华 东 师 范 大 学 软 件 学 院

2015 年软件工程学士学位论文

基于潜在语义的区域交通模式挖掘

Regional Traffic Pattern Mining Based on Latent Semantics

姓　　名：汪诚愚

学　　号：10112510249

班　　级：11 级 2 班

指导教师姓名：何晓丰

指导教师职称：研究员

2015 年 5 月

目 录

摘 要

随着 GPS 技术的发展，车辆采集了海量的轨迹数据。轨迹中隐含了丰富的信息，促进了交通模式挖掘的发展。然而，大部分已有的工作主要关注于挖掘轨迹本身的模式以及兴趣点（POI），与之不同，本论文目的在于挖掘海量 GPS 轨迹数据中的区域性的、集体的交通模式，以及隐含的语义信息。

在本论文中，我们提出了一个数据驱动的框架，它可以利用隐性狄利克雷分配（LDA），从大规模轨迹数据集中挖掘区域交通模式。具体地说，我们首先介绍了 GPS 轨迹数据的建模问题。之后，我们描述了在交通模式挖掘的框架下的模型表示。与 LDA 常用于维度规约和文本聚类不同，我们在轨迹挖掘的场景中描述了数据实例的表示方法（即“文档”和“词”的表示），以及它基于 MapReduce 框架的生成方法。此外，我们使用 Gibbs 采样算法来学习模型的参数。通过发现隐含主题的分布，可以分析交通模式及其语义含义。为了对之前没有出现的数据进行预测，我们在 GPS 轨迹上进行推理，来发现这一区域的主题分布。

我们在大规模真实轨迹数据集上进行了广泛的实验。实验结果显示了本论文中提出的框架有很强的能力来捕获城市动态，为理解人类活动和城市规划提供了宝贵的支撑点。

关键词：模式挖掘，GPS 轨迹，隐性狄利克雷分配，Gibbs 采样

Abstract

The development of GPS technology enables vehicles to generate massive amounts of trajectory data. The rich information contained in the trajectories motivates the discovery of traffic patterns. While most existing work focuses on mining patterns of trajectories themselves or places of interest (POIs), this thesis concerns with mining regional, collective patterns generated by a large number of vehicles with GPS devices mounted and the hidden semantic meanings behind the patterns.

In this thesis, a data-driven framework is introduced to mine regional traffic patterns from large-scale trajectory datasets based on Latent Dirichlet Allocation (LDA). Specifically, we first introduce GPS trajectory modeling. Later, we describe the model representation under the traffic pattern mining framework. Different from the normal usage of LDA for dimension reduction and text clustering, we describe the data instance representation (i.e., document and word representation) in the context of trajectory mining, and its generating methods based on the MapReduce framework. Then, we employ Gibbs sampling algorithm to learn the parameters of the model. By discovering distributions of hidden topics, we can analyze traffic patterns and their semantic meanings. To make prediction on previously unseen data, we preform inference on GPS trajectories to discover topic distributions of the region.

We conduct extensive experiments with real-life trajectory datasets in a large scale. The results show the proposed framework in this thesis has a strong capacity to capture the city dynamics and provide valuable insights for human mobility understanding and urban planning.

Keywords: Pattern mining, GPS trajectory, Latent Dirichlet Allocation, Gibbs sampling

一、 绪论

(一)背景

地理定位技术以及基于无线网络的通讯技术突飞猛进，使得移动计算系统和基于位置的服务（location based service）的发展如火如荼。大量的服务和应用产生了海量的空间轨迹数据。一条轨迹表示一个移动物体在空间和时间上的位置变化，如人、车辆、动物、自然现象等。以下给出两个例子：

1. 人物轨迹：人们在不断地记录在世界上活动的轨迹。旅游者用 GPS 记录自己的旅行轨迹，发布到互联网上，与朋友分享。人们在运动的时候，例如骑车、慢跑时，用软件记录自己的运动轨迹，进行统计和分析。此外，人们在用手机的时候，位置会自动地传给服务运营商的塔台；在刷卡消费的时候，银行会自动记录他的消费轨迹。
2. 交通轨迹：近年来，大量车辆，例如出租车，安装了 GPS 定位设备。大城市的出租车上都有 GPS 传感器，它们实时地以一定频率将自己的位置信息传给出租车公司的数据中心。这些数据构成了海量空间轨迹，这些轨迹能用来进行资源调配、安全管理、交通分析等。

总而言之，轨迹数据给我们带来前所未有的海量信息。这些信息有助于我们理解运动的物体和地点，促使了系统性的研究来处理和挖掘海量轨迹数据，使它们得到广泛应用。其中，轨迹的模式挖掘在数据挖掘领域，是一个快速发展的研究方向。它的目的在于从海量轨迹数据中发现内在的模式，来帮助人们理解数据，进而发现数据中隐含的信息。

在这些方面中，GPS 轨迹数据的研究正越来越受到重视。有了 GPS 定位仪，人们能容易地得到轨迹信息。尤其在交通领域，城市里数以万计的出租车每天产生 GB 级的 GPS 轨迹数据。对这些数据的分析和挖掘，产生了很多的系统和应用，例如给出租车司机推荐接乘客的地点[1]，从轨迹数据中发现流行的路径[2]，发现城市中的异常轨迹[3]等。这些应用利用数据挖掘和普适计算的技术，促进了智慧城市的发展。

(二)论文主要工作

由于 GPS 定位技术的广泛应用，各种安装有 GPS 定位仪设备采集了海量的 GPS 轨迹数据。这些轨迹数据隐含了丰富的信息，促进了城市交通模式挖掘的发展。然而，很多轨迹挖掘的工作着眼于交通分析、流量监测等，只能从统计的角度挖掘出轨迹数

据的模式，例如：“晚上 6 时有大量轨迹在某区域聚集”[4]。但是，这些挖掘工作有两个明显的问题：i）缺乏内在的、隐含的语义信息，例如在晚上 6 时、7 时的大量轨迹的出现，往往意味着晚高峰，可以归为一类，而不是单独的个体；ii）挖掘的模式往往是确定的，而不是基于概率的，在实际上，因为人们的活动复杂多变，不可能只有单一的模式来主导。

在本文中，我们提出了一种由数据驱动的框架，它的目的是挖掘从大量有 GPS 设备的车辆产生的区域性的、集体的交通模式。这些交通模式反映出潜在的语义信息，可以用于理解人们的生活模式。

为了达到上述目的，解决我们提出的两个问题，本文提出的方法基于一种概率主题模型，即隐性狄利克雷分配（Latent Dirichlet Allocation，简称 LDA）[5]。LDA 常用于文本的建模，它将文档看成隐含的主题的分布，将隐含的主题看成词的分布，同时利用狄利克雷先验分布，对实际的语料有很好的拟合。

本文在 LDA 的基础之上，提出了在交通模式挖掘的框架下的模型表示，即怎样在海量轨迹数据集中，找到相应的处理方法，将轨迹的模式表达为“文档”和“词”。为了高效地处理轨迹数据，我们给出了在 MapReduce 框架下的实现算法，使之能在分布式集群上实现快速计算。

由于存在不能被观测到的隐含主题，所以无法直接从数据中得到相应的概率分布。我们利用 Gibbs 采样算法[6]，从数据中进行采样，利用采样的结果学习模型参数。此外，对于之前未知的数据，为了提高模型通用性，避免模型过拟合，我们给出了在原有 LDA 模型上的推理算法。利用上述算法，可以发现在轨迹数据中隐含的主题分布，这些隐含的主题具有一些语义上的含义，反映出人们的城市生活特征。

最后，为了体现出本文提出的方法的有效性，我们在超过 50GB 的真实轨迹数据集上进行了实验，这些数据来自北京市超过 12000 辆出租车 1 个月的 GPS 轨迹。我们在这些轨迹进行了详尽的实验。在实验结果中，我们详细地分析了挖掘的主题分布，对实验结果进行了分析。我们的实验结果显示了本文提出的算法和框架能够挖掘出海量轨迹数据中的隐含主题，并且能对不同区域和不同时间范围内人的活动特征进行分析。它能更好地理解人们在城市的交通和生活模式，为城市的规划和发展提供了宝贵的支撑点。

(三)相关工作

在近年来，在交通模式挖据的方面有很多研究的相关工作。根据研究工作的具体方向不同，以及和我们工作的关系。现在把相关工作细分成三个方面，分别进行仔细介绍，然后与本文的工作相比较。

