TGTM: Temporal-Geographical Topic Model for Point-of-Interest Recommendation

Abstract: 基于位置的社交网络（LBSN）和微博-博客服务的广泛使用产生了大量用户的登记数据，其包括用户ID，文本内容，发布时间戳，地理信息等。兴趣点（POI）推荐是提供有趣地点的个性化推荐以增强LBSN中的用户体验的任务。在本文中，我们提出了2个新颖的时间位置内容感知POI推荐模型，它们共同整合辅助时间，文本和空间信息，以提高POI推荐的性能。具体地，我们利用时间信息将原始用户-POI登记频率矩阵划分为子矩阵，以便可以对类似时间场景中的行为进行分组。然后，我们利用（LDA）模型和空间坐标来推断POI。使用真实数据集进行的综合实验证明了我们的方法的优越性。

## 1 Introduction

鉴于用户群体的显着增长率和具有定位功能的大量移动设备，基于位置的社交网络服务已经变得非常受欢迎，包括Facebook Places，Google latitude，Twitter和Foursquare等在线网站。LBSNs现在不仅允许用户以“签入”的形式共享他们的物理位置坐标和时间戳，而且还可以就他们访问过的POI撰写文本意见。使用位置日志数据挖掘和建模登记行为是对用户和POI都很有价值，因为POI推荐可以帮助人们发现有吸引力的地方，并可能为POI的所有者培养更多潜在的业务。

为此，POI推荐已成为过去几年的热门研究课题。POI建议中最关键的挑战之一是如何应对用户POI登记频率矩阵的极端稀疏性。此外，与仅考虑用户项目评级的传统建议不同，POI推荐系统要复杂得多，因为在这种情况下的移动行为可能是许多方面的混合。 特别是，当存在诸如评论或微博之类的富文本时，POI推荐应该是个性化的，位置内容感知和上下文依赖。有鉴于此，各种类型的信息构成了另一个重大挑战，即如何系统地以统一的方式将它们纳入其中。

直观地说，用户倾向于在工作日的中午附近的办公室附近的餐馆吃午餐。 但是在周末，用户可能会在晚上长时间访问酒吧以获得乐趣。 我们相信在类似的时间背景下可以更容易地找到移动模式。鉴于此，我们首先基于时间戳将原始用户POI信息划分为子组。然后，我们使用LDA算法[2]来推断用户的兴趣和POI的主题分布。在[4]中，作者发现用户的位移（2个连续登记位置之间的距离）趋势可以用幂律分布近似。考虑到文本和地理影响，在矩阵分解模型的基础上，我们可以推断出用户更喜欢POI的程度。

总之，本文的贡献有三个：

1. 我们从文本内容信息，地理坐标和时间影响的角度研究了用户隐性反馈登记行为与LBSN辅助信息之间的关系。
2. 我们提出了2种用于POI推荐的新概率矩阵分解模型，并且每种模型都包含上述三种类型的辅助信息。
3. 我们通过对从Twitter的API中提取的真实世界LBSN数据集进行综合实验来评估所提出的方法。 结果证明了我们的方法的有效性。

我们有通常的组织：调查，问题定义，提出的方法，实验和结论。

## 2 Related Work

### 2.1 Matrix Factorization

推荐系统中的协同过滤（CF）可以预测未消费项目的个性化偏好[1,19]。 协作过滤方法有两大类。基于邻域的解决方案通过寻找志同道合的用户[5,18]来预测用户的潜在兴趣，这些用户使用相似度函数（例如余弦距离或皮尔逊相关性）来计算相似的用户或项目。基于模型的[9,17]解决方案利用观察到的评级或标签来模拟用户-项目交互。基于模型的各种成功实现基于矩阵分解，其分解用户项目评级矩阵。在本文中，我们采用基于矩阵分解方法的方法，因此我们在此简要介绍。

基本上，矩阵分解（MF）方法将评级矩阵RR分解为一个用户特定矩阵UU和一个项目特定矩阵CC [10]。 然后，可以通过将两个分解矩阵相乘来近似原始评级矩阵，同时使用正则化项避免过拟合：



其中||⋅||F2表示Frobenius范数，Iij是指标函数，如果用户i对电影j进行评级则等于1，否则等于0。常数λ控制正则化的程度以减轻过拟合问题。概率矩阵分解（PMF）[16]给出了正则化的概率解释。

