一些高性能的推荐算法，包括协同过滤，矩阵分解, 排序，以及相关的POI推荐方法之间的性能差异，之后对本模型中的涉及到的各种因素进行实验研究。本章将首先简要地对实验环境及实验数据的基本情况进行介绍，并介绍验证推荐效果的评价指标。在4.4节，我们将给出实验的方案以及相应的分析结果。

## 4.1 实验环境及数据情况

### 4.1.1 实验环境

本文的实验环境如表4.1所示：

表4.1 程序运行环境

|  |  |  |
| --- | --- | --- |
| 单机 | 硬件平台 | 16G内存，1T外存 |
| 操作系统 | Centos 7 |
| 开发工具 | Pycharm， eclipse |

### 4.1.2 实验数据描述

本文实验所采用的数据集为真实的LBSN网络数据集，分别为：Foursquare[[3]](#footnote-3)网络数据集和Gowalla[[4]](#footnote-4)网络数据集，其中Gowalla数据集由斯坦福大学社交网络分析项目中获得。Foursquare数据集包含了从2009年12月7号到2013年七月21号美国地区用户在Foursquare上发表的签到信息，Gowalla数据集包含了2009年2月4号到2010年10月23号的签到信息。数据集中的每个签到信息包含了用户ID, item ID, 和时间戳，每个Item包含有经纬度的信息，以及类别的信息。Foursquare数据集中总共包含34个Tag信息，Gowalla数据集Tag信息缺失。

在数据的预处理部分，我们首先去除了访问低于3个的用户，以及被访问次数少于1个的Item，之后我们按照时间顺序划分最早的80%的数据作为训练集和验证集，剩余20%的数据作为测试集。利用训练集中的数据，我们生成用户和POI之间的访问矩阵，POI之间的共现矩阵，以及item和Tag之间的共现矩阵，所有的数据预处理代码可见仓库[[5]](#footnote-5)。经过预处理之后，实验中的输入数据详情见表4-2。

表4.2 实验数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 用户数 | POI数量 | 训练集稀疏度 | 测试集用户数 |
| Foursquare | 3224 | 70765 | 0.141% | 556 |
| Gowalla | 36022 | 58446 | 0.097% | 1355 |

## 4.2 实验评价度量

本实验将采用推荐领域经典的评价度量精度（precision）和召回率（recall）对实验的效果进行验证。精度和召回率是信息检索领域两个最经典的度量模型性能的指标。精度用于评估检索出的信息的相关度，召回率用于评估查全率。

本文中将推荐top-K 个POI给用户得到的精度和召回率分别表示为P@K和R@K，其中P@K 表示推荐的K个POI中，用户喜欢的比例；R@K表示检索出的POI占测试集中所有POI的比例。P@K和R@K的定义如公式(4.1)和公式(4.2)所示。

(4.1)

(4.2)

其中表示测试集中被用户所访问的POI集合，表示推荐给用户的top-K个未访问过且未出现在训练集中的POI集合。

## 4.3基准的方法

为了验证我们提出的模型的有效性，本文将和以下几个推荐模型做对比。

1. USG模型[11]，该模型结合地理位置影响力，社交网络，和用户兴趣到协同过滤模型中，为用户推荐POI。
2. GeoSoCa模型[14]，该模型使用自适应的核密度估计对用户签到的地理位置因素进行建模，同时使用幂律分布对用户的朋友关系，以及位置的类别标签信息进行建模。
3. ItemCF模型[3]，ItemCF即基于物品的协同过滤算法，其同样利用物品之间的相似性做为策略进行推荐。实验中，我们选取经典的余弦相似度来衡量物品之间的相似性。
4. WRMF模型[4]，该模型以不同的信任值度量已观察到的数据和未观察到的数据，通过矩阵分解来最小化二次损失函数。
5. BPR模型[6]，BPR模型将推荐问题建模为偏序问题，该模型优化了对于已观察的POI和未观察到的POI之间的相对排序。