1. 轨迹模式挖据

不同的移动设备，如智能手机、个人导航设备、平板电脑等设备在近年来在我们的生活中越来越流行。这些设备用卫星导航系统定位，比如 GPS、Wi-Fi、基于 RFID 的系统等。这些设备在商务和个人的不同场合广泛使用，积累了大量的轨迹数据。对这些轨迹数据的处理和挖掘产生了大量的应用。

交通优化应用[7]需要找到相似轨迹的集合，这些轨迹体现出不同的物体正在一起移动中。举例来说，拼车的应用把在同一轨迹的司机联系起来，这样，他们可以拼车来降低费用。物流的应用把送货的卡车组合在一起，进行更好的规划。

利用轨迹数据可以利用预测的方法来理解物体的行为特征[8]。通过利用预测算法，我们可以提供更加优质的服务，例如更加精准地向目标受众发送广告，向顾客提供基于位置的服务。

轨迹模式挖掘在科学研究上具有广泛的应用，例如识别动物群体的移动。这在发现动物移动模式上很有用，如黄蜂、鸟类、海龟、鱼、鲸等[9]。挖掘这些模式能研究动物的行为模式。类似的，社交分析研究可以利用类似的方法从人们的移动模式中识别社会经济模式[10]。

团队体育活动（例如足球、棒球、垒球等）产生了宝贵的轨迹数据来体现运动员的移动轨迹[11]。研究一场比赛中轨迹可以让我们更好的认识比赛，分析比赛中用的战术，甚至抽取使用某种战术的时间和地点。

交通分析应用能使用轨迹的集合去研究人群和离群点。在这种场景下，一个移动物体可以是路上的车辆或者人行道上的行人，大量的轨迹的聚集很可能是人群的行为[12]。通过从轨迹中识别人群，我们可以更好地认识他们的行为，如人们聚集和散开的时间和地点。这些信息能用来高效的管理交通基础设施。

此外，挖掘离群点也很有意义[13]。离群点不属于任何轨迹群，能用来发现和去除轨迹数据中的错误，例如发现 GPS 接收器失效的设备。它也能用来识别危险的驾驶行为。

2. 主题模型

在信息检索（information retrieval）领域，主题模型（topic model）常用于对自然

语言文本进行建模，从而发现文本中的隐含的语义主题，进行文本聚类和维度规约等。最著名的主题模型有隐含语义分析（Latent Semantic Analysis，简称 LSA）[14]、概率隐含语义分析（Probabilistic Latent Semantic Analysis，简称 PLSA）[15]和隐性狄利克雷分配（Latent Dirichlet Allocation，简称 LDA）[5]等。

LSA 由 Deerwester 等人提出[14]，利用奇异值分解（Singular Value Decomposition，简称 SVD）[16]的方法对文档-词矩阵进行维度规约，维度规约后的特征可以包含原有文本的语义信息。然而，SVD 用纯数学的方法进行维度规约，所以，我们很难从语义的角度解释 LSA 的结果。这使得 LSA 很难解决本文提出的问题。

Hofmann 极大地改进了 LSA，从概率的角度提出了 PLSA。PLSA 把每个文档建模成若干个主题的混合，每个主题是词上的概率分布，每个词都是由一个固定的主题生成[15]。PLSA 的主要缺点在于参数的数量与训练集的大小相关，它没有对主题和词的先验分布进行建模。所以，PLSA 没有办法处理在训练集中没有的数据，容易过拟合。

LDA 由 Blei 等人提出，用来建模文档集中一系列未被观测到的主题。为了描述文档中的隐含的主题，LDA 认为每个主题是词的分布，每个文档是主题的分布，这些分布可以从狄利克雷分布采样到[5]。所以，LDA 的参数空间与训练集的大小无关。本文中，我们采用 LDA 的方法进行区域交通模式挖掘。与现有工作不同的是,我们把地理上的区域看成文档，把区域内的交通行为特征看成词，来挖掘轨迹数据中隐含的主题。

3. 行为模式识别

从轨迹数据中识别人们的行为模式和本文提出的工作有紧密的关系。这些工作可以分为有监督的学习和无监督的学习。

在有监督的学习方面，很多工作利用轨迹数据进行识别和预测。其中，Zheng 等人利用决策树和序列平滑方法来识别 GPS 轨迹数据的交通模式[17]，例如开车、骑自行车、坐公交车和走路。为了准确地预测交通模式，首先将 GPS 轨迹切成若干段，对于每一段，根据轨迹数据的特征提取特征向量，利用决策树进行预测。此外，为了提高准确度，利用平滑的方法进行后处理。

Yin 等人使用动态贝叶斯网络（Dynamic Bayesian Network，简称 DBN）[18]进行基于位置的活动识别。他们设计了双层的贝叶斯网络，首先在下层估计用户位于的实

际位置，其次在上层利用预测的结果进一步预测用户的具体活动。用户的位置、活动和目标都是隐变量，可以从原始数据中推断出来。

此外，Liao 等人使用条件随机场（Conditional Random Field，简称 CRF）利用 GPS 数据进行基于位置的活动识别[19]。他们将模型分成三个层次：第一个层次是 GPS 轨迹，属于原始数据；第二个层次是具体的活动，例如走路、开车、睡觉；第三个层次指的是重要的位置，例如人们的家或者工作场所等。整个识别过程被整合成一个层次条件随机场中。

有监督的学习方法精度比较高，研究工作比较多。然而，这些方法不能直接运用到本文提出的问题中。本文意在海量轨迹数据中进行挖掘，无法进行人工标注。

在无监督的学习方面，Eagle 和 Pentland 使用主成分分析（Principal Component Analysis，简称 PCA）[20][21]来识别人们日常生活行为的主要组成部分。这些组成部分体现了日常生活的特征，即为 eigenbehavior，排序较高的 eigenbehavior 显示出生活的主要活动，如晚上在家，早晨上班等；而排序较低的 eigenbehavior 只能显示个别人或偶尔出现的生活模式。

与本文工作较接近的是由 Farrahi 和 Gartica-Perez 提出的[22]。他们使用 LDA 在大规模人们活动的轨迹中发现日常的生活模式。在文中，将人们一天的生活记录（包括时间和地点）分别用两种不同的方式表示成词的集合，将连续若干天的记录建模成文档集。除了 LDA，还是用了作者主题模型（Author Topic Model，简称 ATM）[23]进行建模，并在结果上与 LDA 做了比较。

本文的工作同样基于无监督的学习，然而与上述工作不同的是，本文重在对海量轨迹数据集进行处理，从一个宏观的角度发现交通的模式，从而了解人们的城市生活模式。

(四)论文组织

本文的组织结构如下：在第一章中，本文主要介绍了研究的主要背景，特别是海量 GPS 轨迹数据挖掘的工作。我们系统地概括了本文的研究工作和相应的贡献。此外，我们从三个方面详细地介绍了相关工作，分别是轨迹模式挖掘、行为模式识别和主题模型，并与本文提出的框架进行比较。在第二章中，我们介绍了 GPS 数据的建模问题。在第三章和第四章中，本文分别从模型表示、模型学习和模型推理三个方面详细说明如何在隐性狄利克雷分配的基础上解决行为模式挖掘的问题。特别地，在第

三章中，我们详细说明了在轨迹挖掘框架下的词和文档生成问题，以及隐性狄利克雷生成过程。为了解决数据量大的问题，我们给出了基于 MapReduce 的相应算法，用于分布式计算。在第四章中，我们采用 Gibbs 采样的方法对提出的模型进行学习和推理，并且提出在轨迹挖掘框架下的应用，来挖掘轨迹数据中隐含的主题信息。第五章详细描述了本文使用的 GPS 轨迹数据集、具体的实验分析方法和实验结果，并对本文提出的算法的有效性进行了分析。最后，在第六章中，我们总结了全文，并提出了对未来的展望。

二、 轨迹数据的数学模型

(一)GPS 轨迹数据

近年来，GPS 定位设备的应用促使了交通模式挖掘的发展。如今，大城市的出租车上都有 GPS 传感器，它们以一定频率将自己的位置信息传给数据中心，这些数据构成了海量空间轨迹。举例来说，在北京，有超过 6 万辆装有 GPS 定位终端的出租车，每天生成数万条轨迹，这些出租车每天生成的轨迹数据就超过 10GB。这些轨迹数据反映出城市的建设、车辆的移动和人的生活。

为了更好地体现轨迹数据的作用，我们用 ArcGIS[1]对其进行可视化。图 2-1 和图 2-2 分别是北京市 50 辆和 200 辆出租车在 4 天形成的轨迹可视化效果。