### 2.2 POI Recommendation

POI建议首先在GPS轨迹数据上进行了研究[23]。随着LBSN的日益普及，POI推荐引起了很多关注。有关利用地理影响的POI建议的大量文献。现有研究发现，人们往往会去附近的地方进行相互观察。基于用户的CF和基于项目的CF在[11,20]中进行了研究。Ye等人[20]在基于用户的CF框架下共同探讨了社会和地理因素，以提出POI建议。最近有兴趣利用社会和地理属性来提高POI建议的有效性。在[3]中，他们提出了一种矩阵分解方案，通过高斯混合模型（GMM）结合地理和社会信息。在[6]中作者发现，来自类似用户的偏好对于城镇内用户更为重要，而友谊对于城外用户更为重要。

刘等人[13]提出了一种基于图的方法，以综合方式利用时间和地理信息。 袁等人[21]以基于用户的CF方式结合时间影响，并且候选POI的最终偏好分数通过分别基于时间影响和地理影响计算的分数线性组合。

谈到文本内容信息，高等人[7]研究了LBSNs的内容信息。统一POI建议框架下的POI属性，用户兴趣和情绪指示。Liu和Xiong [12]研究了POI相关标签与聚合LDA模型的影响。Hu和Ester [8]根据主题建模方法调查了Twitter和Yelp的用户兴趣。袁等[22]从空间，时间和内容方面共同模拟个人用户的移动行为。他们用概率生成模型捕获它们。

3 Proposed Approach

3.1 Problem Definition

POI建议的问题是推荐可能对用户具有吸引力的POI。为了便于说明，让u = {u1，u2，...，um}为用户集合，c = {c1，c2，...，cn}为POI集合，其中m和n表示 用户数和POI数。每个用户都具有可观察的属性xi（例如，用户的评论和登记历史）。此外，每个POI保持可观察的属性xj（例如，POI的相关文本描述和经度和纬度方面的空间坐标）。R∈Rm×n是登记频率矩阵，其中每个条目Rij表示由ui在cj处所观察到的登记频率。将LDA模型应用于与所有用户和POI相关的所有文本描述可以帮助我们获得针对POI cj的每个用户ui或πj的主题分布θi。

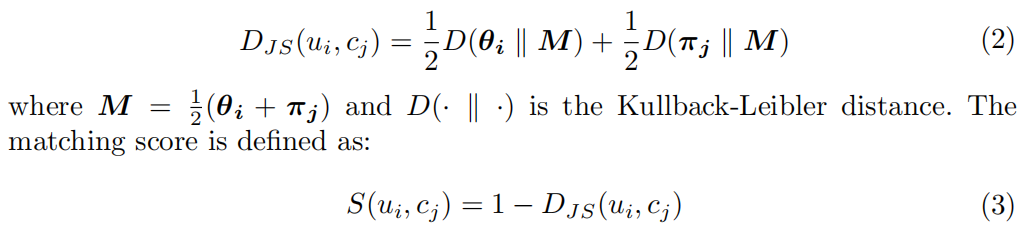
3.2 Time Aware User-POI Subgrouping（时间感知用户-POI子分组）

基于上下文信息对原始用户项目评级矩阵进行分组已被证明是一种改进推荐系统的有前途的方法[15]，因为生成的子组包含具有更高相关性的相似评级。具体地，我们使用时间信息（即，星期几和小时）将原始用户-POI登记频率矩阵划分为4个子组。直观地说，人们的活动在工作日可能更加规律和可预测，而周末则更加多样化和相似。因此，我们首先根据星期几（即工作日与周末）对原始矩阵进行分区。接下来，一天（24小时）可以分为工作段（从上午08:00到下午17:59）和休闲时间（从第二天的下午18:00到下午7:59）。 然后，根据星期几生成的2个子矩阵可以基于时间进一步导出为4。 最后，我们得到4个用户POI登记频率子矩阵，我们可以利用地理和文本信息进行预测。

