分 类 号 学 号 M201472821

学校代码 10487 密 级



学位论文

基

|  |  |
| --- | --- |
| 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**

**Multi-Topic Influence Maximization in Large Scale Social Networks**

|  |  |
| --- | --- |
| **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

**关键词：**推荐，可伸缩性，

# Abstract

**Key words:**

# 目 录

# [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获得实时位置信息。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), Gowalla等社交网络，。 这些位置社交网络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Point-of-Interest (POI)的位置信息。 LBSN中，用户与POI之间巨大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，进行精确个性化的POI推荐也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题。当一个新的POI或者新的用户出现在系统的时候，LBSN网络中没有该POI的访问者信息，以及用户的历史签到信息，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。本文将将首先分析用户签到中的规律，之后从POI的角度寻找潜在位置，并据此给出推荐模型，以期获得更有效的推荐结果。

## 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， 优秀的POI推荐能够准确的识别用户的偏好，以及当前的场景，并给出用户最优的POI推荐列表。 优秀的

推荐算法可以极大的节省用户的时间，提升用户的体验度； 同时商家能够通过推荐系统发现潜在的顾客。 因此，对于POI推荐问题的研究将有助于现实生活中用户的生活体验以及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。

本文将主要着眼于普通的POI推荐问题的研究。 主要原因是： 在实际的LBSN中，用户求助于位置推荐时，一般会给定类别和关键字进行搜索。对于普通的POI推荐的研究将有助于优化此类场景下的推荐效果。目前已有对普通的POI推荐问题的研究大多是从用户的角度出发，考虑影响用户访问POI的朋友关系，地点，兴趣等因素对该问题进行建模。 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。另外也有使用基于用户的协同过滤算法，同时考虑位置等因素对于用户签到的影响。但是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。

本文旨在对真实的LBSN数据进行分析，发现LBSN网络中存在的固有规律，并从POI的角度出发对该推荐问题进行建模分析。通过从POI的角度入手建模，能够一定程度上解决数据稀疏性的问题，同时使模型更具扩展性和更高的推荐效果。

## 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因素。目前一般的POI推荐研究主要从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同一模型中建模，并给出唯一的推荐结果。 另外还有针对连续POI推荐的研究，该研究更侧重于基于用户最近的签到信息，像用户推荐下一个将会访问的POI。

本章节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的研究进展以及矩阵分解技术在POI推荐的应用，之后按照POI推荐中的位置，用户关系，时间因素进行研究现状的介绍。  
**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]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[2]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[3]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，该模型只通过非零值进行参数优化，所有具有较高的学习效率。研究[4]使用排序的方法解决隐式反馈的推荐问题。POI推荐问题也属于隐式反馈， 但是由于POI推荐问题的位置特性，时间特性等原因使得已有的矩阵分解技术很难运用到POI推荐场景种。Lian等[5] 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。

**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[5-14]尝试利用位置影响力去提升POI推荐的效果。目前对位置影响力的建模主要分为三种方法：位置影响力服从幂律分布，高斯分布和核密度估计模型。Ye等[8]使用幂律分布对位置的影响力进行建模，通过学习得到距离为d的两个POI共现的概率，Ye同时使用了对用户的喜好，位置因素通过协同过滤的方法进行融合建模， 实验得出了用户受地理位置因素的影响很大， 用户的签到位置呈现聚集的现象； Cheng等[6]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[12]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[11]结合地理位置关系，社交朋友关系，以及POI的类别关系建立了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的得到用户对POI的相对分数。使用自适应核密度函数对每个用户的位置影响力单独建模将更具有针对性， 但是同时需要付出较大的计算代价。另一方面，Liu[7]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范数作为正则项学习该位置的潜在向量。Wang等[15]同样考虑了位置的区域兴趣，同时使用金字塔模型来解决稀疏性问题。

**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用户喜好这一假设， 已有研究[6-8, 11, 16-20]利用朋友关系来提升POI的推荐效果。 Ye等[19]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[6]通过将社交关系作为规则项融入到概率矩阵分解模型PMF[21]， 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。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[6, 8, 17, 19]。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，人所处位置不同，其访问同一POI的代价也不同。Li等[18]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[9, 22-25]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[26]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[20, 27-29]利用马尔可夫链对该序列模式进行建模。其中研究[27, 28]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[21]来推荐连续的POI。 Zhang等[20]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[26]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建议信息，对POI的评价信息，用户所拍POI的照片信息，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，用户签到的内容信息可能提供可更加细微的用户喜好信息，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[30-34]都利用用户的评论信息来提升POI推荐的效果。Yang等[33]通过分析用户的评论信息分析用户对该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。

除了上述按照不同因素考虑POI推荐问题的研究方法，也有一些研究从另外的一些角度融合多因素对POI问题进行建模。Yin等[35]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[15]同样使用概率图模型对用户的行为进行建模， 模型考虑了用户的个体兴趣和目标区域的整体兴趣，同时对用户的不同的访问身份进行了区分（本地人， 旅游者），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。 Li等[13]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。

尽管上述的研究成果基本上考虑了POI推荐问题中的各种印象因素，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，同时算法的可扩展性差。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来发现用户可能会喜欢的位置。本研究的主要目标是提出一个更健壮，可扩展，性能较好的POI推荐模型。