图 2－1 轨迹数据可视化 I

Figure 2-1 Visualization of Trajectory Data I

[1] ArcGIS 是一套地理信息系统平台，用于地图制作、空间数据分析和管理等功能，参见 http://www.arcgis.com/。

图 2－2 轨迹数据可视化 II

Figure 2-2 Visualization of Trajectory Data II

从图 2-1 中可以看出，50 辆出租车在 4 天内的轨迹已经覆盖了北京的主干道，勾勒出了北京路网的主要信息。我们把注意力集中到北京市区，在图 2-2 中，200 辆出租车在 4 天内的轨迹覆盖了大小街道。

为了进一步体现出 GPS 轨迹数据对城市交通的描述作用，我们考虑城市路网的信息[2]。在图 2-3 中，我们用 ArcGIS 将图 2-2 中的数据和北京市的路网叠加在一起。从图 2-3 和图 2-4 中可以发现，北京市区的道路基本上被覆盖，在郊区，主干道的形状和 GPS 轨迹整齐地叠加在一起。

[2] 北京市完整路网数据见 http://www.datatang.com/data/43855。

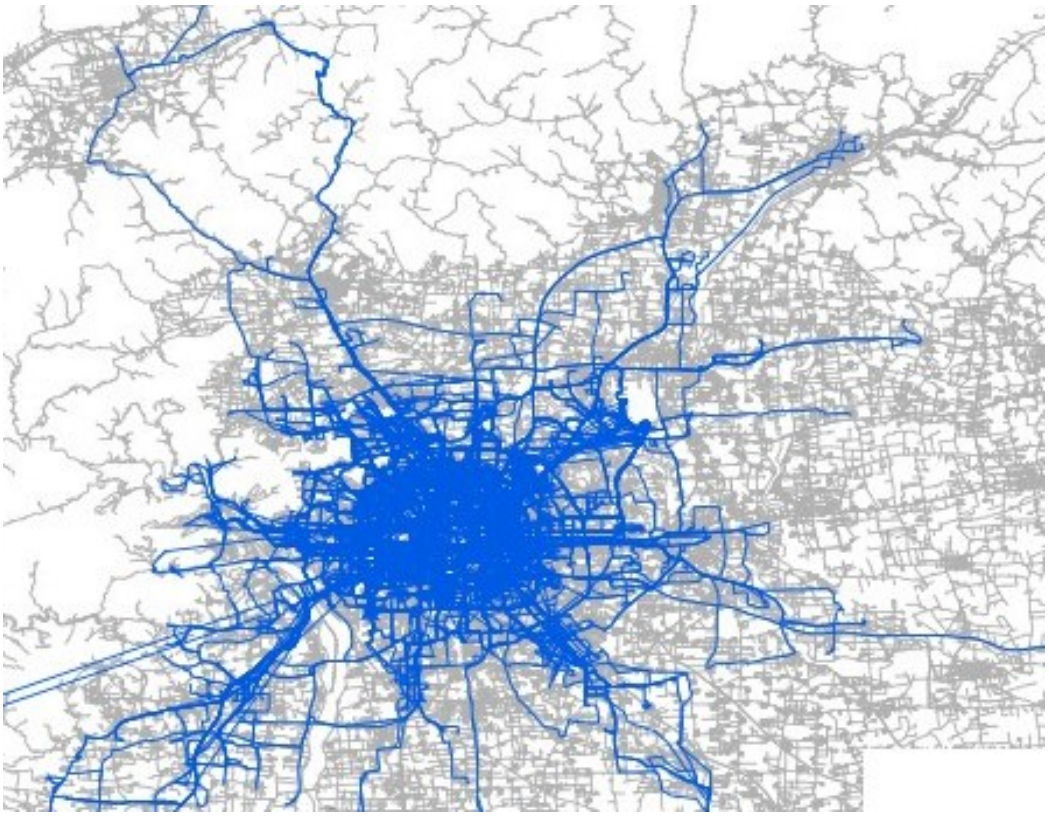

图 2－3 轨迹数据和路网的可视化 I

Figure 2-3 Visualization of Trajectory Data and Road Network I

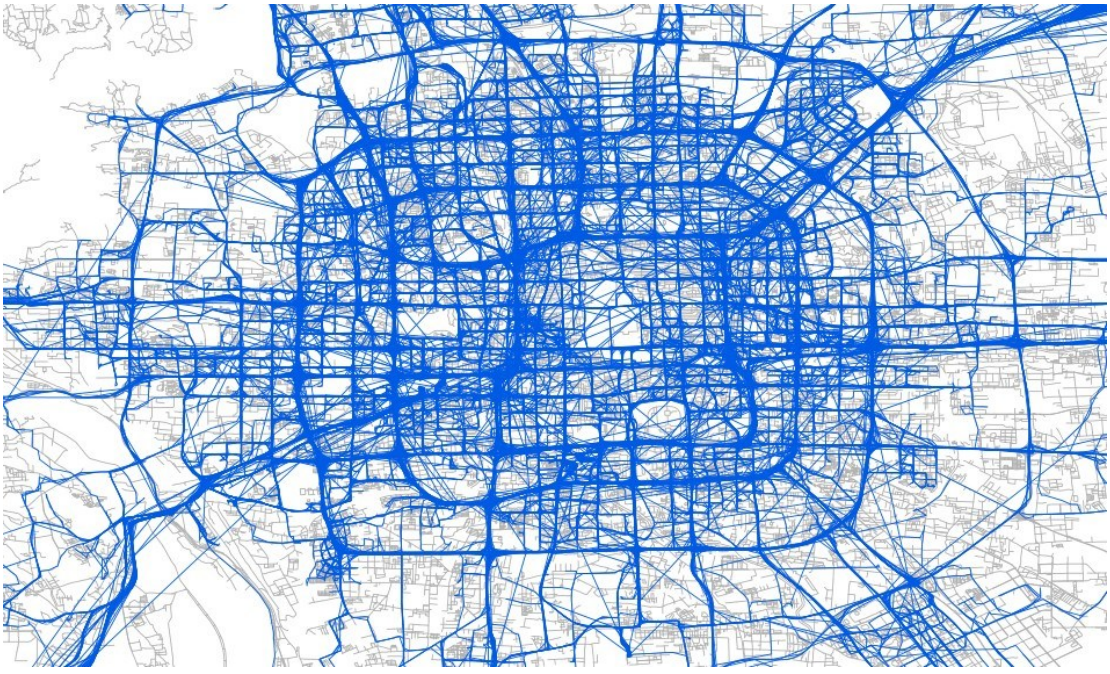

图 2－4 轨迹数据和路网的可视化 II

Figure 2-4 Visualization of Trajectory Data and Road Network II

在以下章节，我们对 GPS 轨迹数据进行建模，并给出详细定义。

(二)轨迹数据的建模与定义

在本节中，我们对 GPS 轨迹数据进行建模，对相关概念给出定义。

定义 1 （GPS 轨迹点） 一个 GPS 轨迹点p是一个结构体，它含有若干属性，包括 ID、时间戳、经度、纬度、状态等。符号和示例表示如下：

表 2－1 GPS 轨迹点示例

Table 2-1 Example of a GPS Trajectory Point

属性	符号	示例	说明
ID	$p.id$	431498	
时间戳	$p.timestamp$	20121101095636	北京时间，格式为 yyyymmddhhnnss
经度	$p.lontitude$	116.4243011	以度为单位
纬度	$p.latitude$	40.0727348	以度为单位
状态	$p.state$	VACCANT	VACCANT（空车）、OCCUPIED（载客）

值得说明的是，在本文中，我们考虑的是出租车的 GPS 轨迹，VACCANT（空车）和 OCCUPIED（载客）指的是出租车是否载客。

定义 2 （GPS 轨迹） 一条 GPS 轨迹$Traj$是由一个个 GPS 轨迹点的有序列表，即为

$$Traj = \{p_0, p_1, \cdots, p_k\}$$

其中，对于任意$0 \le i \le k$和任意$0 \le j \le k$，我们规定

$$p_i.id = p_j.id$$

对于任意任意$0 \le i < j \le k$，我们规定

$$p_i.timestamp < p_j.timestamp$$

即 GPS 轨迹点按时间戳先后顺序排序。

从一条 GPS 轨迹中，除了可以得到状态信息，如空车和载客之外，还可以得到动作信息，如乘客的上客和下客。

定义 3 （上客点） 一条 GPS 轨迹$Traj$中，一个上客点p_i为状态从空车变为载客的点，即

$$p_{i-1}.\text{state} = \text{VACCANT}$$

$$p_i.\text{state} = \text{OCCUPIED}$$

定义 4 （下客点） 一条 GPS 轨迹$Traj$中，一个下客点p_j为状态从载客变为空车的点，即

$$p_{j-1}.\text{state} = \text{OCCUPIED}$$

$$p_j.\text{state} = \text{VACCANT}$$

此外，我们可以考虑加上位置、时间限制的轨迹。

定义 5 （区域） 一个区域S是地理位置上的一块矩形区域，由其经纬度所限定，即为

$$S = \{(x,y)|lontitide_i \le x \le lontitude_j, latitide_i \le y \le latitude_j\}$$

