# 通过在线地图搜索查询分析探索城市感兴趣区域

Exploring the Urban Region-of-Interest through the Analysis of Online Map Search Queries

## 摘要

城市感兴趣区域（ROI）是指具有吸引人们关注和活动的特定功能的综合城市区域，例如娱乐商业区，交通枢纽和城市地标。实际上，在宏观层面，投资回报率是集聚经济的代表之一，在城市商业规划中发挥着重要作用。在微观层面，投资回报率为了解城市生活，需求和人口流动提供了一个有用的场所。然而，由于投资回报率的模糊性和多样性，它仍然缺乏以整体方式调查投资回报率的定量方法。为此，本文通过挖掘大规模在线地图查询日志，提出了ROI分析的系统研究，为ROI检测和分析提供了一种新的数据驱动研究范式。具体来说，我们首先将城市区域划分为小区域网格，并根据从地图查询中提取的转换信息计算其PageRank值作为访问流行度。然后，我们提出了一种基于密度的聚类方法，用于将具有高流行度的邻近区域网格合并到集成的ROI中。之后，为了进一步探索不同ROI的概况，我们开发了一个空间时间潜在因子模型URPTM（Urban Roi Profiling Topic Model）来识别ROI访问者的潜在旅行模式和兴趣点（POI）需求。最后，我们实施了大量实验，根据从北京收集的大规模现实世界数据，对我们的方法进行了实证评估。事实上，通过可视化从URPTM获得的结果，我们可以成功获得许多有意义的旅行模式和有趣的城市生活发现。

**1引言**

随着市场经济的发展，城市群聚集现象越来越普遍，这明显影响了城市商业活动和所有相关组成部分。作为代表之一，城市感兴趣区域（ROI）吸引了公民在日常生活中的关注和活动。与兴趣点（POI）不同，ROI是指具有特定功能的综合城区，例如娱乐商业区，交通枢纽和城市地标。实际上，调查这些投资回报率可以使城市商业的多个方面受益。具体而言，在宏观层面，投资回报率是衡量集聚经济繁荣的指标，可以促进城市管理的城市商业规划。在微观层面，ROI为商业从业者提供了一个有用的场所，可以了解城市生活，人们的日常需求和流动性，从而实现广泛的基于位置的服务，包括商业网站选择和有针对性的广告。

在文献中，基于新近可用的移动大数据，大量研究工作致力于城市相关主题[7,16,24,26]。 然而，由于投资回报率的模糊性和多样性，它仍然缺乏以整体方式调查投资回报率的定量方法。 例如，与计划的城区相比，投资回报率的出现通常取决于城市商业环境中的内部供需关系。 因此，准确确定ROI的范围和边界是一项非常重要的任务。 此外，投资回报率通常包含多样化的功能，可能会吸引不同的旅行需求的不同访客。

为此，在本文中，我们通过挖掘大规模在线地图查询日志[21]提出了ROI分析的系统研究，该日志为ROI检测和分析提供了一种新的数据驱动研究范式。实际上，与城市计算中使用的传统移动数据（如签到和出租车轨迹[27]）相比，在线地图查询包含更丰富的上下文信息，用于描述城市环境中的旅行，可用于实现细粒度的投资回报率分析。具体而言，我们首先将城市区域划分为小区域网格，并基于从地图查询中提取的转换信息（即，起始 - 目的地）来计算它们的PageRank值作为访问流行度。然后，我们提出了一种基于密度的聚类方法，用于将具有高流行度的邻近区域网格合并到集成的ROI中。之后，为了进一步探索不同ROI的概况，我们开发了一个时空潜在因子模型URPTM（Urban Roi Profiling Topic Model）来识别ROI访问者的潜在旅行模式和兴趣点（POI）需求。特别是，URPTM可以揭示ROI吸引人们访问的答案和原因。最后，我们实施了大量实验，根据从北京收集的大规模现实世界数据，对我们的方法进行了实证评估。事实上，通过可视化从URPTM获得的结果，我们可以成功获得许多有意义的旅行模式和有趣的城市生活发现。