## 1.4 论文的主要研究内容

本文以本地POI推荐为研究背景，从POI角度对POI推荐问题进行建模，提出一个高效，强健的POI推荐模型， 以给出用户最可能感兴趣的POI。

本文主要的研究内容包括：

1. 从POI角度，对POI的推荐问题进行建模
2. 基于实际的LBSN社交网络中用户于POI的交互信息，发现POI场景中的新现象。

3. 基于构建的POI推荐模型，给出不同场景下的POI推荐方案

4. 利用真实的LBSN网络数据集对所提出的模型和POI推荐算法进行验证，证明算法的高效和可伸缩性。

## 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、国内外已有研究现状进行总结，并且概括地介绍了所要研究的内容。

第2章，多话题敏感的传播模型。针对现有模型的不足，结合多话题、位置信息等因素，提出了新的话题敏感的传播模型MTSIC。

第3章，多话题敏感的影响力最大化算法。根据所提出的多话题敏感的传播模型的特点，改进经典贪心算法以突出多话题因素在传播过程中的重要作用。

第4章，基于SPARK的并行算法。针对影响力最大化算法在大规模网络中运行的效率低下的问题，利用SPARK的基于内存计算的机制，提出了基于SPARK的并行算法。

第5章，实验与分析。对实验环境及算法评估度量进行说明，并且从多个角度对所提出的传播模型和影响力最大化算法进行了细致的分析。

第6章，总结与展望。对本课题的工作进行概括总结，针对新的影响力最大化算法及传播模型的缺点提出未来可能进行改进的地方。

# 2 多话题敏感的影响传播模型

推荐模型是推荐问题中的核心部分，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网络中的用户交互信息探索数据集中存在的规律，并进一步提出基于概率的，强健的POI推荐模型。

## 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。** POI指LBSN网络中对应的实体商家或者地点，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。例如某一购物广场可能集合了吃饭，电影等功能。为叙述方便起见，本文接下来将混合使用POI和位置两种术语指代POI。

**定义 2.2 用户活动。** 用户活动是一个的二元组, 其表示用户*u*访问了位置。用户活动通过LBSN网路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。**用户通过关注，互粉等行为在LBSN网络中结成朋友关系*f*。本文使用符号表示用户*u* 的朋友集合。 LBSN网络中朋友之间可以相互查看活动，以及签到信息。

**定义2.4 LBSN网络。**LBSN即基于位置的社交网络**。**POI推荐场景下的基于位置的社交网路是一个用户和位置*l*组合成的异质网络，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，以及用户和位置之间的访问关系*R*； 以及位置于地理坐标和类别标签之间的所属关系；*T*代表的是网络中所有的标签集合。为了便于理解，图2-1给出了简单的LBSN网络场景图。

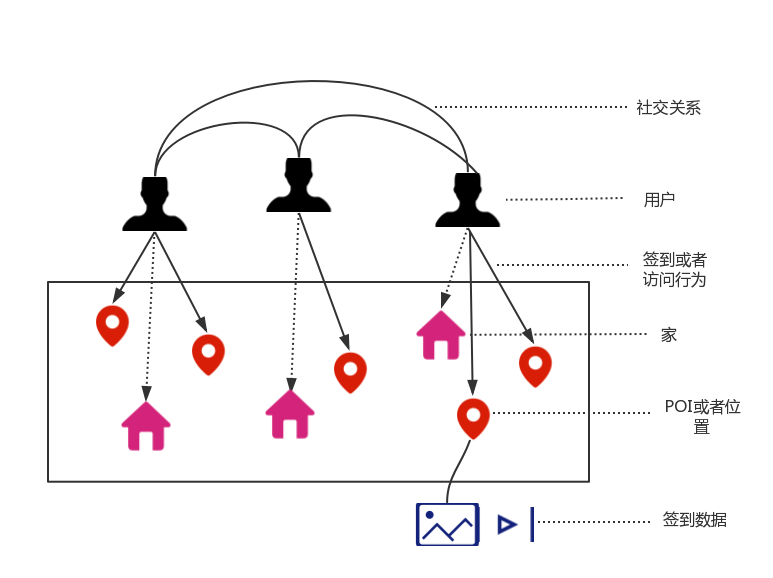
****

图2-1基于位置的社交网络

本文模型所使用的数据集D包含了四个元素： 用户，位置，以及位置的Tag属性, 例如其中， ， 。

为简明清晰，表2-1列出了本文所有数学符号的定义，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 |

为了理解方便，我们使用表示用户*u*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*i*表示：矩阵第i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*i*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。

## 2.2 推荐问题定义

**POI推荐问题**： 给定LBSN网络中的历史签到信息（包含经纬度以及类别信息），POI推荐的目标是：预测用户*u*对于未访问过的位置*l*的评分，并返回top-K个最优评分的位置推荐给用户*u*。

## 2.3 POI中因素的分布规律

为了更好的理解LBSN社交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见章节4.1。数据分析将分别从POI的类别，用户之间的朋友关系，以及地理位置因素三个方面进行。

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，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。通过分析用户所访问过的POI的类别标签可以得到用户对于类别的偏好。 图2-2中是对于Foursquare和Gowalla中签到信息的标签信息的分析图， 图2-2(a ) 是对签到信息的类别进行统计得到的柱状图，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”。