定义 6 （关于区域S的 GPS 子轨迹） 一条关于区域S的 GPS 子轨迹$Traj_S$是一条 GPS 轨迹的片段，即为

$$Traj_S = \{p_0, p_1, \cdots, p_n\}$$

其中对于任意0 ≤ i ≤ n，我们有

$$(p.lontitude, p.latitude) \epsilon S$$

定义 7 （时间段T内关于区域S的 GPS 子轨迹） 一条时间段$T = (t_{START}, t_{END})$内关于区域S的 GPS 子轨迹$Traj_S$是一条 GPS 轨迹的片段，即为

$$Traj_{S,\text{T}} = \{p_0, p_1, \cdots, p_n\}$$

其中对于任意0 ≤ i ≤ n，我们有

$$(p.lontitude, p.latitude) \epsilon S$$

对于时间的限制，满足

$$t_{START} \le p_0.timestamp < p_n.timestamp \le t_{END}$$

三、 交通模式的模型表示

(一)LDA 模型表示

1. LDA 简介

隐性狄利克雷分配（Latent Dirichlet Allocation，简称 LDA）是一种概率主题模型，由 Blei 等人提出，最初用于自然语言处理等领域[5]。LDA 是一个三层的层次贝叶斯模型，它基于如下假设：每一篇文档（document）由有限个隐含的主题（topic）混合构成，而每一个主题由有限个词（word）混合构成。LDA 可以用于对大规模文本语料库（corpus）进行建模，用于发现在文本语料库中隐含的主题，这些主题中包含的单词在语义上有其相似性。利用 LDA，可以自动地、无监督地发现隐含的主题及与每个主题相关的词。

虽然 LDA 被广泛地应用于文本建模的领域，然后，LDA 可以应用于所有满足上述假设的情景。在本文中，我们使用“词”、“文档”和“语料库”的含义可以扩展如下：

定义 8 （词） 一个词w是一个离散的数据单元。

定义 9 （文档） 一个文档$\boldsymbol{w} = (w_1, w_2, \cdots, w_N)$是一个长度为$N$的向量，向量的每个元素都为一个词。

定义 10 （语料库） 一个语料库$D = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \cdots, \boldsymbol{w_M}\}$是一个由$M$篇文档构成的集合。

在不同的领域，“词”、“文档”和“语料库”可以有不同的表现形式，并不局限于文本。在本文中，我们将定义在轨迹挖掘情境下的“词”、“文档”和“语料库”等概念。

2. LDA 模型

LDA 是一个生成的（generative）概率模型[27]。在本节中，我们简要介绍在 LDA 中的概率分布。在 LDA 中，我们假设语料库D中有M篇文档和K个主题，第i篇文档有N_i个词。

LDA 的图示参见图 3－1，方块表示模板，圆圈表示随机变量，箭头表示依赖关系。其中，灰色圆圈表示可观测的随机变量，白色圆圈表示不可观测的随机变量。

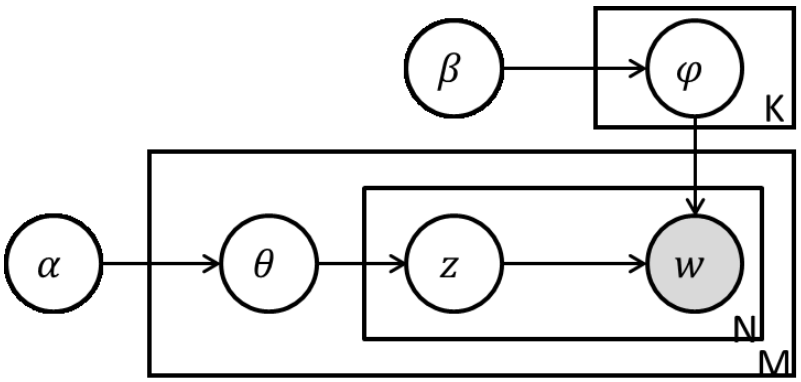

图 3－1 LDA 模型表示

Figure 3-1 Representation of LDA Model

文档的主题由$\boldsymbol{z}$表示，K个主题满足多项分布（multinomial distribution），称为文档-主题分布[5]；第i篇文档的主题分布$\boldsymbol{z_i}$的参数为$\boldsymbol{\theta_i}$，记为$\boldsymbol{z_i}\sim Multinomial(\boldsymbol{\theta_i})$。 同样地，给定主题，对应的词的分布也满足多项分布，称为主题-词分布[5]；第j个主题词分布$\boldsymbol{w_j}$的参数为$\boldsymbol{\phi_j}$，记为$\boldsymbol{w_j}\sim Multinomial(\boldsymbol{\phi_j})$。

在 LDA 中，我们把参数$\boldsymbol{\theta_i}$和$\boldsymbol{\phi_j}$同样看成随机变量。$\boldsymbol{\theta_i}$满足狄利克雷分布（Dirichlet distribution），参数为$\boldsymbol{\alpha}$，记为$\boldsymbol{\theta_i}\sim Dirichlet(\boldsymbol{\alpha})$。同样地，$\boldsymbol{\phi_j}$也满足狄利克雷分布，参数为$\boldsymbol{\beta}$，记为$\boldsymbol{\phi_j}\sim Dirichlet(\boldsymbol{\beta})$。在 LDA 中，我们常常将$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$称为超参数（hyperparameter）。我们在附录中，详细推导了多项式分布和狄利克雷分布的关系及其相关性质。

根据附录，我们知道，K维随机变量$\boldsymbol{\theta_i}$，其分布的参数为$\boldsymbol{\alpha}$，则它的概率密度函数（probability density function）的形式为

$$\mathrm{p}(\boldsymbol{\theta_i}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{k=1}^{K}{\theta_{i,k}}^{\alpha_k-1}$$

同理，与之类似，L维随机变量$\boldsymbol{\phi_j}$，其分布的参数为$\boldsymbol{\beta}$，则它的概率密度函数（probability density function）的形式为

$$p(\boldsymbol{\phi_j}|\boldsymbol{\beta}) = \frac{\Gamma(\sum_{i=1}^{l}\beta_i)}{\prod_{i=1}^{l}\Gamma(\beta_i)}\prod_{l=1}^{L}{\varphi_{j,l}}^{\beta_l-1}$$

3. LDA 的联合分布率

LDA 对文档、主题的联合分布律进行建模。在 LDA 中，联合分布率如下所示：

$$p(\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}) \cdot p(\boldsymbol{z}|\boldsymbol{\alpha})$$

若语料库中第k个主题在第m篇文档的出现次数为$n_{m,k}$，所有词汇中第t个词的被分为第k个主题计数$n_{k,t}$，主题-词分布$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})$为：

$$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}) = \int p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\varphi})p(\boldsymbol{\varphi}|\boldsymbol{\beta})d\boldsymbol{\varphi}$$

根据附录中狄利克雷分布的性质，我们得到：

$$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}) = \int \prod_{z=1}^{K} \frac{1}{\Delta(\boldsymbol{\beta})} \prod_{t=1}^{V} {\varphi_{z,t}}^{n_{z,t}+\beta_t-1}\, d\,\boldsymbol{\varphi}_z = \prod_{z=1}^{K} \frac{\Delta(\boldsymbol{n}_z+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}$$

其中$\boldsymbol{n}_z = \{n_{z,t}\}_{t=1}^{V}$，$\Delta(\cdot)$为狄利克雷分布中的归一化参数，详见附录。

同理，主题分布$p(\boldsymbol{z}|\boldsymbol{\alpha})$产生方式比较类似，为：

$$p(\boldsymbol{z}|\boldsymbol{\alpha}) = \int p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta} = \int \prod_{m=1}^{M} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{K} {\theta_{m,k}}^{n_{m,k}+\alpha_k-1}\, d\,\boldsymbol{\theta}_m = \prod_{m=1}^{M} \frac{\Delta(\boldsymbol{n}_m+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}$$