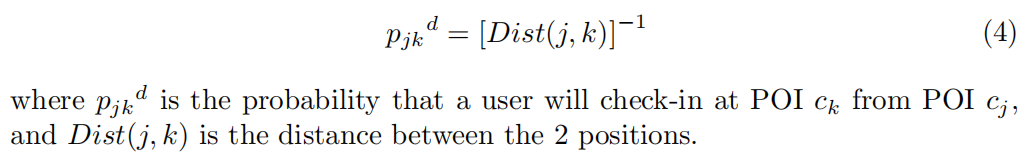
3.3 Exploiting Textual Content and Spatial Information(文本内容与空间信息的开发)

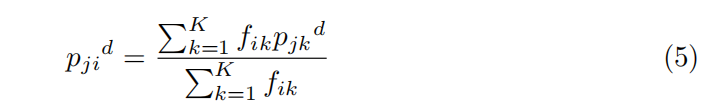
已知LDA算法在诸如短推文之类的短文本文档上具有差的性能。在本文中，我们将与相同POI相关的所有文本注释聚合到POI文档中。我们还将每个用户签入的POI的所有文本注释组合到用户文档中。因此，我们获得了大型文档集，每个文档对应一个POI或用户。

然后利用LDA [2]分析每个文档的主题分布。与[12]类似，我们将用户ui和POI cj之间的匹配分数定义为关于用户的主题分布θi和POI πj的主题分布的相似性。我们采用Jensen-Shannon分歧来测量上述2个多项主题分布之间的相似性。它们之间的对称Jensen-Shannon分歧是：



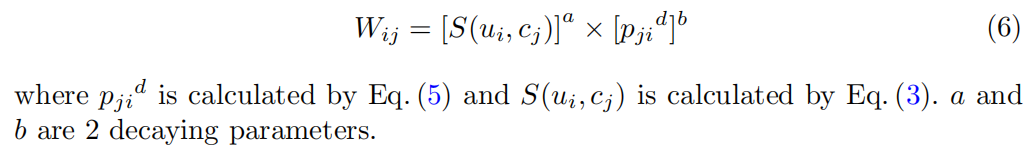
接下来，我们利用空间信息。以前的工作[4]调查了用户连续登记的基于距离的位移。 他们发现关于位移的登记频率的趋势可以通过幂律分布来近似。为了与这种观察到的特性相一致，假设一个用户现在POI cj，我们将用户访问另一个POI ck的概率按距离建模为[13]：



实际上，用户可能不会及时共享他们的位置信息，因此在没有精确空间坐标的情况下，不能简单地应用方程（4）。直观地说，如果用户经常在某个POI办理登机手续，这个POI可能会对他的日常生活产生很大的影响，因为他可能在那里生活或工作。虽然用户的个人行为不太容易依赖用户很少发生的POI。 根据这种现象，我们假设POI cj对用户ui的吸引力是由候选POI cj与用户（用户ui）登记日志列表中的POI之间的等式（4）计算的概率的加权平均值。权重与用户在其历史中的每个POI的登记频率成比例。例如，假设在用户ui的历史列表fi中存在K个POI。POI cj对ui的地理吸引力计算如下：

其中fik是关于POI ck处的用户ui的登记频率的概率，并且通过使用等式（4）来计算pjkd。位移在POI ck和cj之间。

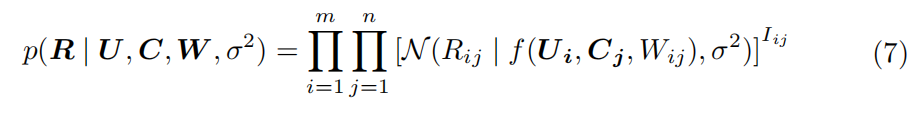
我们假设在POI登记的概率应反映内容和地理影响。因此，我们将这两个因素融合在一起，并将POI cj对用户ui的综合吸引力推导出来：



**3.4 Temporal-Geographical Topic Model for Personalized POI**

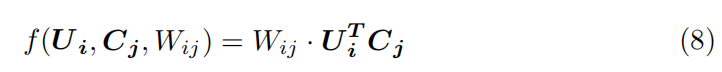
**Recommendation**

为了共同利用时间，空间和内容信息，我们将它们整合到概率矩阵分解模型（TGTM-1）中。 图1（a）显示了TGTM-1的图形模型。我们将观察到的登记频率的条件分布定义为：