## 4.4 实验结果及分析

在本小节中，我们使用精度，召回率等指标对提出的模型在标准POI推荐任务进行了评估。

### 4.4.1 标准推荐的性能

图4-1到图4-4 展示了所有推荐方法在两个数据下的实验性能。实验结果分别取top5，top8，top10，top15，top20，top25以及top30推荐下的精度和召回率。实验中参数取值为用户对于该位置的访问频率，同时仅仅使用节点的直接邻居作为候选节点。后续实验若未作特殊的说明，实验参数一律采用上述默认的设置。从实验结果中，我们可以看出，模型PBG的的推荐效果在精度和召回率两个指标上都优于其他对比的推荐模型。

在Foursquare数据集中，PBG模型的性能要略胜于UG以及WRMF模型的性能，而在Gowalla数据集中PBG模型的性能要远远优于其他两个模型, 这是因为在Gowalla数据集中具有相同喜好的朋友之间的相似度更高。UG模型使用了基于用户的协同过滤算法，同时利用幂律分布对位置因素建模。UG模型使用幂律分布对位置建模的过程隐含的利用了位置的共现关系，但是并未深入的挖掘位置之间的相似关系，而是仅仅通过距离衡量两位置同时出现的概率。PBG模型优于UG模型表明利用POI间的相似性比利用用户的相似性更好。

在Foursquare数据集中，PBG模型的性能优于GeoSoca模型，其原因可能是：线上朋友之间的兴趣相似度非常低，对线上的朋友关系进行建模的益处不大；同时因为Foursquare数据中只包含粗粒度的类别标签时，对类别标签直接建模并不能捕捉到用户真正的兴趣。在Gowalla数据集中，因为标签信息缺失，GeoSoca模型仅仅使用了朋友关系和位置两个因素进行推荐。

图4-1 Foursquare数据集推荐精度

图4-2 Foursquare数据集推荐召回率

图4-3 Gowalla数据集推荐精度

图4-4 Gowalla数据集下的推荐召回率

从图4-1到图4-4的实验结果可知，PBG模型模型的性能要远远优于ItemCF模型的性能。这说明了PBG模型的建模方式，相似性度量方案更适合稀疏的LBSN网络下POI的推荐。

BPR模型对用户和位置通过偏序的方式建模，这能够解决数据稀疏性的问题，但是BPR模型仅仅考虑了用户与已访问的位置之间的偏序关系，而未考虑潜在的位置。 而PBG模型将共现图中节点的邻居节点做为候选节点进行挖掘推荐。因而PBG模型的推荐性能好于BPR模型。

### 4.4.2 权重信息的影响

由公式中(3.10)可以得出，用户对于潜在位置的评分受两方面的影响： (1) 用户已访问过的位置与位置的相似性即的大小, 相似性是由位置的距离和兴趣点提供的服务的相似性两个潜在因素影响的，其属于客观因素。(2)评分同样受参数的影响，权重可以解释为对于已访问的位置的喜爱程度，当用户对于位置越喜爱，那么根据人们爱屋及乌的心里，用户就越有可能访问与位置相近的位置。本小节将针对参数*w*的的取值进行研究分析。我们共将参数*w*的取值设置为四组：(1)用户对于位置的喜爱程度，用户的喜爱程度近似取值为用户对位置的访问频率，我们用PBG\_LOVE表示(2)将参数统一设置为1，即进行无区分性对待，我们使用PBG\_SAME表示。(3)按照用户访问的时间对位置赋予不同的权重，我们使用PBG\_TIME进行表示。具体的方法是：按照某一时间阈值对访问的位置进行划分，对于该阈值之前的访问的位置，设置为0.9即进行访问概率的衰减，而对于之后访问的位置，设置为1。(4)用户对于类别的喜好程度，用户对于类别的喜好程度近似为用户对于类别的访问频率，我们使用PBG\_TAG表示。