其中$\boldsymbol{n}_m = \{n_{m,k}\}_{k=1}^{K}$。

则 LDA 的联合分布律为：

$$p(\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{z=1}^{K} \frac{\Delta(\boldsymbol{n}_z+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta(\boldsymbol{n}_m+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}$$

因为 LDA 对文档和主题的联合分布率进行建模（$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$为超参数），所以 LDA 为概率生成模型。

4. LDA 的文档生成过程

在给出数学表达后，我们说明 LDA 的文档生成过程。假设语料库D中有M篇文档和K个主题，第i篇文档有N_i个词。语料库D的生成过程如下：

1. 对于 $i\epsilon\{1,2,\cdots,M\}$，采样得到主题分布$\boldsymbol{\theta_i} \sim Dirichlet(\boldsymbol{\alpha})$
2. 对于 $k\epsilon\{1,2,\cdots,K\}$，采样得到主题下的词分布$\boldsymbol{\phi_k} \sim Dirichlet(\boldsymbol{\beta})$
3. 对于每篇文档的每个词$w_{i,j}$，其中 $i\epsilon\{1,2,\cdots,M\}$，$j\epsilon\{1,2,\cdots,N_i\}$

   首先对主题采样$z_{i,j} \sim Multinomial(\boldsymbol{\theta_i})$

   确定主题$z_{i,j}$后，对词采样$w_{i,j} \sim Multinomial(\boldsymbol{\phi}_{\boldsymbol{z_{i,j}}})$

(二)交通模式的文档表示

在本节中，我们给出对于本文中提出的框架中，针对交通模式挖掘领域的文档的表示方法。

一个海量的轨迹数据集可以看成由k条轨迹构成，即为

$$D_{Traj} = \{Traj_1, Traj_2, \cdots, Traj_k\}$$

然而，一条轨迹可以经过多个区域。首先对D_{Traj}中所有轨迹数据点的经纬度进行统计分析，得出这些轨迹点的经纬度范围。即对于任意轨迹点p，满足

$$lontitide_{START} \leq p.lontitude \leq lontitude_{END}$$

$$latitide_{START} \leq p.latitude \leq latitude_{END}$$

这个经纬度范围构成了一个矩形，我们将这个矩形平均划分成$M \times N$个全等的矩形，每个矩形是一个区域[3]。整个矩形可以看成$M \times N$个区域构成的集合：

$$\mathbb{S} = \{S_1, S_2, \cdots, S_{M \times N}\}$$

在本文提出的框架中，一个语料库即为区域的集合$\mathbb{S}$。在这个语料库中，每篇文档即为一个区域S_i。

为了便于理解，我们给出一个示例。在图 3-2 中，a_i为经度值，b_i为纬度值，S_i为区域。图 3-2 中共有四个区域，以经纬度作为它们的边界。则S_1的边界点为(a_1, b_1)，(a_2, b_1)，(a_1, b_2)和(a_2, b_2)，其余区域依次类推。如图 3-4，整个区域可以看成一文档集，包含S_1至S_4四个文档。

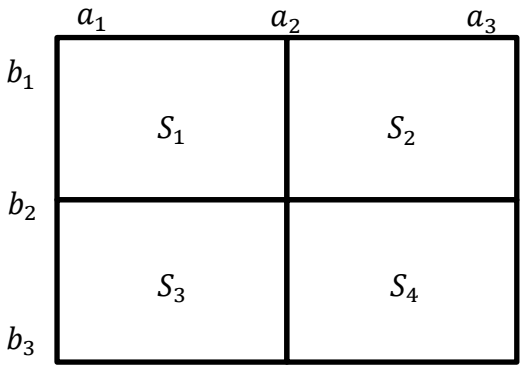

图 3－2 区域示例

Figure 3-2 Example of Regions

(三)交通模式的词表示

[3] 实际上，经纬度不同，相同经纬度差表示的实际地理距离可能因经纬度而异。然而，当范围比较小的时候（本文考虑北京市区），这种差异性可以忽略不计。我们可以近似地认为，这些矩形区域的长度和宽度都相同，即面积是相等的。

我们说明文档表示之后，考虑文档中词的表示。在一个区域S_i中，我们将一种描述交通的行为(behavior)看成一个词。具体地说，在区域S_i中，我们可以得到关于S_i的子轨迹数据集

$$D_{S_i} = \{Traj'_1, Traj'_2, \cdots, Traj'_k\}$$

对于这些子轨迹，我们需要提取统计量，来表征这个区域交通的活跃程度。单纯考虑这些子轨迹的数量是不合理的，例如下图：

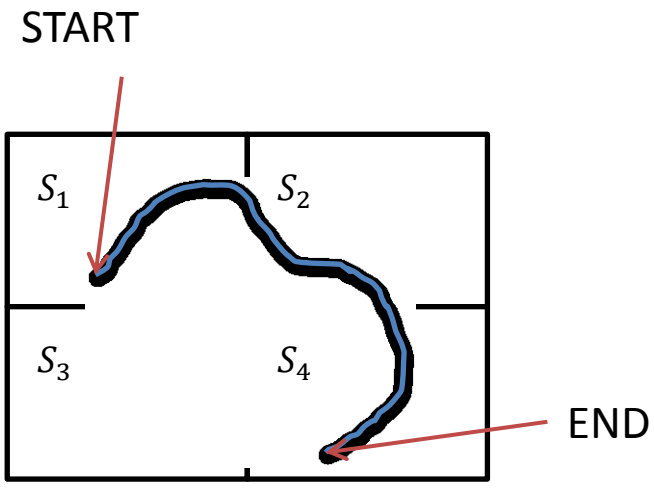

图 3－3 轨迹示例

Figure 3-3 Example of a Trajectory

在图 3-3 中，轨迹从S_1开始，途径S_2，最后到S_4。在本文的数据集中，对于一条出租车的载客轨迹，START和END分别为上客点和下客点。上客点和下客点可以表征人们的行为特征，在图中为S_1与S_4，而S_2只是中途路过，不能反映出人们实际的外出的起始点和目标。在本文中，我们要挖掘的是交通行为中，隐含的人们生活模式，所以，我们只考虑S_1与S_4。

特别地，在本文中，我们考虑在某个时间片段$T = (t_{START}, t_{END})$中，对于区域S_i，交通热度可以表示为所有子轨迹中的上客点和下客点的数量总和。我们只考虑时间段在T之间的子轨迹：

$$D_{S_i,\ T} = \{Traj''_1, Traj''_2, \cdots, Traj''_l\}$$

详细定义如下：

定义 11 （**交通热度**）在某个时间段$T = (t_{START}, t_{END})$ 中，对于区域S_i，交通热度定义为

$$H_{T,\ s_i} = \sum_{Traj \in D_{S_i,\ T}} \sum_{p_i \in Traj} (\delta(p_{i-1}.state = VACCANT \wedge p_i.state = OCCUPIED) + \delta(p_{i-1}.state = OCCUPIED \wedge p_i.state = VACCANT))$$

其中$\delta(\cdot)$为指示函数（indicator function），输入为一个二值函数，如果返回值为真，函数值为 1，否则，函数值为 0，即为

$$\delta(p) = \begin{cases} 1, & p = true \\ 0, & p = false \end{cases}$$

$H_{T,\ s_i}$的分布是明显不平衡的。在某些热门地域，如商业区，其值非常大；而在其他区域，如郊区，其值很小。在实际生活中，由于热门地域很稀少，而郊区较多，所以分布是明显不平衡的，所以，我们在原有交通热度的基础上，定义了交通热度级。

定义 12（交通热度级） 交通热度级是交通热度的离散化表示，将交通热度值映射为{XL, VL, L, M, H, VH, XH}。具体映射规则见表 3-1：

表 3－1 交通热度级

Table 3-1 Traffic Intensity Levels

交通热度	交通热度级	全称
[0,2)	XL	Extremely Low
[2,5)	VL	Very Low
[5,10)	L	Low
[10,15)	M	Medium
[15,20)	H	High
[20,50)	VH	Very High
[50, +∞)	XH	Extremely High

在定义上述概念之后，我们提出针对交通模式挖掘领域的词的表示。一个词是一个二元组，由时间片段和交通热度级构成。 在一个区域中，包含不同时间段的不同交通热度级，这可以类比为 LDA 中文档和词的概念。

综上所述，我们给出本框架中“词”、“文档”和“语料库”的定义。

定义 13 （区域交通模式中的词） 一个区域交通模式中的词$w = (T, Level)$是一个包括时间片段和交通热度级的数据二元组。

定义 14 （区域交通模式中的文档） 一个区域交通模式中的文档$\boldsymbol{w} = (w_1, w_2, \cdots, w_N)$是一个长度为$N$的向量，是一个区域内所有上述二元组的集合。