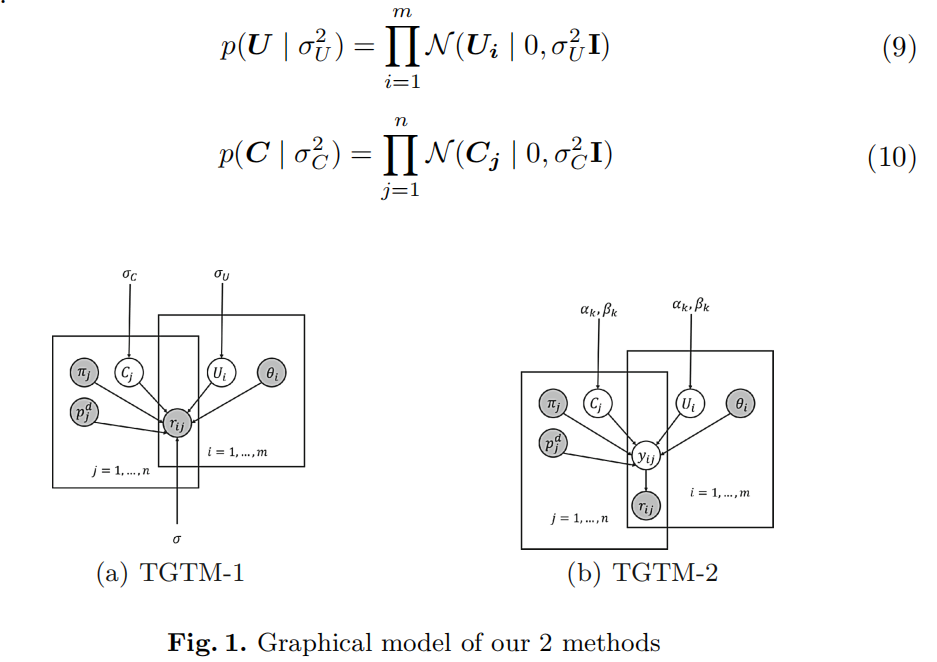
其中N（x |μ，σ2）是具有平均μ和方差σ2的高斯分布的概率密度函数，并且Iij是如果用户ui访问POI cj则等于1的指示符函数，否则等于0。我们使用函数f(Ui，Cj，Wij)近似于POI Cj的用户界面签入频率。