图4-5 Foursquare中权重对于精度的影响

图4-6 Gowalla数据集中权重对于精度的影响

从实验结果图4-5和4-6中可以发现一个有趣的现象：模型PBG\_LOVE的值在大多数的情况下推荐的效果要略低于模型PBG\_SAME取得的值。这说明使用用户的访问频率作为喜好信息对于未来进行推荐可能并不是最好的策略，与现实中对应的现象是当用户工作地点或者居住地发生变化时，此时寻找与用户历史的位置附近相似的位置进行推荐就不是合理的。而从图4-5和4-6中我们发现PBG\_TIME模型的效果要略好于PBG\_SAME和PBG\_LOVE，这说明考虑时间因素确实会对用户未来的访问决策产生影响。

另外一个值得注意的结果是：Foursquare数据集中PBG\_TAG模型不仅没有提升模型的推荐精度，反而降低了推荐的精度，而当推荐个数增长时，模型PBG\_TAG与模型PBG\_SAME之间的性能差距在缩小。这可能是由于Foursquare数据集本身造成的问题。Foursquare数据集中的类别标签只是粗粒度的类别标签，不能够真正的反应用户的喜好。例如: “Food”这个标签并不能反映出用户对于” Chinese food”的喜爱。当使用用户对于粗粒度类别标签的喜爱程度作为权重时，会缩小相同类别位置间的差距，因此造成性能的降低是可能的；但是当推荐的top-K增大时，用户最喜爱类别的位置更多的被推荐给用户，此时类别权重的影响在弱化，因此两个模型之间的精度差距在缩小。

通过对权重信息的分析实验，我们可以得出：用户最近一段时间所访问的位置会影响接下来的访问行为，因此利用用户最近时间段的访问行为构造权重信息将更加合理；另外用户对粗粒度类别标签的访问频率不适合作为权重信息，因为粗粒度标签不能真正代表用户的喜好信息。

### 4.4.3 推荐的类别

在4.4.2中，PBG\_TAG模型的效果促使我们对PBG模型推荐的位置的类别进行研究。本节实验的目的是：研究PBG模型推荐的位置类别与用户历史访问的类型间的关系。

首先，我们在Foursquare数据集中使用PBG\_LOVE模型进行推荐，并统计推荐的top-K个位置中，与用户历史访问类别不符的位置的平均个数。实验结果如表4-3所示。

表4-3 推荐与用户历史访问类别不符的平均个数

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| top-K | 5 | 15 | 25 | 30 |
| N | 1.2 | 2.9 | 4.5 | 5.2 |

从表4-3中可以发现，在top30的推荐中模型平均推荐了5.2个新类别的位置，但是大部分位置的类别与用户历史访问的类别相同。为了验证推荐的新类别是否有益于提升推荐的效果，我们对比了剔除推荐中新类别位置的方法PBG\_NO的性能，实验结果如图4-7所示。从图中可以看出，PBG\_NO模型的性能低于PBG模型的性能，因此可知推荐新类别的位置提高了POI推荐的性能。

图4-7推荐的新类别对于精度的影响

### 4.4.4 地理位置与相似性的影响分析

位置因素在用户访问某一具体位置的过程中扮演着重要的角色。本部分我们将通过实验比较PBG模型中对位置因素进行建模对推荐性能的提升效果。为了显示PBG模型中位置因素的效果，我们将原有融合的推荐模型按照因素进行如下划分: 仅仅使用位置因素的模型G（Geographical），仅仅使用节点相似性的模型S（Similarity），以及二者融合的推荐模型SG（Geographical and Similarity）。图4-8是三种模型在两个数据集中不同的参数下得到的推荐精度。