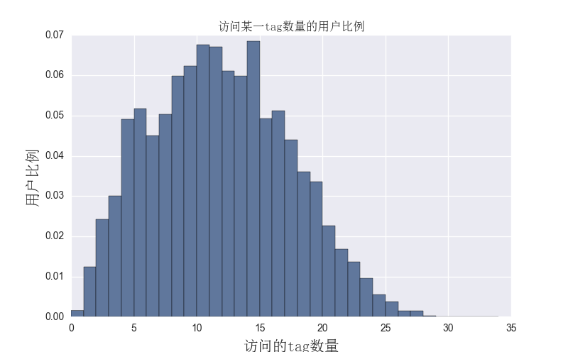
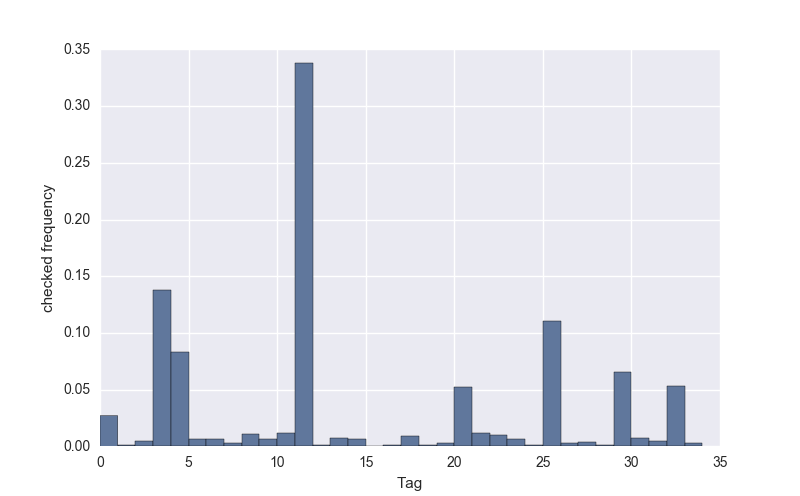


图2-2（a）Foursquare签到信息的类别统计 图2-2(b) Foursquare用户签到的类别统计

图2-2(b) 是对用户所访问的类别数进行统计得到的直方图，从图中可以得出每个用户平均会访问10个类别的POI，这说明用户的兴趣是广泛而非单一的。\*\*\*\*

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时间的前后顺序排列，并对其五等分；之后对每等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各个时间段的兴趣占比变化图，如图2-3(a )所示，以及统计得到的每个用户前后两个时间段的兴趣相似度，如图2-3所示。我们使用了余弦相似度度量用户前后兴趣的相似性，并且在实验中过滤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。

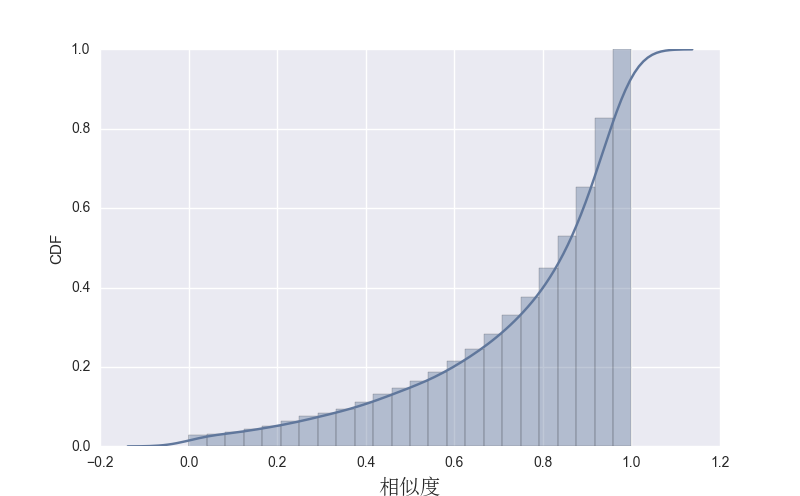
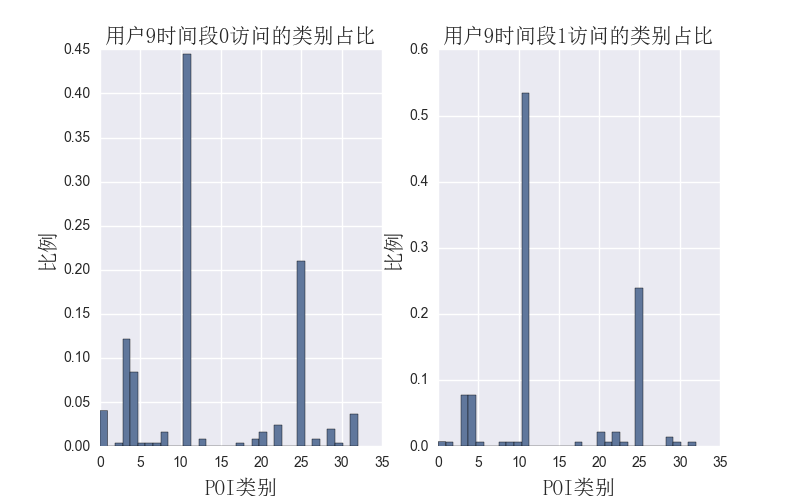


图2-3(a)用户9先后两个时间段访问类别比占 图2-3(b) 所有用户前后两个兴趣的相似度

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，且用户兴趣程多元化发展，同时个人兴趣随时间的变化幅度相对较小。\*\*\*\*\*\*\*，研究方案

**2.3.1 朋友关系**

在LBSN网路中，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。文章中【16】将朋友关系进一步的划分为了线上社交关系形成的朋友， 访问过相同位置形成的朋友关系，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类别标签两种因素所促成的朋友关系。2-4中，给出了线上朋友关系之间, 以及基于相同喜好而形成的朋友之间的相似度。 这里，我们使用预先相似度的计算朋友之间的相似性，同时给出了累计分布函数（cumulative distribution function、CDF）[[1]](#footnote-1)值。

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；具有相同喜好的朋友之间的相似度也非常低，相似度的均值大概在0.03左右。