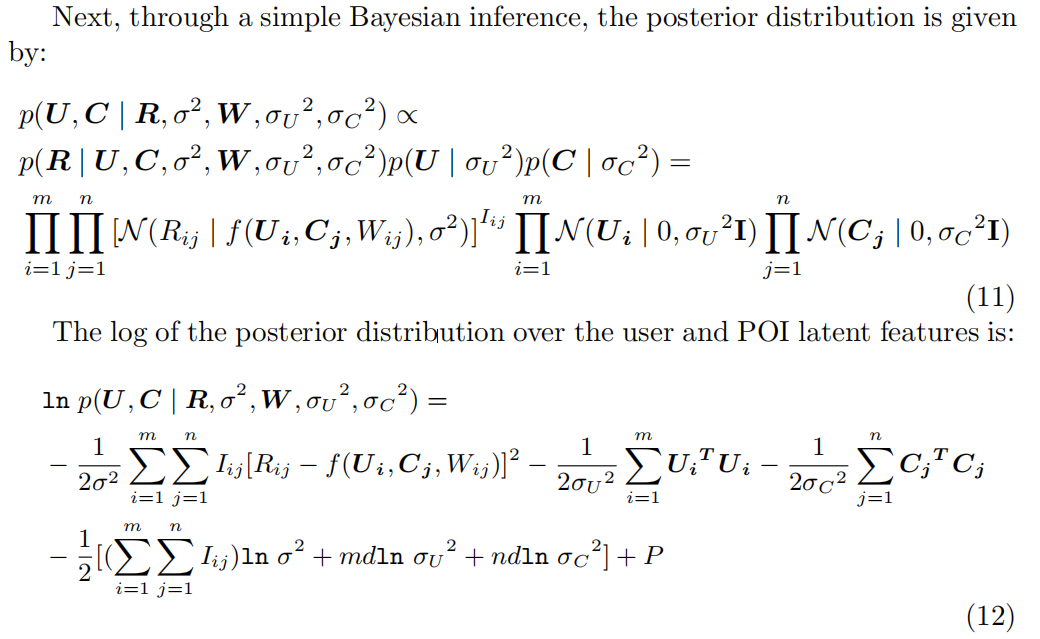
考虑到空间和内容的影响，我们定义：



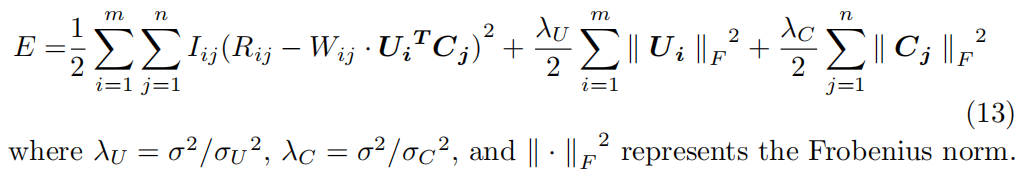
其中Wij是使用方程（6）计算的。用户特定和POI特定潜在特征向量的加权乘积允许我们的模型充分利用所有信息并进行个性化POI推荐。

我们在用户和POI潜在特征向量上放置零均值球面高斯先验：

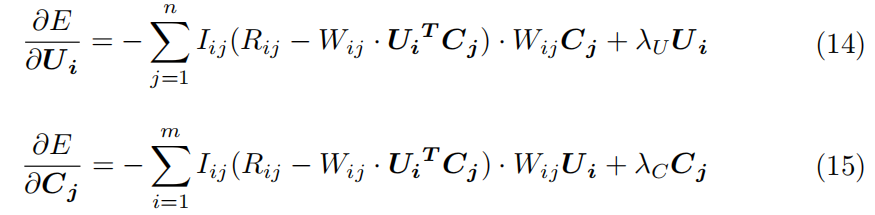




其中P是不依赖于参数的常数，d是潜在特征的数量。 使用固定超参数最大化具有固定超参数的对数的后向矢量等效于使用二次正则化项最小化随后的平方误差目标函数：



将梯度下降算法交替应用于潜矩阵U和C，可以得到局部极小值。



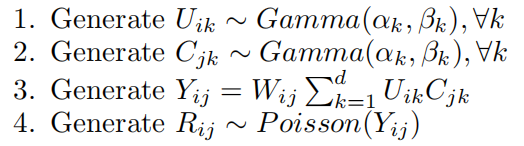
对于登记频率矩阵的非负特性，我们通过在每次更新迭代中将负参数值投影为0来使用预测策略。

在完成上述所有优化过程之后，用户POI签到矩阵中的缺失值可以预测为：

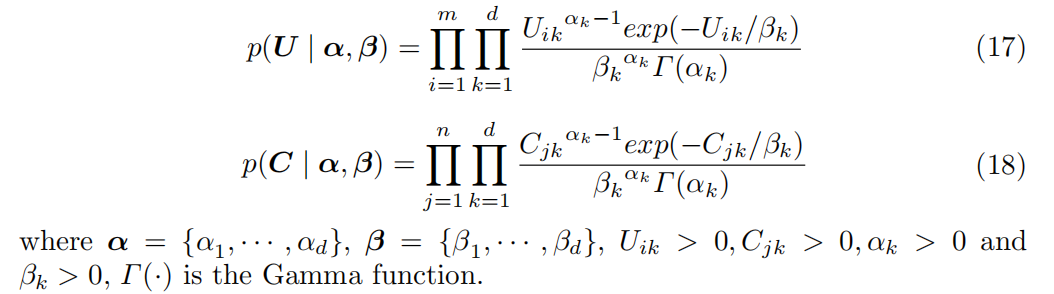


为了减轻高斯分布产生的潜在负值，我们必须使用预测策略。我们从[14]中汲取灵感，并提出了一种更适用于非负值建模的改进模型（TGTM-2），如图1（b）所示。在该模型中，我们假设矩阵R中观察到的登记频率Rij遵循泊松分布，且矩阵Y中的平均值为Yij。矩阵Y中的元素Yij与等式(8)相同的方式定义。但Uik和Cjk被赋予Gamma分布作为经验先验。