图 4-8 数据集中各因素对于推荐性能的影响

从实验结果中可以发现：融合的推荐模型相比使用节点相似性进行推荐的模型性能较好，但是提升幅度有限。这主要是因为在使用节点的一阶相似度寻找候选节点时，候选节点是用户所访问过的节点的直接邻居。因为用户会喜欢访问周边的节点，所以使用相似度时隐含的考虑了位置的因素。然而，需要指出的是两者并不是等同的。使用核密度估计对位置因素进行建模时，因为使用的是密度的方法，所以其更多的是从区域层面而非具体的位置层面进行建模。当两位置所处的区域相同时，使用核密度进行估计得到的访问概率也是相同的，不具有区分性。但是利用核密度对用户的签到位置进行单独的建模使得用户之间具有很好的区分性，这主要是由于不同的用户在实际的位置访问活动中，具有不同的习惯。

从实验结果中可以发现的另外一个现象是：仅仅使用位置因素进行推荐的效果较差。这说明仅仅假设离用户最近的位置被用户访问的概率就高是不合适的，距离用户经常访问的位置非常近的位置被用户访问的概率也可能非常低。这种现象可以在现实的生活中得到解释。当用户经常访问某一位置时，该位置隔壁的位置相对于用户来说是已观察到的，但是未被用户访问，一个解释是相对于用户来说位置相优于位置。这种现象也已经被Lian等[7]运用到实际的POI推荐模型中去。

### 4.4.5 一阶与二阶相似性

在共现图中，我们提出了两种相似性用以度量图中节点之间的相似程度，即一阶相似性与二阶相似性。二阶相似性的提出主要是为了发现共现图中节点之间潜在的边。本节，我们将通过实验的方式对比一阶相似性和二阶相似性的推荐效果。

如图4-9是PBG模型在两个数据集中分别使用一阶相似性和同时使用一阶相似性与二阶相似性的推荐效果。

图4-9 相似性度量对于推荐效果的影响

从实验结果可以发现，二阶相似性对推荐性能提升非常有限，一个可能的解释是：LBSN网络中受位置因素的影响，相近和相似的节点在POI共现图中更多的是通过直接的方式进行连接，此时利用一阶相似性的候选节点较多，因而一阶推荐性能较高。为了证明这个解释，我们进一步的统计了测试集中的用户未来访问的位置在用户所已经访问的节点的一阶和二阶邻居的中占比情况。实验的统计结果如表格4-4所示，从表中可发现用户未来探索的位置大部分位于已访问位置的一阶邻居中，小部分的位置位于二阶邻居中。因此可以得出用户更倾向于访问与其具有共同访问历史的用户所访问的POI。用户访问的位置在共现图中的分布情况从另外一方面证明了我们模型设计的有效性。

表4-4 用户访问的节点在共现图中的分布情况

|  |  |  |
| --- | --- | --- |
| 数据集 | 一阶邻居占比 | 二阶邻居占比 |
| Foursquare | 48.98% | 20.26% |
| Gowalla | 87.73% | 8.94% |

目前构成的LBSN网络中，Foursquare数据集中用户平均访问的POI个数为99.7，Gowalla数据集中用户平均访问的POI个数为56.6，此时存在充足的一阶邻居作为候选节点推荐给用户，因此利用二阶相似得到的效果也较低。但是当用户已访问节点少，共现网络稀疏时，利用二阶相似性挖掘二阶邻居是非常有必要的。

### 4.4.6 算法效率

在大规模社交网络中，算法的效率至关重要。本小节，我们将通过实验的方法对比PBG模型与已有模型在效率方面的差异。

我们对比PBG模型与已有的POI推荐模型在算法执行上的差异，实验结果如图4-10所示。我们没有与矩阵分解，排序的算法进行对比，因为这些算法在不同的阈值下的运行时间将有很大的不同。

从图4-10中可以看出，在两个数据集中，PBG模型都具有最优的时间效率。GeoSoCa模型融合使用用户关系，地理位置因素，以及类别因素，其计算的时间复杂度为。ItemCF算法关键在于计算Item之间的相似性矩阵，因网络的稀疏性，其时间复杂度为。UG模型在计算用户对于某一位置的得分时，需要计算用户与其他所有用户的相似度，因此其时间复杂度为。 PBG模型的需要计算共现图中节点间的一阶相似性，时间复杂度为，计算二阶相似性的时间复杂度约为。