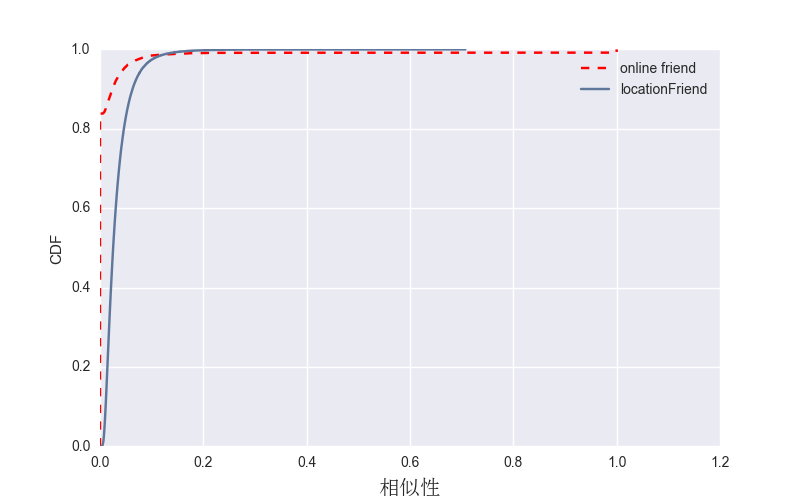


图 2-4 朋友之间的相似性

朋友之间较低的相似性是由LBSN网络中稀疏的数据，以及用户访问量的不均衡性造成的。

我们将使用图模型对该种潜在的POI关系进行建模。

我们进一步的分析朋友关系的家庭位置，从中可以看出，朋友的位置非常

2.3.3地理位置因素

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，我们将用户的签到位置信息隐射到地图中。 这里我们随机选取了Foursquare数据集中签到数量较多的两个用户，如图2-5所示。 从图中可以看出，用户在地理位置上的签到行为程多中心聚簇的现象。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。 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。因此考虑某一位置的POI被用户访问的可能性，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，所以用户5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。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。如用户5访问的C点通过谷歌地图可以发现，其对应于美国峡谷国家公园。

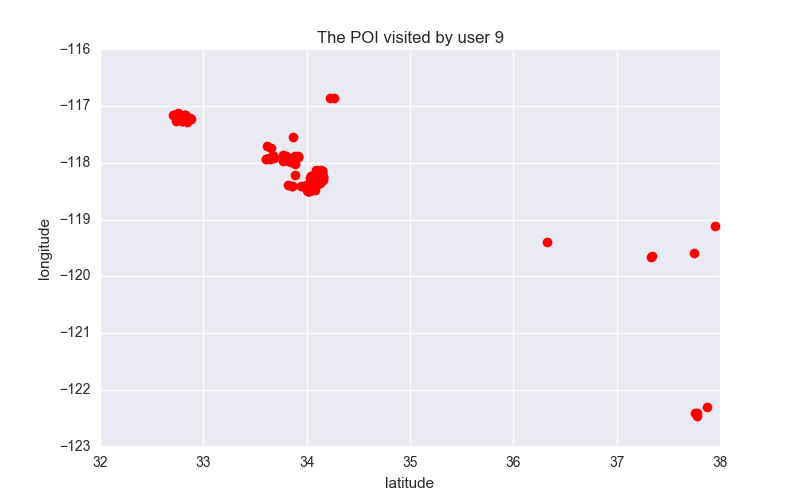
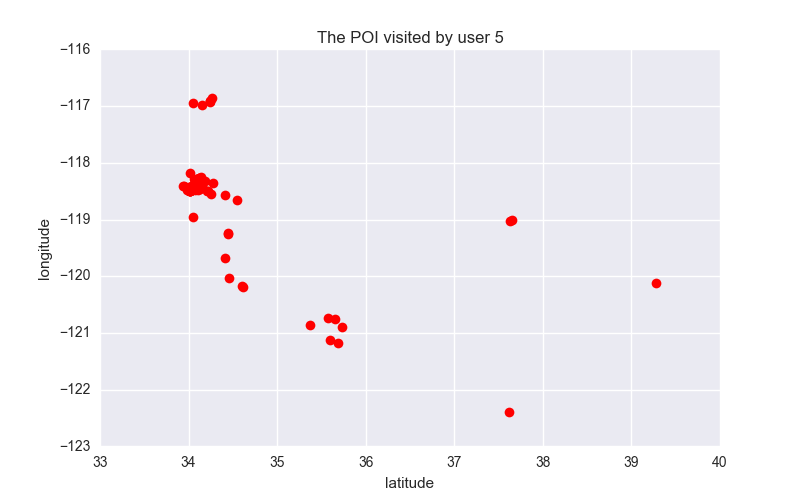


图2-5 Foursquare中两个用户的签到位置

总的来说，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， 用户关系，以及地理位置因素三个方面分析了用户的行为规律。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效的POI推荐模型。

发现异常点，剔除异常点之后的推荐效果会怎么样。

# 第三章 基于图的相似性搜索模型

LBSN网络中用户访问POI的行为呈现出与传统社交网络不同的现象。 POI的类别，地理位置，及用户的社交关系在其中都呈现不同的规律。本章将基于这些潜在的规律对POI推荐问题进行建模。 本章将分别对各个因素：位置因素， 用户喜好，进行建模， 之后给出POI推荐的总体模型。