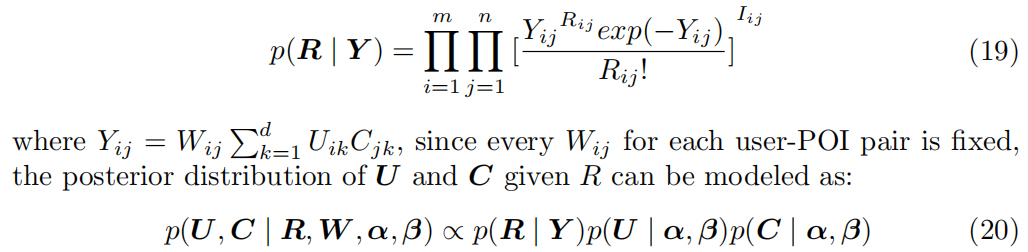
每个观察到的用户-POI计数的生成过程如下：



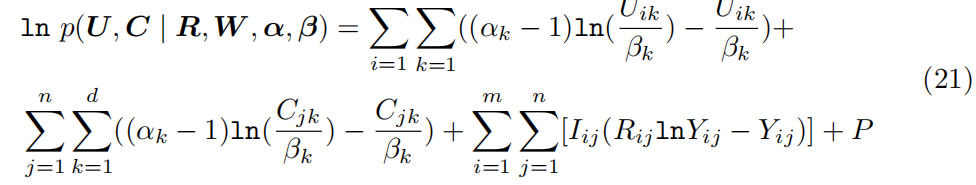
U和C的伽马分布如下：



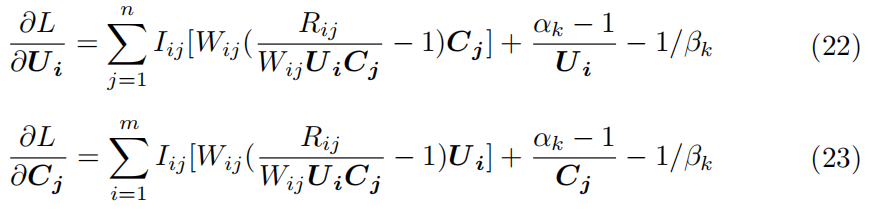
给出了给定Y的R的泊松分布：



对用户和POI潜在特性的后验分布的日志是：



其中P是常数。 取得关于U和C的方程（21）的导数：



等式（16）可用于进行预测，直到最优过程得到收敛。

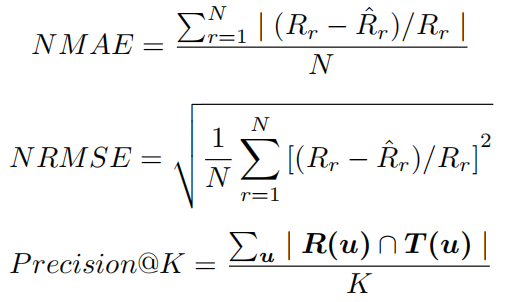
4 Experimental Analysis

4.1 Description of Datasets

我们使用[4]提供的部分数据集。由于Twitter状态消息支持包含地理标签（纬度/经度）以及支持第三方位置共享服务（如Foursquare和Gowalla），空间坐标数据使用Twitter的API进行爬行，其他数据来自Foursquare，详细的数据收集方法可以在[4]中找到。Foursquare是一个基于大规模位置的社交网站。它允许用户在不同位置登记，使用时间戳和空间信息编写文本注释。实际上，用户的登记行为可能并不总是依赖于时间，空间和内容的影响。例如，有些人可能会出差，目的地可能离他们共同的活动区域很远。为了清理数据并删除很少发生的异常值，我们过滤了少于10次签到的用户，并要求每个POI至少应访问10次。此外，我们还限制用户应至少访问每个POI 3次。在实践中，大规模POI可以具有略微不同的地理坐标。在我们的实验中，POI根据其唯一标识符被截断。修剪后，我们获得了3373个独特用户和9333个独特的POI。修剪后的数据集包含31727个签到记录。 用户POI登记矩阵的密度为0.1008％。

4.2 Metrics

在我们的实验中，我们将数据集分为两部分：训练（80％）和测试数据集（20％）。我们使用3个度量来评估性能：归一化平均绝对误差(NMAE)，归一化均方根误差（NRMSE）[14]和P recision @ K。Precision @ K是为top-K POI推荐设计的度量。它们被定义为：