**2数据描述**

在本研究中，我们使用从中国主要商业在线地图提供商收集的两组真实数据，即地图搜索查询日志和城市POI。

具体来说，地图搜索查询日志包含北京地图用户的城市路线搜索历史记录。例如，表1显示了查询日志的玩具示例，其中每个记录由匿名用户ID，详细查询时间以及查询的起点和终点组成。特别地，包括查询中的原点和目的地的详细GPS坐标，并且大多数位置可以链接到特定POI。图1（a）和1（b）分别展示了地图查询中的起源和目的地的地理分布。从图中可以看出，查询起源在城市中均匀分布，而目的地往往集中在一些聚集区域，这自然会激励我们检测投资回报率。此外，地图查询可以反映公民的旅行模式。例如，图2显示了从地图查询中提取的工作日上午（上午8:00至上午12:00）北京地区之间的城市旅游流量。直观地，根据旅行流量，我们可以根据基本访问频率或图表分析方法（如PageRank）轻松估算不同城市地区的人口。

在城市POI数据中，除了详细的位置信息之外，还为每个POI提供了许多标签。 标签代表POI的功能类别，例如咖啡馆，综合医院和购物中心。 图1（c）显示了地图查询中出现的POI的地理分布，同时，图1（d）显示了相应POI的标签频率的词云表示。

**3问题定义**

基于现实世界的数据，本文的目标包括两个任务，即1）ROI检测，2）ROI分析。 具体而言，对于ROI检测的任务，我们的目标是基于从地图查询中提取的旅行流信息来发现具有高访问流行度的综合城市区域。 对于ROI分析的任务，我们的目标是为每个ROI的访问者的时空偏好建模，这可以揭示ROI吸引人们访问的答案和原因。

图3显示了我们的ROI分析框架的概述。具体来说，在第一项任务中，我们首先将城市区域划分为小区域网格。然后，通过从查询数据中提取这些网格中的行进流，我们形成转移矩阵，其中每个元素表示相应网格之间的查询频率。之后，在矩阵上进行PageRank算法，以计算每个网格的访问流行度。随着网格的普及，我们进一步使用基于密度的算法对网格进行聚类，以检测ROI。在第二项任务中，我们设计了一个空间时间潜在因子模型URPTM，用于发现ROI访问者的潜在旅行主题。在模型中，每个ROI被视为文档，而地图查询中的时间，来源和POI标签（其中目的地位于ROI中）被视为单词。在学习模型之后，我们可以获得每个ROI的访问者的时空偏好，这可以用于许多应用，例如旅行需求分析和目标ROI分割。

**4技术细节**

在本节中，我们将介绍ROI分析方法的技术细节。 为了便于演示，我们在图4（a）中绘制了一些地标。

**4.1 ROI检测算法**

在这里，我们引入了一种基于密度的聚类方法来有效地检测ROI。 表2中显示了一些符号。具体来说，想法是首先将城市区域划分为小区域，然后将它们与聚类算法合并为ROI。 请注意，这里我们使用基于网格的方法来分割城市区域[9]，而不是基于街道的分割[23]，因为ROI的边界通常非常模糊并且与计划的街区不一致。 图4（b）显示了基于网格的分区的示例。

**4.1.1计算区域网格的普及程度。**

由于投资回报率是吸引人们关注和活动的城市区域，我们需要一个衡量访问受欢迎程度的指标，以便我们可以将具有较高访问受欢迎度的网格合并为一组。虽然访问频率是流行的直接指标，但它无法捕捉不同地区之间的过渡关系。直观地说，一个经常被其他热门地区的人询问的地区应该更具吸引力。此外，当计划到POI的旅行时，用户可以将他们的查询目的地定位在附近的地铁站，然后在他们到达附近的地点之后向特定的POI提交查询。这些区域通常是交通枢纽，并不是用户的真正目的地。因此，在本文中，我们建议使用PageRank值来衡量访问流行度。使用PageRank算法，流量集线器的访问流行度（通常作为具有大流量和大流量的中间点）将收敛到相对较小的值。具体地，基于查询数据，我们首先构造邻接矩阵，其中每个元素表示相应网格之间的查询频率。然后基于矩阵进行PageRank，并将每个网格的访问流行度表示为其PageRank值 。

同时，我们引入指标热值𝐼𝐼来表示网格的受欢迎程度如何受到周围网格的影响。 具体来说，我们首先定义一个热核来计算how如何影响𝑗，其定义如下：



其中表示网格中心与网格i 和 g之间的欧几里德距离。同时， 会随着的增加而衰减。之后，我们将i的热值定义为，其中表示围绕i的网格集。 直观地说，周围的网格越受欢迎,的价值越高。