**3.1 潜在位置发现**

POI推荐中固有

基于已有的相似性度量的方法

基于第二章中用户的签到规律中可以得出：位置因素，用户喜好，以及类别标签信息共同影响用户的访问决策。

已有的研究【】通过计算用户之间的相似性，来推荐与用户最相似的朋友所访问的过的POI。但是LBSN网络中，稀疏的数据会影响寻找与用户相似的朋友，同时算法的扩展性差， 当LBSN网络中用户 针对如下图所示的情景：

因为用户更偏向于访问和其具有相同爱好的用户所访问的物品，因此这些用户所访问过的POI集则可以作为推荐的候选集合。推荐问题转化为从候选集中找出Top-K个POI推荐给用户。

**地理位置因素建模**

通过第二章中的地理位置因素， 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。 基于这些规律， 我们使用无参的概率密度估计方法--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估计用户地理位置因素的方案。

**核密度估计简介**

核密度估计【】是一种非参数化概率密度估计方法。 相比参数化的概率建模方法， 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，而直接从数据中估计潜在的概率密度分布。因而核密度估计方法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， 可以使用如下的函数统计位于区域R内数据点的数量K：

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。

**基于高斯核函数的概率密度估计**

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*概率为：

其中高斯核函数的计算公式如下所示：

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。我们采用了数据驱动的带宽选择方法，而不是众所周知的数学规则【】。我们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概率密度图形。

**计算复杂度**

值得注意的是使用高斯核函数对用户活动的概率进行密度估计时，计算代价会随着用户签到数量的增长而增大。从计算公式×××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N个位置的访问概率，以返回概率最高的K个POI推荐给用户，其时间复杂度为*。*但是使用高斯核进行密度估计是可取的

在本文中，使用高斯核函数对用户活动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, 在发现用户可能喜欢的潜在位置时，我们仅仅选取中的一部分位置，即且*。的选取规则见章节3.2.2。*

**基于共现网络的潜在POI发现**

本章将从POI的角度出发，构建POI的共现网络，并将推荐问题建模为图中的节点的相似性搜索问题。

**定义2.4** POI共现**网络。**POI共现网络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共现网络的构建方法如下： 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, 以表示共同出现过在一个用户访问列表中。因此可以得到边中的*w*的计算公式如下：

其中当且仅当用户*u*同时访问POI 和时为1，否则为0。

例如简单的示例如下：

表 3-1 简单的访问示例

|  |  |
| --- | --- |
| 用户 | 访问POI的列表 |
|  |  |
|  |  |
|  |  |

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标注该节点。

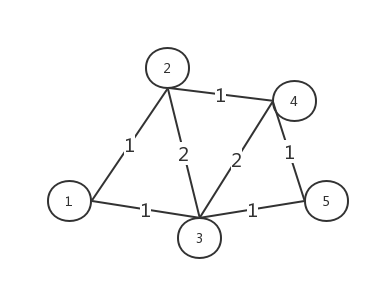


图3-2 POI共现网络示例图

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。 共现网络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。

基于已构建的POI共现网络，我们将定义POI之间的相似度。以用于在推荐模型中，寻找与某一POI相似的POI。

**定义3.2** 一阶相似性。POI共现网络中，两个POI的一阶相似性，定义了两个POI的两个直接相关性。对于POI共现网络中的任何边，*w*的大小是共同访问和的用户数。因此可以认为*w*越大，和越相近。因此可以用来度量节点间的一阶相似性，当节点和之间无直接边相连时，。但为了便于分析和计算我们使用概率来度量节点和之间的一阶相似性。 表示节点和同时出现的概率，计算公式如下：

其中*W*为POI共现网络中所有边的权重和，在此其作为归一化因子。由联合概率，和贝叶斯公式我们可以得出条件概率公式如下：

条件概率定义了已知节点的条件下，得到节点的概率。值得注意的是：条件概率是非对称的，即一般不等于。 这主要是由于POI的流行度不同以及所处地理位置的差异所造成的。

一阶相似性度量了共现网络中两个POI的直接相似性。但是共现网络是我们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，共现网络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，我们可以认为在共现网络中的边仅仅是观察到的一小部分关系，大部分节点之间的关系还未被用户发掘出来。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。在共现网络中，两个可能本来就非常相似，但是因为某些原因却未直接相连的两个节点和，因为，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中节点之间的相似性。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：两个节点拥有越多相同的邻居节点，则两个两个节点越相似。例如在单词共现网路中，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的共现网络中， 这个直觉同样成立。因此，我们将定义节点的二阶相似性，其作为一阶相似性的补充，以挖掘潜在的关系。

**定义3.3 二阶相似性。** 网络中的节点对和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一个节点直接相连，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，计算公式如下：

从杰卡德相似度可以发现，当两节点共享的邻居占比越重，节点之间越相似。另外一个角度可以理解为从出发有较高的概率到达。 因此对于不直接相连的两节点和之间仍使用条件概率定义相对的相似性。 二阶相似性的公式定义如下所示：

对于更高阶的节点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性的节点是没有必要的，其主要有两方面的原因：第一，在2.3.4节中 ， 我们从实际的签到数据中发现，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共现网络的构建原则，从节点出发可以一步到达所有和访问过的节点，因此计算一介相似度包含了这部分的所有节点。第二，POI推荐场景下，POI具有地理位置属性。当两节点没有直接的邻居时，两节点间距离很大， 被用户同时访问的可能性很低

**融合框架**

经过上文的讨论， 可以得出用户喜好和位置因素都可以用来进行POI的推荐。一个直观的做法是分别利用喜好和位置因素构建不同的推荐系统，之后将各个因素推荐的结果进行融合。但是我们这里相乘的方式得到一个统一的喜好分数