其中N是预测总数，Rr是项目的实际评级，Rr是相应的预测评级。R（u）是前K推荐POI的列表，T（u）是用户u的访问位置。Precision @ K w.r.t每个用户表示实际访问过的前K个推荐位置中POI的百分比。

4.3 Impact of Spatial Information, Content and Temporal Influence

由于预测是在每个子组中独立进行的，因此我们可以准确地了解模型在每个子组中的表现。但是，我们并不知道所有数据的总体性能。鉴于此，在所有子组中的所有独立训练和测试程序之后，我们使用所有子组的所有测试数据集计算度量作为模型的整体性能。以度量标准NMAE为例，符号N是所有子组中预测的总数，而不仅仅是一个单个子组。可以以相同的方式导出其他符号。第4.3节中的实验结果是2个提出的模型的整体性能。

我们使用网格搜索和五重交叉验证来找到Gamma分布的最佳参数，以获得最佳的整体实验性能。Gamma分布的形状参数α= 100，比例参数β= 0.05。β越大，Gamma分布产生的数值范围越宽。 在我们的实验中，我们发现大β会损害模型性能。

参数a和b是非常重要的，因为它们决定了文本内容和空间影响在多大程度上影响了预测的准确性。图2和图3显示了我们的2个模型的预测性能的3D图。内容参数a平衡文本内容的权重：a越大，我们使用文本内容进行预测的越多。空间参数b平衡了地理影响的影响，b越大，推荐时就越考虑距离。考虑到所有指标，当a = 10时，我们的方法获得最佳预测准确度，对于TGTM-1，b = 0.2，而对于TGTM-2，a = 9，b = 0.3。我们还观察到性能对空间影响更敏感，参数b的微小变化可以使得NMAE和NRMSE的性能波动。这表明文本内容和空间影响都对模型性能有贡献，但空间信息影响更大。 因此，POI的地理吸引力在做出旅行决策时起着更重要的作用。

在我们的实验中，我们发现尽管关于NMAE和NRMSE的性能可能较差，但是度量P recision @ K仍然相对较好（参见图2和3）。这是合理的。直观地，在POI推荐场景中，因为登记频率处于不同的数量级，所以我们更关心除了准确的丢失条目值之外的推荐结果的等级。这种现象也证明了我们以指数方式对加权矩阵中的条目进行建模的有效性。虽然这种策略可能会产生相对较大或较小的预测，但指数策略可以很好地区分用户广泛的检查频率行为。 但是，这不能简单地应用于传统的评级预测情况。由于在这种情况下，额定值的数值范围固定在1-5，这远远小于登记频率的范围。

然后，我们研究时间影响对我们提出的模型的影响。在我们提出的模型中，我们根据时间戳（即，星期几和小时）将原始用户-POI登记矩阵子组划分为4个子矩阵。为了学习时间效应，我们进行了以下实验：使用相同的空间和文本参数配置，将我们的完整模型（TGTM-1和TGTM-2）与比较（GTM-1和GTM-2）进行比较，无论时间如何影响。这意味着比较使用原始矩阵进行预测而不使用子分组策略。在本实验中，使用上述结果，我们为TGTM-1和GTM-1设置a = 10，b = 0.2。我们为TGTM-2和GTM-2设置a = 9，b = 0.3。从图4可以看出，时间影响确实对模型性能产生了很大影响。 与不考虑时间影响的模型相比，TGTM-2和TGTM-1的性能提高了近10％。 考虑到时间的Gamma先验可以获得更好的预测性能。该结果显示了基于时间的用户-POI子组的有效性。

4.4 Comparisons

在本节中，我们介绍了我们的模型和一些最先进的POI推荐算法之间的性能比较。

BasicMF: 根据P≈UTC预测缺失值，其仅使用用户-POI登记矩阵而没有任何其他辅助信息（例如，地理，时间信息和文本内容）。这个用作基线。

GeoCF: [20] 通过假设幂律分布并联合集成基于用户的协同过滤算法来考虑空间影响。 该建议是空间信息和用户偏好的线性组合。

UPT: [13] 考虑时间和地理影响，并将辅助信息集成到矩阵分解模型中。 在这种方法中，时间戳也用于子分组。

TLA: [12] 将项目内容和空间信息合并到矩阵分解模型中，从而根据项目内容，空间影响和登记频率进行预测。在该模型中，通过使用区域级流行因子而不是地理坐标来考虑空间信息。