**4.1.2基于密度的聚类用于ROI检测。**

我们基于密度的聚类方法的一般聚类过程如下。 首先，我们启动一个新的集群，其中包含一个新的核心网格，它不会集群到现有集群，然后将当前核心网格可访问的核心网格和流行网格（参见表2）添加到此集群中。 然后，我们重复此过程，直到所有核心网格都已群集，并且不会生成新的群集。 与传统的DBSCAN算法类似[3]，我们的聚类算法对起始核心网格不敏感，无论我们从哪个网格开始，我们都可以获得相同的聚类结果。 具体来说，我们使用广度优先搜索（BFS）来实现聚类算法，详细过程如算法1所示。

请注意，我们还将集线器网格视为上述过程中的核心网格。从一开始，整个城区被划分为非常小的网格，并且可能在某些情况下，ROI中的流行网格被一些不受欢迎的网格分开。例如，大型购物中心可能包含在其中心区域具有入口和不受欢迎的网格的流行网格，因为用户经常在他们搜索时在入口处定位他们的目的地。虽然并不总是很受欢迎，但这些网格通常具有较大的热值。我们将热值高于阈值的网格命名为“中心网格”。此外，我们在图5（a）中给出了一个示例，以显示中心网格如何影响群集过程。橙色网格是流行的网格，黄色网格是不受欢迎的中心网格。直观地说，中间的橙色网格形成了投资回报率。如果我们不考虑黄色网格，为了检测这种形状的ROI，我们应该设置𝑂𝑂不大于2（或者在中间的4个网格中没有核心网格）。这将导致结果成为图5（a）中的上图，其中我们发现红色ROI和蓝色ROI。然而，外部蓝色网格环明显不够密集以形成ROI。如果我们使用黄色轮毂网格作为核心网格，那么我们可以将𝑂𝑂设置为4以获得更密集的结果ROI，如图5（a）中的底部图形所示。

在现实世界的应用中，访问流行度并非在整个城市地区均匀分布。以北京为例，如图1（b）所示，我们可以看到北京东部地区更频繁地被用户查询。因此，这些区域的网格将比其他网格具有更高的访问受欢迎程度。如果我们使用相同的阈值*PRth*来确定整个城区的活动网格（参见表2），则聚类算法将无法获得合理的结果。一小部分*PRth*可能会检测到东部地区过多的热门网格，从而导致非常大的集群，而*PRth*的大量价值会导致许多集群丢失，例如，北京南部的一些ROI可能会被忽略。直观地说，我们应该为不同的区域设置不同的*PRth*值来检测所有的ROI。 *Hth*存在类似的问题。为解决此问题，我们建议限制群集的大小以自动确定阈值。也就是说，如果我们发现算法1的簇输出的大小太大，即核心网格的数量大于阈值*Msize*，那么该区域的阈值*PRth*和*Hth*的值增加。重新划分过程的示例如图5（b）所示，蓝色的重新划分为两个较小的簇，最后获得三个ROI。在算法2中示出了完成的具有详细自适应重新分区过程的ROI检测算法。注意，随着过程的进行，我们也增加了*Msize*，因为直观地，较高流行度的ROI可能具有更大的尺寸。

**4.2 ROI分析模型**

在这里，我们建议进一步探索不同ROI的概况，并开发一个名为URPTM（Urban Roi Profiling Topic Model）的时空潜在因子模型，以识别ROI访问者的潜在旅行模式。 表3说明了UPRTM的一些符号和概念。