定义 15 （区域交通模式中的语料库） 一个区域交通模式中的语料库$D = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \cdots, \boldsymbol{w_{MXN}}\}$是一个由$M \times N$个区域构成的文档的集合。

(四)交通模式挖掘的文档生成过程

为了便于理解，我们给出一个示例来解释整个过程。在某区域内，我们对经过该区域的次轨迹进行统计分析。其中，我们把每个时间段设为 1 小时，则$T = \{0,1,2,3, \cdots 23\}$。将统计得到的交通热度值按表 3-1 转化成交通热度级，则$Level = \{XL, VL, L, \cdots XH\}$。表 3-2 给出一个简单的例子[4]：

表 3－2 统计量示例

Table 3-2 Example of Statistics

区域 ID	日期	小时	交通热度级
REGION_0_1	2013.11.1	0	VL
		1	VL
		2	XL
	2013.11.2	0	XL
		1	VL
		2	VL
REGION_0_2	2013.11.1	0	L
		1	VL
		2	XL
	2013.11.2	0	VL
		1	L
		2	M

根据表 3-2 的数据，我们可以生成如下文档，见表 3-3。例如，文档 ID 是区域的编号，每个区域的编号都是唯一的。文档内容是这些二元组的集合。

[4] 表 3-2 和表 3-3 的例子并非产生自真实数据，仅供解释和说明用。

表 3－3 文档示例

Table 3-3 Example of Documents

文档 ID	文档内容
REGION_0_1	(0, VL) (1, VL) (2, XL) (0, XL) (1, VL) (2, VL)
REGION_0_2	(0, L) (1, VL) (2, XL) (0, VL) (1, L) (2,M)

对于所有区域的所有时间都进行上述转换，我们可以利用轨迹数据集，生成语料库，以供 LDA 求解。然而，由于 GPS 轨迹数据规模一般较大，在下一节，我们介绍在分布式集群上的实现技术。

(五)分布式实现方法

本章节介绍从海量轨迹数据集中生成文档和词的实现方法。由于 GPS 轨迹数据一般具有采样密集、数据量大的特点。从统计中发现，在一个大型城市，出租车每天都会生成超过 1GB 的数据。所以，基于单机的实现对于本文涉及的轨迹处理并不合适。为了解决这个问题，本文采用 MapReduce 计算框架来完成。

1. MapReduce 简介

MapReduce 是一种编程模型，用来处理海量数据集，它可以在集群上运行并行的、分布式的算法，它有两个主要阶段，分别为 Map（映射）阶段和 Reduce（化简）阶段[28]。在本文中，我们使用 Apache Hadoop 项目的开源实现[5]。

在 Map 阶段，处理 Map 任务的集群节点接受输入$< K_1, V_1 >$的键值对，进行处理后输出$< K_2, V_2 >$的键值对[28]。在这一过程中，我们可以把一个大的问题切分成很多的小的问题，分给不同的节点分布式地执行。然后相同键K_2的键值对会合并起来，传入 Reduce 阶段。

在 Reduce 阶段，处理 Reduce 任务的集群节点接受输入具有相同键K_2的键值对，值分别为$V_{2,1}$、$V_{2,2}$等，最后输出$< K_3, V_3 >$键值对[28]。在这一过程中，我们把小的次问题的解整合起来，最终解决原来的大问题。

2. 处理轨迹数据的 MapReduce 实现

[5] http://hadoop.apache.org/

以下介绍如何利用 GPS 轨迹数据，生成对应的交通热度级。

（1）Map 阶段

在 Map 阶段，输入为每一条 GPS 轨迹数据，我们对每个区域生成唯一的序号 sid，对每个时间段生成唯一的序号 tid。映射用的键为 sid 和 tid 的组合，值为对应的轨迹数据。实现算法如下：

算法 3-1：生成交通热度级算法（Map 阶段）
输入：轨迹数据集 D 输出：Map 函数输出的键值对<区域时间段序号，轨迹点> 步骤： 1. 令 p 为轨迹数据集 D 的每一个轨迹点 2. 令 sid 为轨迹点 p 的经纬度映射到地图的区域序号 3. 令 tid 为轨迹点 p 的时间戳对应的时间段 4. 令 key 为 sid 和 tid 的组合 5. 输出键值对(key,p)

（2）Reduce 阶段

在 Reduce 阶段，输入为相同 key 的 GPS 轨迹数据的集合。由于 key 相同，保证了这些数据位于的区域和时间段都相同。根据时间戳的先后，首先生成一条条的子轨迹，然后根据各点的状态字段，检测上客点和下客点，最后计算这一区域和时间段的交通热度级。实现算法如下：

算法 3-2：生成交通热度级算法（Reduce 阶段）

输入：键值对<区域时间段序号，轨迹点集合>

输出：Map 函数输出的键值对<区域时间段序号，交通热度级>

步骤：

1. 令 count 为计数器，初始化为 0
2. 令 map 为映射表，键为车辆 ID，值为轨迹
3. 对于轨迹点集合的每个轨迹点 p
   3.1. 令 cid 为轨迹点的车辆 ID
   3.2. 以 cid 为键，加入 map 中 cid 对应的轨迹中
4. 对于 map 中的每个轨迹列表 traj
   4.1. 令 traj_s 是 traj 中的轨迹点按照时间戳排序的列表
   4.2. 对于 traj_s 内的第 i 个轨迹点 p_i
      4.2.1 如果 p_i 为上客点，则 count 自增 1
      4.2.1 如果 p_i 为下客点，则 count 自增 1
5. 令 level 为 count 映射到的交通热度级
6. 令 key 为区域时间段序号
7. 输出键值对(key,level)

2. 文档生成的 MapReduce 实现

海量轨迹数据经过上述处理后，可以得到不同区域、不同时间段的交通热度值级。我们可以根据统计后的数据生成上文提出的“词”和“文档”。因为数据量不大，所以可以在单机上完成，也可以在集群上实现。为了不失完整性，我们给出其 MapReduce 实现版本。

（1）Map 阶段

在 Map 阶段，输入的键为区域序号 sid 和时间段的序号 tid，值为交通热度级。首先，对数据进行解析，得到时间段 timespan。其次，将时间段和交通热度级相结合生成“词”。输出的键为区域序号 sid，值为生成的“词”。实现算法如下：

算法 3-3：文档生成算法（Map 阶段）

输入：键值对<区域时间段序号，交通热度级>

输出：Map 函数输出的键值对<区域序号，词>

步骤：

1. 令 sid 为区域时间段序号中的区域序号
2. 令 timespan 为区域时间段序号中的时间段
3. 令 level 为交通热度级
4. 令 word 为词(timespan, level)
5. 输出键值对(sid,word)

（2）Reduce 阶段

在 Reduce 阶段，输入的键为区域序号 sid，值为“词”的集合。我们将其直接组合起来，就可以完成生成工作。实现算法如下：

算法 3-4：文档生成算法（Reduce 阶段）

输入：键值对<区域序号，词的集合>

输出：Map 函数输出的键值对<区域序号，文档>

步骤：

1. 令 sid 为区域序号
2. 令 document 为空文档
3. 对于词的集合的每个词 word
   3.1. 将 word 加入 document
4. 输出键值对(sid, document)

四、 区域交通模式挖据

(一)交通模式的主题挖掘

本节首先从利用 Gibbs 算法，学习 LDA 模型参数，然后介绍如何在交通模式挖掘的框架下，挖掘轨迹数据中隐含的主题分布和语义信息。

1. 模型的参数学习

在 LDA 模型中，最重要的是联合分布率$p(\boldsymbol{w},\boldsymbol{z})$[6]，只要求得这个联合分布率就可以求解任何模型参数。然而，根据链式法则，

$$p(\boldsymbol{w},\boldsymbol{z}) = p(\boldsymbol{w})\, p(\boldsymbol{z}|\boldsymbol{w})$$

$p(\boldsymbol{w})$为语言模型，可以直接从数据观测到，而$p(\boldsymbol{z}|\boldsymbol{w})$是主题关于词的条件分布率。因为主题为隐变量，观测不到，所以$p(\boldsymbol{z}|\boldsymbol{w})$不能直接求得。

LDA 模型的求解有多种方法，包括变分推理（variational inference）、Gibbs 采样（Gibbs sampling）等[5]。在本文中，我们在 Gibbs 采样的基础上实现区域交通模型的学习和推理。