其中表示在位置因素的影响下，用户*u*访问位置*l*的概率。表示在用户喜好的影响下，位置*l*被访问的概率。值得说明的是：在这里更像是一个过滤器，其过滤掉因为位置因素而影响用户访问的POI。

**相似性搜索推荐模型：**

在对用户*u*进行推荐时， 目的是寻找POI集合中，用户感兴趣的POI进行推荐。

而用户的兴趣是未知的，只有用户所访问过的POI可以反映用户的兴趣爱好。所以问题转化为在POI共现网络中，寻找与用户已经访问过的POI相近的POI，并将其推荐给用户。

基于已有的相似性度量模型，推荐方案如下：

给定用户*u*已访问过的位置集合， 则用户*u*访问某一位置*w*的可能性得分为：

其中为相应的权重信息。可选的赋值方案有：

（1）按照用户对该POI的喜爱程度（访问频率）

（2）平等化对待，即令

（3）将用户访问的所有POI按照时间段划分，给予不同时间段的的POI不同的权重。一个直观的做法是，赋予较近时间段的物品较大的权重，因为其更大可能代表用户的喜好。

（4）按照用户对类别的喜好程度进行赋值。

**3.4 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，将潜在位置发现的过程建模为在POI共现网络中相似性性搜索的过程。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。在下一章中，我们将针对提出的模型，设计并实施大量的实验以验证模型的有效性。

本文提出的模型区别于已有模型的之处在于：

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之间的共现网络。因为共现网络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之间的共现网络图，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推荐模型仅仅从用户所访问的POI的一阶邻居和二阶邻居中选择候选节点。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。

# 第四章 [实验与分析](#_Toc283328287)

在本章中， 我们设计并实施了大量的实验以对比我们提出的PBG模型与一些先进的POI推荐算法，包括协同过滤，矩阵分解的方法，之后对本模型中的涉及到的因素进行实验研究。实验设计的目的主要是主要是。 本章将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。 之后在5.节给出实验的分析结果。

## 5.1 实验环境及数据情况l

### 5.1.1 实验环境

本文的实验环境如表5.1所示：

表5.1 程序运行环境

|  |  |  |
| --- | --- | --- |
| 单机 | 硬件平台 | 8G内存，1T外存 |
| 操作系统 | Centos 7 |
| 开发工具 | Pycharm， eclipse |

### 5.1.2 实验数据描述

本文实验所采用的数据集为LBSN网路中的真实数据集，分别来源于：Foursquare[[2]](#footnote-2)网络数据集和Gowalla[[3]](#footnote-3)网络数据集，两个数据集均可从网上直接下载。其中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\*\*月到\*\*\*年几月的信息。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数据集中总共包含\*\*个tag 信息。

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[[4]](#footnote-4)。经过预处理之后，实验中的输入数据详情见表5-2。

表5.2 输入数据整体情况

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 用户数 | POI数量 | 测试集稀疏度 | 测试集用户数 |
| Foursquare | 3224 | 70765 |  | 556 |
|  |  |  |  |  |
| Gowalla |  |  |  |  |

## 5.2 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），召回率（recall）和NDCG（normalized discounted cumulative gain ）值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。NDCG是搜索领域的用于度量排序效果的指标。

本文中将POI推荐top-K 个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的定义公式如下：

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。

NDCG指标表示归一化折损累积增益。其由累积增益CG(cumulative gain)经过折算和归一化而得来，其主要评估检索出的排序信息是否和期望的排序信息一致。其计算公式如下：

公式不正确

通过精度和召回率能够验证POI推荐算法推荐POI的相关度，通过NDCG能够验证推荐结果的排序的好坏。

## 5.3基准的方法

为了验证我们提出的模型的有效性，本文将和以下几个模型做对比。

* USG模型[8]，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。
* Geo-SAGE模型[15]，该模型使用概率图模型的方式对用户的日常活动进行建模； 同时为了减少数据的稀疏性，Geo-SAGE使用金字塔模型来减少数据的稀疏性。
* ItemCF模型【】，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作为物品之间的
* PMF模型[21]， 该模型对用户和item之间评分矩阵进行矩阵分解，得到低维空间向量。PMF与已有模型不同之处在于，其假设评分值相对于用户和物品item的低维向量满足高斯分布， 该模型主要用于解决协同过滤算法中的低效率和稀疏性问题 。
* WRMF模型[2]， 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。
* BRP模型[4]， BRP模型将推荐问题建模为偏序问题，该模型优化了对于以观察的POI和未观察到的POI之间的相对排序。
* Cofactor模型[36]， 该模型对评分矩阵利用矩阵分解得到低维向量， 但是同时利用词嵌入的方式对推荐效果进行进一步的增强。该模型也是解决隐式推荐问题。

## 5.4 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。

5.4.1 标准推荐的性能

图5-4 展示了所有推荐方法在两个数据下的实验性能。 实验结果分别为top5, top10, top20推荐下的精度和召回率。从实验结果中，我们可以看出，模型×××的性能总是优于××××模型。

利用item之间的相似进行推荐更加有效的原因，

我们进一步的统计了相关的在测试集中用户喜好的哪些

相比WRMF模型，PBG模型利用集体智慧的方式度量

一部分的原因在于：在POI推荐的场景下，POI的分类和内容相对的的简单，新颖性不高；相比而言，线上的商品或者文章分类更加的精细，地理位置，POI的类别，以及服务的质量因素是三个影响用户选择的最重要因素。同时受到地理位置因素的限制，用户所访问的POI的位置相对比较集中。