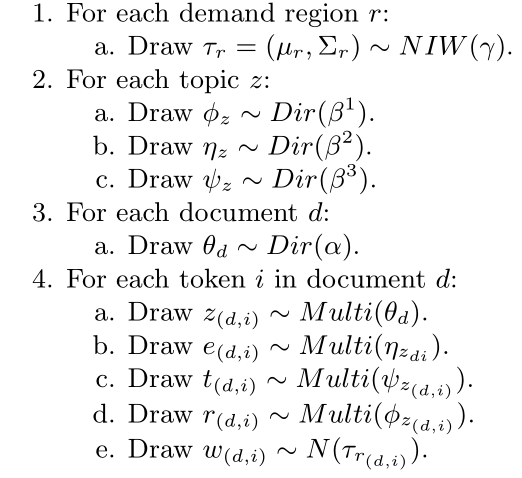
**4.2.1 ROI概要主题模型。**

首先，我们想介绍如何使我们的ROI分析问题适应主题建模设置，主题建模设置考虑空间和时间信息。 具体地，每个ROI被视为文档*d*，并且单词标记形成为*（e，t，w）*，从查询数据和POI数据中提取，其中*e*表示搜索到的POI的类别，*w*表示GPS坐标 原点位置，*t*表示查询的时间戳。 同时，我们将时间分成几个时隙，以便可以将mode建模为离散值。

我们假设有*K*个主题反映用户的旅行模式，直观地，在每个主题中，来自特定地点的访问者希望在特定时间查询特定的POI类别。 因此，我们对这些主题进行建模，以揭示POI类别，时隙，查询起点和查询数据中显示的ROI之间的关系。

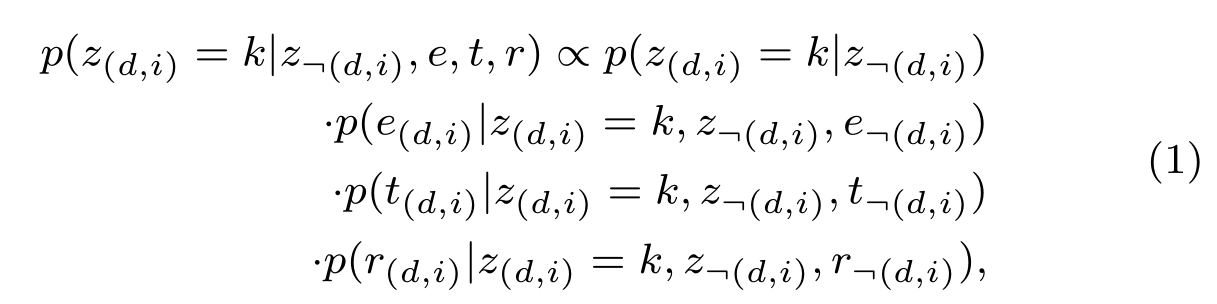
虽然POI类别和时间段可以通过多项分布轻松地与主题相关联。 由于查询起源是空间数据，如何将非离散空间信息集成到我们的模型中成为一个巨大的挑战。 直观地说，来自同一地区的游客往往有类似的旅行需求。 例如，在某些地区，人们会对购物产生兴趣，因为他们可以方便地去购物中心。 为此，我们创造性地建议使用高斯分布对查询起点进行建模。 具体来说，我们假设市区中存在有限数量的主题区域，并且每个主题区域𝑠被赋予具有参数）的高斯分布。