TGTM-1 TGTM-2: [Sect. 3.4] 使用时间戳进行用户POI子分组。 在子组中，他们将涉及登记频率，文本内容和空间信息的预测纳入概率矩阵分解模型。 假设TGTM-1的用户和POI潜在特征矩阵符合高斯先验，而假设TGTM-2的矩阵符合Gamma先验。

在该实验中，我们设定a = 10，对于TGTM-1，b = 0.2，对于TGTM-2，a = 9，b = 0.3。从表1和图5中，我们可以看出应用Gamma之前的TGTM-2确实优于具有高斯先验的TGTM-1。表1还显示UPT，TGTM-1和TGTM-2在每个子组中表现不同。原因是当时间场景发生变化时，话题兴趣和地理距离对人类移动行为的影响略有不同，尤其是人们倾向于访问匹配的POI 他们的利益在周末以长途为代价更好。由于Basic MF，GeoCF和TLA不考虑时间信息，因此我们使用它们的整体性能来表示每个子组的性能。我们还可以看到，关于度量P recision @ K的性能相对较高，原因是我们的数据预处理过程过滤掉了几乎没有相关性的数据，并且测试部分中的所有用户都没有太多的签到。我们将大规模的比较实验作为我们未来的工作。尽管所有比较都可以接受，但当K = 1时，我们的TGTM-2模型仍然可以将基线性能提高30％以上。

BasicMF实现了最低的精度和精度，因为这种通用模型不考虑任何辅助信息。对于GeoCF模型，由于我们的实验数据中没有社会关系信息，我们只考虑基于用户的协同过滤因子和空间影响因子。当空间影响力达到70％时，我们发现GeoCF获得最佳性能，而基于用户的协同过滤因子占30％。 这意味着空间信息在POI推荐场景中起主导作用。GeoCF比BasicMF模型表现更好，但它仍然不如我们的方法准确。原因是尽管GeoCF使用协同过滤方法考虑了更多相关数据并考虑了地理影响，但POI推荐的一些潜在重要因素仍然丢失。例如，文本内容信息明确地示出了目标POI是否匹配用户的兴趣。此外，时间信息直观地解释了为什么用户的活动范围在周末更大。UPT利用时间信息将原始数据分为4个子群，并将空间影响整合到矩阵分解模型中。实验结果证明了将时间和地理影响结合起来的好处。TLA利用文本和空间信息来推荐POI。UPT优于TLA和GeoCF，这意味着时间信息比使用用户协同过滤方法推断的显式文本信息和不明确的用户偏好更有助于用户的行为。可能的原因是，大多数人在工作日有规律地工作，工作日比周末更长时间。因此，话题的兴趣表现出相对较弱的影响。

表1和图5还表明，我们提出的2个模型始终优于比较，表明联合充分利用文本内容，时间和地理信息确实提高了POI推荐的性能。虽然有些比较在NMAE和NRMSE上表现良好，但它们仍然无法比Precision @ K上的2个模型表现更好。因为它们不能适当地处理大范围的检查频率值，这表明以指数形式对地理影响和内容信息进行建模使得我们的模型对于大范围的用户POI登记频率更加灵活。

5 Conclusions

在本文中，我们提出了两种概率矩阵分解算法来进行POI推荐。 提出的推荐方法有几个优点。首先，时间戳用于对原始数据进行子组化，以便深入理解不同时间场景中的用户行为模式。其次，模型捕获空间影响，这已被证明对用户的移动行为非常重要。第三，深入开发文本信息符合真正个性化推荐系统的标准，推荐符合用户兴趣的POI。最后但并非最不重要的是，所提出的方法以指数形式模拟文本和地理影响，适用于大范围 用户的隐式登记反馈数据。最后，对通过Twitter API采集的真实数据进行了大量实验，验证了本文方法的实用性。

我们未来的工作包括满足在线实时需求并进行大规模比较实验以进一步评估性能。