图4-10各算法在不同数据集上的运行时间

PBG模型并不需要进行模型的训练过程，其主要的时间花费在通过共现图构建一阶相似性，以及通过一阶相似性计算得到用户的推荐列表的过程。

## 4.5 本章小结

本章首先对实验的环境以及数据进行了简单的介绍，然后给出了度量实验性能的评价指标以及实验的对比算法，最后给出了实验的设计方案以及结果分析。针对不同的目的，实验总共分为六个任务，分别是验证PBG模型在标准POI推荐任务上的性能；研究PBG模型中参数对推荐性能的影响；研究PBG模型推荐的位置类别与用户历史访问的类别的关系；研究PBG模型中地理位置因素以及相似性对于推荐性能的影响；研究一阶和二阶相似性的效果，研究PBG模型的效率。实验结果表明了本文所提出的PBG模型具有较好的性能和效率。

# [5 总结与展望](#_Toc283328293)

## 5.1 论文总结

随着最近几年位置社交网络的发展，POI推荐问题成为工业界以及学术界热门的研究方向。针对POI推荐问题的研究也如雨后春笋般涌现。POI推荐问题涉及到用户，时间，地点，类别等诸多因素，这给POI推荐问题的建模带来了巨大的挑战。同时POI推荐中的冷启动，稀疏性，以及实时性推荐等也是推荐问题中最棘手难以解决的问题。

通过对已有的研究分析发现，当前已有针对POI推荐的研究都是从用户的角度出发对用户的兴趣直接进行建模，建模的方法大概分为两种：基于用户之间的相似度和基于概率图模型的方式对用户的行为进行建模。基于用户相似度的建模时间开销较大，难以适应大规模的位置社交网络中的实时推荐；同时使用用户相似度时，推荐结果的可解释式性较差。基于概率图模型的建模方式受数据稀疏性的影响较大，当用户的签到数据较少时，难以准确的对用户的行为习惯和兴趣进行建模。针对以上建模方式的缺点，本文的研究重点是从POI的角度出发，利用POI之间的相似性给出一个高性能的推荐模型。本文的主要工作大致如下：

1. 针对POI推荐问题的国内外研究现状做了深入的调研，分析和总结。根据分析，发现已有的研究基本都从用户的角度出发对POI推荐问题进行研究。用户角度建模会有伸缩性差，健壮性不强等缺陷。因此，本文展开了针对POI推荐问题的研究。
2. 针对真实的LBSN网络中的签到数据进行分析，分析从POI的类别，用户之间的相似性，以及用户签到的地理位置三个方面出发。通过分析得到了不同于传统推荐场景的特殊规律。
3. 根据分析得到的用户签到规律，首先利用核密度对用户签到的地理位置因素进行了建模，之后提出了一种新的处理用户签到数据的模型­——共现图， 并在共现图的基础之上，提出了两种度量节点之间相似性的方式：一阶相似性与二阶相似性。最后，提出了融合的推荐框架和算法的时间复杂度分析。
4. 在Foursquare和Gowalla两个真实数据集上与已有的POI推荐算法进行了充分的对比试验。实验中采用了推荐领域经典的度量指标：精度和召回率对推荐的效果进行衡量。实验从：推荐的效果，推荐模型中权重的影响，地理位置以及相似性对于推荐效果的影响，共现图中两种相似性对于推荐性能的影响，以及模型的效率几个方面进行。实验结果表明了我们提出的模型的有效性。

## 5.2 论文展望

鉴于本人学术水平和时间所限，论文中还存在着需要进一步探讨和完善的地方，以下几点是未来工作的主要出发点：

1. 尽管本文提出的模型从POI的角度出发，通过构建共现图，以及度量图中两节点之间的相似性进行推荐得到了较好的推荐效果，但是本文并未进一步的细分POI之间的相关性。POI之间的相关性其实进一步可以分为顺序性相关和类别相关。顺序性相关指某一POI经常在另一类型的POI之后访问，例如人们经常吃完饭看电影。类别相关是指两个POI是因为提供的服务相当而体现出相关性。针对POI中存在的这两种相关性，本文将在未来进行深入的研究。