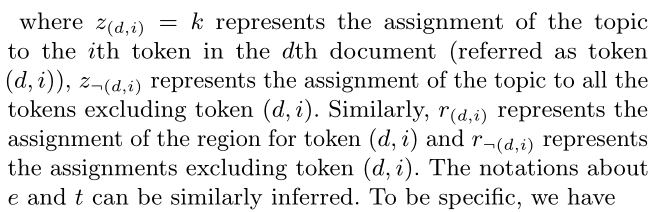
URPTM的图形模型如图6所示，生成过程如下所示：

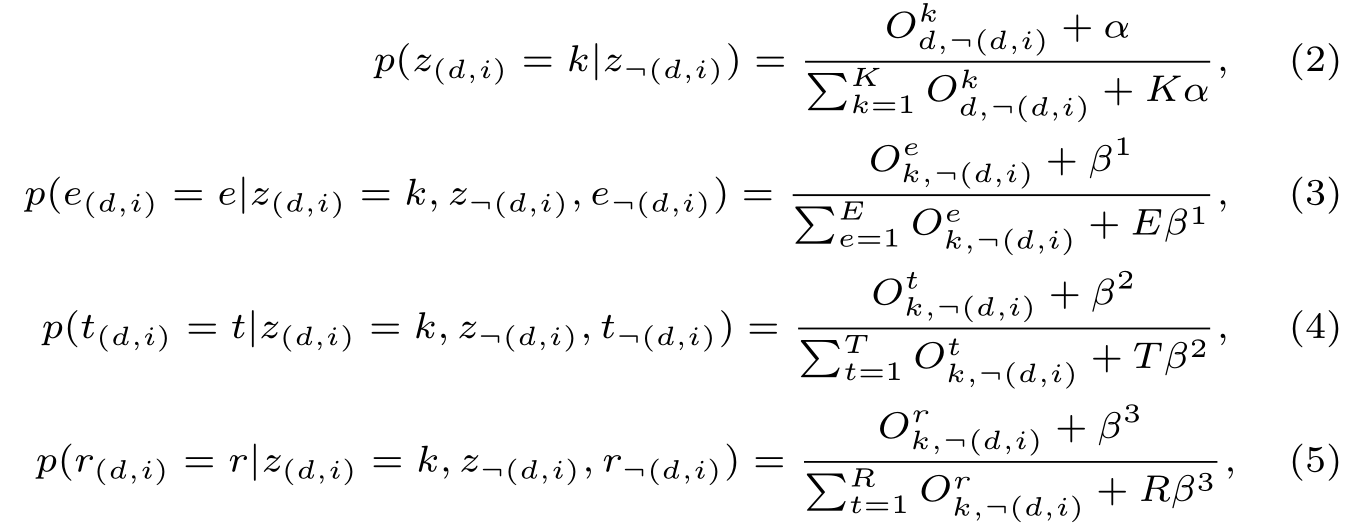


**4.2.2模型推理。**

为了学习我们的模型URPTM，采用吉布斯采样[5]来推断所有参数。 具体来说，为了更新*A*，我们有

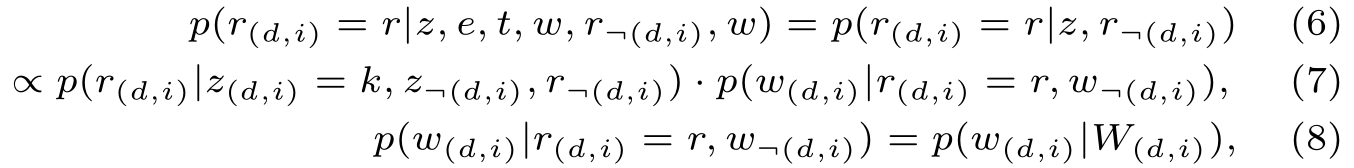






其中𝑃𝑙𝑒，¬𝑗表示第topic主题在文件中分配的频率𝑒除了令牌（𝑒，𝑗）。 𝐿表示主题的数量。 𝑃𝑓𝑙，¬（𝑒，𝑗）表示第9个主题分配给除了令牌（𝑒，𝑗）的语料库中具有第PO个POI类型的令牌的频率。 𝐹表示POI类型的数量。 𝑃𝑢𝑙，¬（𝑒，𝑗）表示

第9个主题分配给令牌中第time时间戳的令牌的频率，不包括令牌（𝑒，𝑗）。 𝑈表示时间戳的数量。 𝑃𝑠𝑙，¬（𝑒，𝑗）表示第9个主题分配给语料库中第region区域的令牌的频率，不包括令牌（𝑒，𝑗）。 𝑆表示区域数量。 对于地区𝑠，我们有



其中𝑋（𝑒，𝑗）表示令牌中除了（𝑒，𝑗）之外的所有点的集合，它们被赋予与令牌（𝑒，𝑗）相同的区域𝑠。 后验分布𝑞（𝑥|𝑋）遵循多变量Student-T分布[18]，其参数与[19]中的分布相似。

**5 实验**

我们在三周内收集了北京的在线地图搜索查询数据。 过滤北京以外的失败查询和查询，并在短时间内删除同一用户提交的同一目的地的重复查询（本文使用十分钟），我们获得了超过1,500,000个查询。 另一方面，我们通过从在线地图API搜索关键字和目的地来获取目的地的POI类别，并且获得了具有158种标签的大约120,000个POI。 对于这个POI数据集，我们还进行了一些预处理以删除没有标签的POI并合并一些标签，例如“清华大学东门”和“清华大学”合并为“清华大学”。

根据地理信息系统（GIS）坐标，我们将北京市区划分为大小为𝑀×𝑀𝑀2的小区域网格。 直观地，𝑀的不同设置将导致ROI检测的不同准确性。 为了进行细粒度的ROI分析，我们在实验中根据经验设定𝑀= 200。