5.4.2 权重信息的影响

在公式中\*\*\*， 用户对于潜在的位置*w*的评分受两方面的影响： (1) 用户访问的每个位置与位置*w*的一介相似性与二阶相似, (2)用户对于相似性度量的

5.4.3 位置因素对于推荐效果的影响

位置因素在LBSN网络中，用户访问的过程中扮演着重要的角色。本部分我们将用实验比较PBG模型中对位置因素进行建模的效果。 为了显示PBG模型中位置因素的效果，我们将原有的融合推荐模型划分为两个部分仅仅使用位置因素，仅仅使用节点相似性，以及二者融合的推荐模型。

从实验中可以发现：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提升幅度较小。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一阶邻居。由用户访问的习惯可以知道，

而仅仅使用位置因素进行推荐的效果较差。这说明距离用户经常访问的位置非常近的位置被用户访问的可能性很低。这种也可以在现实的生活中发现，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可见的，但是未被用户所访问因而可以用户对于该位置不感兴趣。这种现象也已经被【】运用到具体的POI推荐中去。

5.4.4 一阶与二阶相似度量

实验中利用二阶相似性对推荐性能提升有限的原因主要有以下几个方面： （1）LBSN网络中受位置因素的影响，相近和相似的节点在POI共现网络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，（2）当网络中数据稀疏时，基于二阶的相似度进行度量时，节点之间的边

但是文本提出节点的二阶相似性的度量方法，我们认为可能在以后的

可能只有当数据存在于一定的规模时，二阶相似度将有助于推荐的效果。

但是经过数据的累积，网络变密之后，二阶的相似性可能并不适用。

5.4.5 效率与可伸缩性

PBG模型相对于与已有的

5.4.6 分中心的推荐呢？

5.4.6 推荐中一些有趣的现象？？？

推荐的策略：

# [6 总结与展望](#_Toc283328293)

## 6.1 论文总结

影响力最大化问题一直以来都是数据挖掘领域中的热门的研究方向，对于影响力最大化问题的相关研究也是从出不穷。然而随着像FaceBook、Twitter等社会网络的快速壮大，已有的影响力最大化的模型与算法已经无法满足当前的需求，更不能达到预期的效果。然而，随着社会网络的盛行，利用“病毒式营销”、“口碑传播”等方式推广商品或事件就有着巨大的商业前景，这就使得影响力最大化问题面对着前所未有的挑战。

通过对已有工作的研究发现，之所以已有研究并不太适用于当前的社会网络的原因主要有两点：第一，绝大多数的已有研究忽略了用户话题的信息，即使有少部分研究考虑了话题元素，但却忽略了任何的商品或事件都不是话题单一的且用户的兴趣也具有多样性；第二，目前的社会网络都是大规模网络，已有研究在处理大规模网络时的效率问题尤为突出。上述出现的问题也就是本文所需解决的问题。因此，本文的主要工作大致如下：

1. 对于影响力最大化问题的国内外现状做深入的了解，根据已有研究分析它们的不足与缺陷，发现已有研究都忽略了任何商品或事件都包含多个话题，因此没有考虑多话题因素。针对这种情况，展开多话题敏感的影响力最大化问题的研究。

2. 基于对于现实世界的5个观察，提出MTSIC模型。该模型融入了多话题因素，同时考虑用户节点的权威度与中心度，使得结果更加符合真实情况。考虑到现实应用往往存在地理位置的限制，将地理位置因素加入其中，使得结果更加准确，更具有针对性。

3. 由于已有影响力最大化算法无法适用于基于多话题的MTSIC模型，因此将多话题因素与经典的贪心算法相结合，提出了ANSM度量，并根据该度量给出了ANS算法。

4. 考虑到当前的社会网络皆为大规模网络，已有的算法在大规模网络下的运行效率较差，而基于内存计算的SPARK计算框架适用于图迭代的特性，因此给出了基于SPARK的并行算法，以解决影响力最大化问题的效率问题。

5. 在DBLP和Twitter真实数据集上进行不同传播模型与影响力最大化算法的对比实验，使用最后激活节点数与本文所提出的新的度量分别对不同模型与算法进行分析，给出新模型MTSIC与算法ANS的性能优势。通过分析在分布式平台的与单机运行所花费的时间，给出所提出的并行算法PMT的效率优势。

## 6.2 论文展望

鉴于本人学术水平和时间所限，论文中还存在着一些需要进一步完善和探讨的地方，以下是几点未来可以着手改进的地方：

1. 尽管本文所提出的模型将多话题、地理位置、权威度和中心度等因素融入其中，但该模型依然是静态的。事实上，随着时间的推移，用户节点的兴趣和网络结构也在随之改变，因此考虑动态因素就成为一个需要研究的方向。

2. 本文是利用Topical HITS算法来计算用户节点的权威度和中心度，但已有研究有很多其他可选方法来进行模型的参数化学习。以后可以研究其他更有效地方式来估计这些参数值，使得模型更准确且更贴近现实。

3. 由于本文仅仅将多话题因素融入到经典的贪心算法中，而并未考虑其他高效的先进的算法，也并没有利用启发式、剪枝等策略，因此算法的效率依然很低，仅仅单纯满足了模型的需求。未来可以研究更加高效的算法，来提升运行效率。

4. 本文所提出的模型和算法只是作为研究，并未上升到实际应用。因此可以设计一个平台，去响应商品或事件的影响力最大化的请求，并记录结果。对于相似商品或事件的请求，可以将已经计算得到的之前相似的商品或事件的结果返回，避免重复计算，节约运行成本。

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

# 附录1 攻读硕士学位期间投稿的论文

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

**参考文献**

[1] Koren Y. Collaborative filtering with temporal dynamics, 2009[C].