2. 本文的工作主要集中于标准的POI推荐问题，但是现实的生活中，智能实时的POI推荐系统将更有用武之地。当用户通过LBSN网络搜索食物时，早上和晚上用户的需求可谓天差地别。因此在现有工作的基础之上，研究时间敏感的推荐算法也是一个重要的方向。

3. 本文提出的模型并未考虑到文本信息，这主要受限于能够得到的数据。然而用户签到的文本信息在隐式推荐具有重要的地位，其能够反映出用户对于此次活动的满意程度，以及其他额外的信息。用户的满意程度可以用于修正共现图中边的权重信息，提升模型推荐的效果。

最后，由于本人能力有限，文中未免会出现一些错误或纰漏，本人愿意接受老师和同学们的悉心指导和批评指正。

# 致 谢

时光荏苒， 三年硕士研究生的学习生涯转眼就要结束了。三年前作为刚从本科毕业的大学生，带着对学术的热情和向往，我来到了华中科技大学，来到了智能分布与计算实验室。研究生的三年是我人生中重要的三年，三年的时光充满了酸甜苦辣，三年我付出了很多，更收获了亲情，友情，以及知识等无价之宝。借此机会，感谢陪我一起度过美好时光的每位尊敬的老师和亲爱的同学，感谢你们的帮助，感谢这段时间有你们的陪伴。

在华中科技大学智能分布于计算实验室学习的三年，是我学术生涯的启蒙期。三年来，我逐步学会了如何学术搜索，查阅文献资料，阅读学术论文，做学术报告，学术论文的写作，学术规范，以及科研基金的写作。这些知识为我日后的科研生涯奠定了重要的基础。这最应该感谢的我导师李玉华老师。李玉华老师对于学术研究具有极大的热情，这在很大程度上影响我。同时，不管在研究方向上的建议，以及研究中的困难和生活中的困难，李老师给予了我良好的建议和及时的帮助。在科研过程中， 李老师也能够经常检查询问我的想法，以及实验的进度，和效果。在论文写作的过程中，李老师经常督促我，并认真仔细的帮我修改论文中的不足，终使我及时有质量的完成学术论文的写作。在这里，由衷的感谢李玉华老师三年来为我所做的一切。

感谢实验室主任李瑞轩老师给我们的严格要求，以及为我们创造的良好学术氛围。

李瑞轩老师经常请国外的专家和学者来实验室交流和访问， 这使我能够与优秀的专家学者面对面交流自己的科研想法，以及学习到国外不同的科研工作方法，这让我受益匪浅。 同时李瑞轩老师敏锐的学术眼光和严谨的学术态度深深的感染了我。同时我也要感谢章衡老师老师对博士研讨会所作的努力，这让我学习到很多知识。感谢文坤梅老师和辜希武老师，感谢他们对我学术生涯的帮助。

感谢我的师兄牛进宝，张德文和王凯在学术和生活上对我的帮助和建议。感谢我的同学杨琪，周李鹏，刘其磊，王格，黄凤玲，李晶晶，黄奕伟对我学习和生活中的帮助。感谢学弟袁请亮和梁天安以及学妹徐明丽的帮助。

特别感谢我的家人和朋友常年对我的支持和理解! 他们是最爱我的人，也是我亏欠最多的人，他们默默的奉献是我求学深造来的支持和动力，谢谢。

最后，由衷地感谢各位答辩评委们对我提出的宝贵意见，真心感谢老师们对我论文的悉心指正。

# 参考文献

[1] Koren Y. Collaborative filtering with temporal dynamics. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009. 447-456