1. Gibbs 采样

Gibbs 采样是一种采样方法，利用随机模拟的方法来生成样本。Gibbs 采样利用的是马尔可夫链（Markov chain）的平稳分布的性质进行采样[29]。

利用马尔可夫链的细致平稳条件（detailed balance condition）[29]，我们可以对于不同的状态，如$p(X)$和$p(Y)$，构造概率转移矩阵Q，来进行 Gibbs 采样。

$$p(X)Q(X \to Y) = p(Y)Q(Y \to X)$$

在附录中，我们推导了 Gibbs 采样算法。Gibbs 采样可以从二维的情况推广至n维，即对于对于n维随机变量的联合分布率$P(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n)$进行采样。此处不再详述，参见附录。

2. LDA 的 Gibbs 采样方法

在 LDA 中，我们利用 Gibbs 采样的方法对$p(\boldsymbol{z}|\boldsymbol{w})$进行采样。记$i = (m,n)$，第$m$篇文档的第$n$个词为$w_i$，主题为$z_i$。记除了$i$之外其他所有的词为$\boldsymbol{w}_{\neg \boldsymbol{i}}$，所对应的主题为$\boldsymbol{z}_{\neg \boldsymbol{i}}$。

在附录中，我们详细推导了 LDA 的 Gibbs 采样方法。对于文档$\boldsymbol{w}$的第i个词，给定其他词的主题$\boldsymbol{z}_{\neg \boldsymbol{i}}$，第$i$个词的主题为$k$的概率为：

$$p(z_i = k|\boldsymbol{z}_{\neg \boldsymbol{i}}, \boldsymbol{w}) \propto \hat{\theta}_{m,k} \cdot \hat{\varphi}_{k,t}$$

[6] 为了简单起见，我们省略超参数$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$不写。

现在需要对$\theta_{m,k}$和$\varphi_{k,t}$的后验分布进行参数估计。在附录中，我们推导了以狄利克雷分布为先验的多项式分布的贝叶斯估计量：

$$\hat{\theta}_{m,k} = \frac{n_{m,k} + \alpha_k}{\sum_{i=1}^{K}(n_{m,i} + \alpha_i)}$$

$$\hat{\varphi}_{k,t} = \frac{n_{k,t} + \beta_t}{\sum_{n=1}^{V}(n_{k,n} + \beta_n)}$$

其中，第k个主题在第m篇文档的出现次数为$n_{m,k}$，第t个词的被分为第k个主题计数为$n_{k,t}$。所以，我们可以得到 LDA 中 Gibbs 采样的方法：

$$p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{w}) \propto \hat{\theta}_{m,k} \cdot \hat{\varphi}_{k,t} = \frac{n_{m,k} + \alpha_k}{\sum_{i=1}^{K}(n_{m,i} + \alpha_i)} \cdot \frac{n_{k,t} + \beta_t}{\sum_{n=1}^{V}(n_{k,n} + \beta_n)}$$

3. 模型训练

在本节中，我们利用 Gibbs 采样在训练数据集中学习 LDA 模型，它的本质任务是获得文档-主题分布矩阵$\boldsymbol{\theta}$和主题-词分布矩阵$\boldsymbol{\varphi}$，其中，$\boldsymbol{\theta_m}$是第m篇文档的主题分布向量，$\boldsymbol{\varphi_k}$是第k个主题的词分布向量[31]。

在 LDA 的训练过程中，需要获得第k个主题在第m篇文档的出现次数$n_{m,k}$（主题-文档计数器），以及第t个词的被分为第k个主题计数$n_{k,t}$（词-主题计数器）。从上述计数器中，我们可以推知词-主题计数和$n_k = \sum_{t=1}^{V} n_{k,t}$以及主题-文档计数和$n_m = \sum_{k=1}^{K} n_{m,k}$。求得这些计数之后，$\boldsymbol{\theta}$和$\boldsymbol{\varphi}$即可求出。

LDA 的训练过程包括三个阶段：i）初始化阶段，对变量进行初始化并对文档中的词随机指定主题；ii）Gibbs 采样阶段，利用 Gibbs 采样重新得到词的主题，更新相应计数器，不断迭代上述过程；iii）计算模型参数阶段，利用计数器最终的值，计算模型参数。

LDA 的训练过程如下所示：

算法 4-1：LDA 模型学习算法

输入：主题数量K，超参数$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$，迭代次数n

输出：文档-主题分布矩阵$\boldsymbol{\theta}$，主题-词分布矩阵$\boldsymbol{\varphi}$

步骤：

// 初始化阶段

1. 对计数变量进行初始化，即对于所有的$n_{m,k}$和$n_{k,t}$赋值

$$n_{m,k} := 0$$

$$n_{k,t} := 0$$

2. 对于语料库中的每篇文档的每个词w，随机指定一个主题k

$$k \sim Multinomial(\frac{1}{K})$$

对主题k的计数变量进行更新

$$n_{m,k} := n_{m,k} + 1$$

$$n_{k,t} := n_{k,t} + 1$$

// Gibbs 采样阶段

3. 进行n次迭代

3.1. 对于语料库中的每篇文档的每个词w，进行更新

3.1.1 对当前的词w及其对应主题k的计数变量自减

$$n_{m,k} := n_{m,k} - 1$$

$$n_{k,t} := n_{k,t} - 1$$

3.1.2 利用 Gibbs 采样，对主题进行采样

$$p(z_i = k|\boldsymbol{z}_{\neg \boldsymbol{i}}, \boldsymbol{w}) \propto \frac{n_{m,k} + \alpha_k}{\sum_{i=1}^{K}(n_{m,i} + \alpha_i)} \cdot \frac{n_{k,t} + \beta_t}{\sum_{n=1}^{V}(n_{k,n} + \beta_n)}$$

3.1.3 更新对应词w的新的主题k的计数变量

$$n_{m,k} := n_{m,k} + 1$$

$$n_{k,t} := n_{k,t} + 1$$

// 计算模型参数阶段

4. 利用计数变量，估计模型参数

$$n_m := \sum_{k=1}^{K} n_{m,k}$$

$$n_k \coloneqq \sum_{t=1}^{V} n_{k,t}$$

$$\hat{\theta}_{m,k} \coloneqq \frac{n_{m,k} + \alpha_k}{n_m + K\alpha_k}$$

$$\hat{\varphi}_{k,t} \coloneqq \frac{n_{k,t} + \beta_t}{n_t + V\beta_t}$$

5. 输出文档-主题分布矩阵$\boldsymbol{\theta}$和主题-词分布矩阵$\boldsymbol{\varphi}$

2. 交通模式的挖掘

在本节中，我们介绍利用 LDA 模型挖掘交通轨迹中的隐含主题。

从上章的定义中，我们在本文提出的框架内提出了“语料库”、“文档”和“词”的定义。给定模型的超参数，我们认为，对于地理位置上的每一块区域，都有一个区域-主题分布，对于每一个主题，都有一个主题-词分布。每个词是时间段和交通热度级的二元组。

在本文中，进行区域交通模式挖掘主要有两大任务，详细说明如下：

1. 计算主题-词分布

在 LDA 中，每个主题是词的多项式分布。在我们的定义中，每个词包括时间段和交通热度级。我们用一个简单的例子解释它的具体含义，见表 3-4[7]。其中，我们把每个时间段设为 1 小时，则$T = \{0,1,2,3,\cdots 23\}$。为了简便起见，我们只写出前 4 个词。

表 4－1 主题示例

Table 4-1 Example of Topic

词		主题-词分布概率
时间段	交通热度级	
18	H	0.04
19	VH	0.03
20	VH	0.03
21	H	0.02

[7] 表 4-1 的例子并非产生自真实数据，仅供解释和说明用。

在表 4-1 中，我们模拟了一个主题。在这个主题中，我们可以发现从 18 到 21 时之间交通热度级很高。我们可以推断中，这是一个晚高峰的主题。如果一个区域的这个主题比重比较大，这个区域人们在这一时段比较活跃。

2. 计算区域-主题分布

在一个城市中，一般不同的区域内功能不同，生活的人不同，人们的行为模式一般不相同。举例来说，在热闹的商业区，一般在晚上活动比较活跃，这个时候人们打车行为比较多，一直到半夜逐渐减少；在公司聚集的区域，一般上午和下午的早高峰和晚高峰比较明显，和人们上下班的行为相一致；而更多的是郊区，在这些区域，从整体看打车行为都比较少，但不排除有少数情况稍微偏多。上述每一种活动都可以看成是一个隐含的主题。在一片区域，主题并非是单一的，利用 LDA，我们将其假设为多项式分布。利用学习算法，可以找出区域对于主题的分布，从分布中，可以推断城市不同区域的功能、人们的活动模式。