在ROI检测中，可以根据不同的应用需求调整超参数。例如，如果我们想获得更密集的ROI分布，我们可以增加𝑂𝑂。同样，如果我们增加𝑄𝑆𝑄𝑆和𝐼𝐼，将获得具有较高访问流行度的ROI，反之亦然。在我们的实验中，我们根据经验设定τ= 0。 2当计算热值时，𝒪𝒪为8个网格与𝑗共享顶点或边。此外，我们设置𝑂𝑂= 3，𝑁𝑡𝑗𝑨𝑓，𝑄𝑆𝑄𝑆，𝐼𝐼的初始值为3,0。 00003,0。 0003让他们增加0。 2,0。 0000005,0。每轮0000005。生成令牌时，我们在检测到的ROI中使用带有目标的查询数据，并将时间分成48个时隙，表示工作日和周末24小时。对于URPTM，我们假设有20个主题和200个主题区域。我们选择先验β= 2。 5，γ1= 0。 25，γ2= 1，γ3= 0。 3进行初步实验，并将T分布中的最高参数设为ν0=Σ︀︀，𝑗𝑗（𝑒，𝑗）𝑂，Λ0= [︂0.0010.0005 0.0005 0.001]︂，λ0= 2，ξ0 = 2，其中𝑂表示令牌数。 Gibbs采样中的最大迭代次数设置为1,000。

**5.2 ROI检测评估**

在我们的ROI检测算法中，我们使用PageRank值来表示每个网格的访问流行度。 为了验证PageRank的优势，我们通过使用查询频率作为基线来选择另一个流行度标准。 比较结果的一些例子如图7所示，其中蓝色和红色网格是北京同一地区的前500个流行网格，分别用频率和PageRank生成。

从图中可以看出，根据PageRank值，一些具有高访问频率的网格不被认为是受欢迎的。 例如，在图7（a）中，蓝色网格位于公交车站附近，而红色网格位于北京交通大学。 虽然蓝色网格中还有一座建筑物，但实际上它位于校园内。 同样，在图7（b）中，与蓝色网格相比，红色网格包含更多建筑物，餐厅和游泳池。 在这两幅图中，红色网格更有可能成为游客的最终目的地。

图8显示了我们的算法在北京检测到的总共188个ROI，其中不同的颜色用于识别ROI。 总的来说，北京的北部和东部地区的投资回报率高于南部和西部地区。 此外，还有一些着名的投资回报率，如国贸CBD 1和中关村2。 同时，我们可以发现不同的投资回报率具有不同的POI组成，例如中关村的写字楼和商场，西单3的商场，五道口4的餐厅和国贸CBD的办公楼。

**5.3 ROI分析评估**

在本小节中，我们评估了模型URPTM在ROI分析中的性能，包括主题建模和ROI的访问者推断。

5.3.1 ROI的主题建模。 在表4中，我们展示了URPTM学习的5个随机选择主题的详细信息。 具体来说，第一行显示主题的空间分布，我们使用热图来表示主题吸引不同位置的访问者的概率，即𝑞（𝑥|𝑨）=Σ︀︀𝑠（𝑥| 𝑠）𝑞（𝑠|𝑨）。 第二行显示主题中包含的主要POI类别。 第三行示出了主题的时间分布，其中x轴包含48个时隙（即，工作日的上午0:00至下午23:00，周末的上午0:00至下午23:00）。 请注意，由于工作日的时段数是周末的2.5倍，为了便于演示，我们通过将工作日的值除以2.5来重新调整范围。

从结果中，我们可以直观地了解这些主题。 以主题18为例，我们可以发现它是关于“上班”的旅行模式，主要发生在工作日的早晨。 从空间分布来看，我们可以发现这个话题吸引了更多来自北京东北部的人们（即北京的主要住宅区之一）。 实际上，通过这些主题，我们可以了解人们的城市生活。

示例1.在工作日的早晨，患者去医院进行登记，其中一些甚至在凌晨5点出发（基于主题17）。 几个小时后，人们将从家里上班（基于主题18）。 在工作了一整天之后，他们会去一些带家人的餐馆在一个购物中心吃晚餐（基于主题7），而他们中的大多数人想在周末购物（根据主题15）。 对于大学生来说，他们的旅行模式在工作日和周末之间有一个模糊的界限（基于主题14）。