[2] Herlocker J L, Konstan J A, Borchers A, et al. An Algorithmic Framework for Performing Collaborative Filtering. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieva, 1999. 230-237

[3] Sarwar B M, Karypis G, Konstan J A, et al. Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web, 2001. 285-295

[4] Hu Y, Koren Y, Volinsky C. Collaborative Filtering for Implicit Feedback Datasets. In: Eighth IEEE International Conference on Data Mining, 2008. 263-272

[5] Gopalan P, Charlin L, Blei D M. Content-based recommendations with Poisson factorization. In: Advances in Neural Information Processing Systems, 2014. 3176-3184

[6] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian Personalized Ranking from Implicit Feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, 2012. 452-461

[7] Lian D, Zhao C, Xie X, et al. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014. 831-840

[8] 余永红, 高阳, 王皓. 基于Ranking的泊松矩阵分解兴趣点推荐算法. 计算机研究与发展, 2016(08):1651-1663.

[9] Cheng C, Yang H, King I, et al. Fused matrix factorization with geographical and social influence in location-based social networks. In: Proceedings of the Twenty-Sixth Conference on Artificial Intelligence, 2012.

[10] Liu Y, Wei W, Sun A, et al. Exploiting Geographical Neighborhood Characteristics for Location Recommendation. In: ACM International Conference on Conference on Information and Knowledge Management, 2014. 739-748

[11] Ye M, Yin P, Lee W C, et al. Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011. 325-334

[12] Yuan Q, Cong G, Ma Z, et al. Time-aware point-of-interest recommendation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013. 363-372

[13] Zhang J D, Chow C Y. iGSLR: personalized geo-social location recommendation:a kernel density estimation approach. In: ACM Sigspatial International Conference on Advances in Geographic Information Systems, 2013. 334-343

[14] Zhang J D, Chow C Y. GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015. 443-452

[15] Zhao S, King I, M L. Capturing Geographical Influence in POI Recommendations. In: International Conference on Neural Information Processing, 2013. 530-537

[16] Li X, Cong G, Li X, et al. Rank-GeoFM: A Ranking based Geographical Factorization Method for Point of Interest Recommendation. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015. 433-442

[17] Liu B, Fu Y, Yao Z, et al. Learning geographical preferences for point-of-interest recommendation. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013. 1043-1051

[18] Wang W, Yin H, Chen L, et al. Geo-SAGE: A Geographical Sparse Additive Generative Model for Spatial Item Recommendation. In: Proceedings of the 21th ACM International Conference on Knowledge Discovery and Data Mining, 2015. 1255-1264

[19] Gao H, Tang J, Liu H. Exploring Social-Historical Ties on Location-Based Social Networks. In: Proceedings of the Sixth International Conference on Weblogs and Social Media, 2012.

[20] Gao H, Tang J, Liu H. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks. In: ACM International Conference on Information and Knowledge Management, 2012. 1582-1586

[21] Li H, Ge Y, Hong R, et al. Point-of-Interest Recommendations: Learning Potential Check-ins from Friends. In: The ACM SIGKDD International Conference, 2016. 975-984

[22] Ye M, Yin P, Lee W C. Location recommendation for location-based social networks. In: ACM Sigspatial International Symposium on Advances in Geographic Information Systems, 2010. 458-461

[23] Zhang J D, Chow C Y, Li Y. LORE: exploiting sequential influence for location recommendations. In: ACM Sigspatial International Conference on Advances in Geographic Information Systems, 2014. 103-112

[24] Salakhutdinov R, Mnih A. Probabilistic matrix factorization. In: Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, 2007. 1257-1264

[25] Cho E, Myers S A, Leskovec J. Friendship and mobility:user movement in location-based social networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011. 1082-1090

[26] Gao H, Tang J, Hu X, et al. Exploring temporal effects for location recommendation on location-based social networks. In: Proceedings of the 7th ACM conference on Recommender systems, 2013. 93-100

[27] Liu Y, Liu C, Liu B, et al. Unified Point-of-Interest Recommendation with Temporal Interval Assessment. In: The ACM SIGKDD International Conference, 2016. 1015-1024

[28] Zhang J D, Chow C Y. TICRec: A Probabilistic Framework to Utilize Temporal Influence Correlations for Time-aware Location Recommendations. IEEE Transactions on Services Computing, 2016, 9(4): 633-646.