[2] Hu Y, Koren Y, Volinsky C. Collaborative Filtering for Implicit Feedback Datasets, 2008[C].

[3] Gopalan P, Charlin L, Blei D M. Content-based recommendations with Poisson factorization[J]. Advances in Neural Information Processing Systems, 2014,4:3176-3184.

[4] Rendle S, Freudenthaler C, Gantner Z, et al. {BPR:} Bayesian Personalized Ranking from Implicit Feedback[J]. CoRR, 2012,abs/1205.2618.

[5] Lian D, Zhao C, Xie X, et al. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation, 2014[C].

[6] Cheng C, Yang H, King I, et al. Fused matrix factorization with geographical and social influence in location-based social networks, 2012[C].

[7] Liu Y, Wei W, Sun A, et al. Exploiting Geographical Neighborhood Characteristics for Location Recommendation, 2014[C].

[8] Ye M, Yin P, Lee W C, et al. Exploiting geographical influence for collaborative point-of-interest recommendation, 2011[C].

[9] Yuan Q, Cong G, Ma Z, et al. Time-aware point-of-interest recommendation, 2013[C].

[10] Zhang J D, Chow C Y. iGSLR: personalized geo-social location recommendation:a kernel density estimation approach, 2013[C].

[11] Zhang J D, Chow C Y. GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations, 2015[C].

[12] Zhao S, King I, Lyu M R. Capturing Geographical Influence in POI Recommendations[M]. 2013.

[13] Li X, Cong G, Li X, et al. Rank-GeoFM: {A} Ranking based Geographical Factorization Method for

Point of Interest Recommendation, 2015[C].

[14] Liu B, Fu Y, Yao Z, et al. Learning geographical preferences for point-of-interest recommendation, 2013[C].

[15] Wang W, Yin H, Chen L, et al. Geo-SAGE: {A} Geographical Sparse Additive Generative Model for Spatial

Item Recommendation, 2015[C].

[16] Gao H, Tang J, Liu H. Exploring Social-Historical Ties on Location-Based Social Networks[J]. 2012.

[17] Gao H, Tang J, Liu H. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks, 2012[C].

[18] Li H, Ge Y, Hong R, et al. Point-of-Interest Recommendations: Learning Potential Check-ins from Friends, 2016[C].

[19] Ye M, Yin P, Lee W C. Location recommendation for location-based social networks, 2010[C].

[20] Zhang J D, Chow C Y, Li Y. LORE: exploiting sequential influence for location recommendations, 2014[C].

[21] Salakhutdinov R, Mnih A. Probabilistic matrix factorization, 2007[C].

[22] Cho E, Myers S A, Leskovec J. Friendship and mobility:user movement in location-based social networks, 2011[C].

[23] Gao H, Tang J, Hu X, et al. Exploring temporal effects for location recommendation on location-based social networks, 2013[C].

[24] Liu Y, Liu C, Liu B, et al. Unified Point-of-Interest Recommendation with Temporal Interval Assessment, 2016[C].

[25] Zhang J D, Chow C Y. TICRec: A Probabilistic Framework to Utilize Temporal Influence Correlations for Time-aware Location Recommendations[J]. IEEE Transactions on Services Computing, 2015(1):1.

[26] Zhao S, Zhao T, Yang H, et al. {STELLAR:} Spatial-Temporal Latent Ranking for Successive Point-of-Interest

Recommendation, 2016[C].

[27] Cheng C, Yang H, Lyu M R, et al. Where you like to go next: successive point-of-interest recommendation, 2013[C].

[28] Feng S, Li X, Zeng Y, et al. Personalized ranking metric embedding for next new POI recommendation, 2015[C].

[29] He J, Li X, Liao L, et al. Inferring a Personalized Next Point-of-Interest Recommendation Model

with Latent Behavior Patterns, 2016[C].

[30] Gao H, Tang J, Hu X, et al. Content-Aware Point of Interest Recommendation on Location-Based Social

Networks, 2015[C].

[31] Hu B, Ester M. Social Topic Modeling for Point-of-Interest Recommendation in Location-Based

Social Networks, 2014[C].

[32] Lian D, Ge Y, Zhang F, et al. Content-Aware Collaborative Filtering for Location Recommendation

Based on Human Mobility Data, 2015[C].

[33] Yang D, Zhang D, Yu Z, et al. A sentiment-enhanced personalized location recommendation system, 2013[C].

[34] Yin H, Cui B, Sun Y, et al. {LCARS:} {A} Spatial Item Recommender System[J]. {ACM} Trans. Inf. Syst., 2014,32(3):11.

[35] Yin H, Zhou X, Shao Y, et al. Joint Modeling of User Check-in Behaviors for Point-of-Interest Recommendation, 2015[C].

[36] Liang D, Altosaar J, Charlin L, et al. Factorization Meets the Item Embedding: Regularizing Matrix Factorization

with Item Co-occurrence, 2016[C].

**校对报告**

当前使用的样式是 [中华人民共和国国家标准\_GBT\_7714-2005]

当前文档包含的题录共68条

有0条题录存在必填字段内容缺失的问题

所有题录的数据正常

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. https://sites.google.com/site/dbhongzhi/ [↑](#footnote-ref-2)
3. http://snap.stanford.edu/data/index.html [↑](#footnote-ref-3)
4. https://github.com/zjlearn/PBG [↑](#footnote-ref-4)