同时，这些主题可以为城市商业规划提供指导，如商业选址。 例如，可以根据主题18的空间分布来构建办公楼，并且可以根据主题7来构建医院。

5.3.2投资回报率的全球访客推理。 了解不同投资回报率的访问者来源可以使许多商业应用受益，例如有针对性的广告。 直觉上，我们可以通过计算地图查询中的起源来获得此分布。 然而，即使有大规模的查询数据，观察到的ROI访问者的起源通常也非常稀少。 实际上，我们的模型URPTM可用于全球干预ROI的访客分布。 具体来说，我们可以使用概率𝑞（𝑥|𝑒）=Σ︀𝑠𝑞（𝑥|𝑠）𝑞（𝑠|𝑒）来估算ROI中访客的来源。

特别是，我们采取了三个着名的投资回报率，即中关村，西单和三里屯5，进行模型评估。 图9（a），9（b）和9（c）显示了我们数据集中这些ROI的确切查询起源。 我们可以观察到，虽然它们是着名的商业区，但只能识别当地游客，其他地区的游客是非常孤儿的话。

图9（d），9（e）和9（f）显示了URPTM推断的访问者分布，这些分布更加合理和全面。 以西单为例，我们发现它可以吸引地铁1号线（即北京中间的东西直线）的游客，这在西单位于1号线时是合理的，方便这些游客 去那里 因此，西单的一家商店可以在地铁1号线上做广告，以吸引更多的顾客。 此外，西单还可以吸引北京北部的游客，那里有许多大学。

5.3.3 ROI的主题意识访问者推断。 除了ROI的全球访问者推断之外，URPTM还可用于推断ROI访问者关于不同主题的起源。 例如，访问西单吃饭的人的起源在哪里？ 具体来说，对于ROI𝑒和主题𝑨，我们可以通过计算得到原点分布𝑞（𝑥|𝑒，𝑨）=Σ︀︀𝑠（𝑥|𝑠）Σ︀︀𝑨（𝑠|𝑨）𝑞（𝑨|𝑒））。

特别是，由于经典的潜在Dirichlet分配（LDA）[2]也可以为主题分配令牌，这里我们将其用作比较的基线。 与URPTM类似，我们将ROI视为文档，并将每个查询的POI类别视为令牌。 在可视化步骤中，基于具有热图的ROI主题中的大多数可能查询的POI类别的来源绘制位置。

我们在Xidan中展示了不同模型的结果，图10中有3个主要学习主题。我们可以发现URPTM获得的结果更合理，不同主题的访问者来源更清晰。例如，图10（e）显示了URPTM学习的主题7的访问者，其中POI主要是关于“餐馆”。图10（f）显示了主题15的访问者，其中POI主要是关于“购物”。与主题7相比，主题15吸引了来自北京东部的游客较少。事实上，许多购物中心位于北京东部，因此西单在购物方面对游客来说并不那么吸引人。与此同时，主题7对于海淀6区的游客来说并不像主题15那样具有吸引力，因为那里的人们更喜欢五道口的餐馆。实际上，基于不同主题的访问者推断，我们还可以获得不同ROI中访问者的主题感知分解​​。图11显示了以国贸为例的访问者。 ROI的访问者由具有不同主题的访问者组成。

5.4 ROI细分的案例研究在本小节中，我们介绍了一个关于ROI细分的案例研究，旨在将ROI细分为不同的集群。实际上，ROI细分在城市商业规划中发挥着重要作用。例如，如果两个投资回报率具有相似的主题，则为其执行类似的促销和广告活动是合理的。具体而言，基于URPTM的结果，我们可以根据主题分布将ROI划分为不同的集群。图12显示了K-Means算法的ROI分割结果，其中生成了15个簇，并且同一簇中的ROI具有相同的颜色。通常，通过手动检查结果，我们认为分割结果是合理的。例如，在三里屯和大望路，有许多购物中心和酒店，他们的访客来自北京东部。因此，如果有人想在三里屯开一家咖啡馆，但由于某种原因必须改变位置，大王路是另一个不错的选择，因为它们都属于同一个集群，因此有类似的访客。