[29] Zhao S, Zhao T, Yang H, et al. Stellar: Spatial-Temporal Latent Ranking for Successive Point-of-Interest Recommendation. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016. 315-322

[30] Cheng C, Yang H, Lyu M R, et al. Where you like to go next: successive point-of-interest recommendation. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, 2013. 2605-2611

[31] Feng S, Li X, Zeng Y, et al. Personalized ranking metric embedding for next new POI recommendation. In: International Conference on Artificial Intelligence, 2015. 2069-2075

[32] He J, Li X, Liao L, et al. Inferring a Personalized Next Point-of-Interest Recommendation Model with Latent Behavior Patterns. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016. 137-143

[33] Gao H, Tang J, Hu X, et al. Content-Aware Point of Interest Recommendation on Location-Based Social Networks. In: Proceedings of the Twenty-Ninth Conference on Artificial Intelligence, 2015. 1721-1727

[34] Hu B, Ester M. Social Topic Modeling for Point-of-Interest Recommendation in Location-Based Social Networks. In: 2014 IEEE International Conference on Data Mining, 2014. 845-850

[35] Lian D, Ge Y, Zhang F, et al. Content-Aware Collaborative Filtering for Location Recommendation Based on Human Mobility Data. In: 2015 IEEE International Conference on Data Mining, 2015. 261-270

[36] Yang D, Zhang D, Yu Z, et al. A sentiment-enhanced personalized location recommendation system. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media, 2013. 119-128

[37] Yin H, Cui B, Sun Y, et al. Lcars: A spatial item recommender system. ACM Transactions on Information Systems (TOIS), 2014, 32(3): 11.

[38] Yin H, Zhou X, Shao Y, et al. Joint Modeling of User Check-in Behaviors for Point-of-Interest Recommendation. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015. 1631-1640

[39] 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(01):28-51.

[40] Xie M, Yin H, Wang H, et al. Learning Graph-based POI Embedding for Location-based Recommendation. In: ACM International on Conference on Information and Knowledge Management, 2016. 15-24

[41] Feng S, Cong G, An B, et al. POI2Vec: Geographical Latent Representation for Predicting Future Visitors. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017. 102-108

[42] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality. In: Advances in neural information processing systems, 2013. 3111-3119

[43] Walters R M. Density estimation for statistics and data analysis. CRC press, 1986. 296-297

# 附录1 攻读硕士学位期间投稿的论文

[1] ××××××, et al. Multi-Topic Tracking Model for dynamic social network. Physica A: Statistical Mechanics and its Applications, 2016, 454(1): 51-65.

# 附录2 攻读硕士学位期间参与的项目

[1] 基于FPGA的大数据挖掘系统

[2] 大数据分析与挖掘平台

[3] 面向大数据的社交网络话题和群组驱动的推荐技术研究，国家自然科学基金面上项目，项目编号：61572221

[4] 基于话题和群组驱动的在线社交网络推荐技术研究，华中科技大学校自主创新基金

1. CDF是对[连续函数](http://baike.baidu.com/item/%E8%BF%9E%E7%BB%AD%E5%87%BD%E6%95%B0)，所有小于等于随机变量x的值，其出现概率的和. [↑](#footnote-ref-1)
2. 自相关性指一个随机过程中两个不同时间点的相关性。 [↑](#footnote-ref-2)
3. https://sites.google.com/site/dbhongzhi/ [↑](#footnote-ref-3)
4. <http://snap.stanford.edu/data/index.html> [↑](#footnote-ref-4)
5. <https://github.com/zjlearn/PBG> [↑](#footnote-ref-5)