此外，区域-主题分布也可能与日期有关。例如，在工作日，公司聚集的区域整体上人们的活动更加频繁；在周末，商业区活动较为频繁。如果对日期加以考虑，可以计算不同日期的不同分布，有助于我们理解交通模式和时间的关系。我们将在实验中进行详细分析。

(二)交通模式的主题推理

本节首先从经典的 LDA 推断算法出发，讨论基于 Gibbs 采样的算法流程，然后讨论如何在交通模式挖掘的框架下，分析潜在的主题分布和语义信息。

1. LDA 推断算法

LDA 是一个无监督的生成模型，可以对语料库的文档进行主题进行建模。然而，对于训练过程中语料库里没有的文档，无法直接对其建模。在本节中，我们考虑如下问题：对于之前没有遇到的文档$\boldsymbol{w}'$，利用已训练好的 LDA 模型，即文档-主题分布矩阵$\boldsymbol{\theta}$和主题-词分布矩阵$\boldsymbol{\varphi}$，对$\boldsymbol{w}'$进行建模。

我们注意到，对于训练集中没有的文档，文档-主题分布矩阵$\boldsymbol{\theta}$是没有意义的，因为$\boldsymbol{\theta}$中不包含$\boldsymbol{w}'$对主题的分布；而主题-词分布已经求得。我们的目的是给定文档$\boldsymbol{w}'$，利用 LDA 模型，输出这个文档的文档-主题分布。

由于文档$\boldsymbol{w}^{'}$中的词的主题未知，所以，我们不能直接通过参数估计的方法计算参数，我们同样可以用 Gibbs 采样的方式进行求解。我们已经求得

$$p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{w}) \propto \hat{\theta}_{m,k} \cdot \hat{\varphi}_{k,t} = \frac{n_{m,k} + \alpha_k}{\sum_{i=1}^{K}(n_{m,i} + \alpha_i)} \cdot \frac{n_{k,t} + \beta_t}{\sum_{n=1}^{V}(n_{k,n} + \beta_n)}$$

而$\boldsymbol{\varphi}$已知，不需要估计，我们可以将其改为

$$p(z_i = k|z_{\neg i}, w) \propto \hat{\theta}_k \cdot \varphi_{k,t} = \frac{n_k + \alpha_k}{\sum_{i=1}^{K}(n_i + \alpha_i)} \cdot \varphi_{k,t}$$

其中，n_k是文档$\boldsymbol{w}^{'}$中被标为主题k的词的计数。

LDA 的推断过程同样包括三个阶段：i）初始化阶段，对变量进行初始化并对输入文档中的词随机指定主题；ii）Gibbs 采样阶段，利用 Gibbs 采样重新得到词的主题，更新相应计数器，不断迭代上述过程；iii）计算模型参数阶段，利用计数器最终的值，计算文档的主题分布。

所以，我们得到 LDA 的推断算法：

算法 4-2：LDA 模型推断算法

输入：文档$\boldsymbol{w}^{'}$，主题-词分布矩阵$\boldsymbol{\varphi}$，迭代次数n

输出：文档$\boldsymbol{w}^{'}$的主题分布$\boldsymbol{\theta}^{'}$

步骤：

// 初始化阶段

1. 对数变量进行初始化，即对于所有的n_k赋值

$$n_k := 0$$

2. 对于输入文档的每个词w，随机指定一个主题k

$$k \sim Multinomial(\frac{1}{K})$$

对主题k的计数变量进行更新

$$n_k := n_k + 1$$

// Gibbs 采样阶段

3. 进行n次迭代

3.1. 对于输入文档的每篇文档的每个词w，进行更新

3.1.1 对当前的词w及其对应主题k的计数变量自减

$$n_k := n_k - 1$$

3.1.2 利用 Gibbs 采样，对主题进行采样

$$p(z_i = k|z_{\neg i}, w) \propto \hat{\theta}_k \cdot \varphi_{k,t} = \frac{n_k + \alpha_k}{\sum_{i=1}^{K}(n_i + \alpha_i)} \cdot \varphi_{k,t}$$

3.1.3 更新对应词w的新的主题k的计数变量

$$n_k := n_k + 1$$

// 计算模型参数阶段

4. 利用计数变量，估计模型参数

$$n := \sum_{k=1}^{K} n_k$$

$$\hat{\theta}_k := \frac{n_k + \alpha_k}{n + K\alpha_k}$$

5. 输出文档$\boldsymbol{w}'$的主题分布$\boldsymbol{\theta}'$

2. **交通模式的推理**

在本节中，我们介绍利用 LDA 的推断算法，利用对区域内的轨迹数据，分析其主题分布。

对于某区域内的 GPS 轨迹，首先将其映射到训练数据中相同的某区域S，可以使用过滤器将区域外的轨迹点过滤掉。然后，对于这些轨迹，抽取对应的“词”，即对每个时间段计算交通热度值。这些“词”的集合即为一篇“文档”，是 LDA 模型的输入。最后，利用上述基于 Gibbs 采样的推断算法，得出主题分布。

当需要处理的轨迹数据数量不大的时候，我们可以在单机上进行处理；当这些数据也无法在单机上完成计算的时候，我们同样可以使用 MapReduce 框架进行计算。我们可以将上述算法分为两个阶段来实现：第一为文档生成阶段，第二为模型推理阶段。

1. 文档生成阶段

Map 阶段：在 Map 阶段，对输入的每个轨迹点进行检查，检查它是否在区域内。如果在这个区域内，抽取数据点所在的时间段 tid，作为键，而值即为轨迹点本身。实现算法如下：

算法 4-3：交通模式的主题推断算法（Map 阶段）

输入：子轨迹数据集 D

输出：Map 函数输出的键值对<时间段，轨迹点>

步骤：

1. 令 p 为轨迹数据集 D 的每一个轨迹点
2. 令 tid 为轨迹点 p 的车辆 ID
2. 令 inRegion 为 p 的经纬度是否在区域内的布尔变量
   2.1. 如果 inRegion 为真，输出(tid,p)

Reduce 阶段：在 Reduce 阶段，输入为相同 tid 的 GPS 轨迹数据的集合。由于 key 相同，保证了这些数据的和、时间段都相同。根据时间戳的先后，首先生成一条条的子轨迹，然后根据各点的状态字段，检测上客点和下客点，最后计算交通热度级。输出的键为时间段，值为交通热度级。实现算法如下：

算法 4-4：交通模式的主题推断算法（Reduce 阶段）

输入：键值对<时间段，轨迹点>

输出：Map 函数输出的键值对<时间段，交通热度级>

步骤：

1. 令 count 为计数器，初始化为 0
2. 令 map 为映射表，键为车辆 ID，值为轨迹
3. 对于轨迹点集合的每个轨迹点 p
   3.1. 令 cid 为轨迹点的车辆 ID
   3.2. 以 cid 为键，加入 map 中 cid 对应的轨迹中
4. 对于 map 中的每个轨迹列表 traj
   4.1. 令 traj_s 是 traj 中的轨迹点按照时间戳排序的列表
   4.2. 对于 traj_s 内的第 i 个轨迹点 p_i
      4.2.1 如果 p_i 为上客点，则 count 自增 1
      4.2.1 如果 p_i 为下客点，则 count 自增 1
5. 令 level 为 count 映射到的交通热度级
6. 令 tid 为时间段序号
7. 输出键值对(tid,level)

2. 模型推理阶段

在模型推理阶段，只要将每一对 MapReduce 阶段输出的键和值组合起来即为输入的“文档”，载入 LDA 模型后，利用 Gibbs 采样算法进行推理，最后得到文档-主题模型。这个过程和 LDA 模型学习阶段基本相同，所以，在本文中不再赘述。

五、 实验和分析

(一)实验环境

本节简要介绍本文算法实现和实验的具体环境。实验环境包括单机环境和集群环境。我们在海量 GPS 轨迹数据处理方面使用集群环境，我们在单机上实现基于交通轨迹的 LDA 模型学习和推断，以完成相关实验。现将实验环境介绍如下：

1. 单机环境

CPU：Intel(R) Core i7-3632QM @ 2.20GHz （四核心、八线程）

内存：8GB

硬盘：1TB

操作系统：Windows 7 Professional 64bit

JDK 版本：1.6

2. 集群环境

IP 地址：10.11.1.201-10.11.1.210

NameNode：10.11.1.201

DataNode 数量：9

Hadoop 版本：1.0.3

JDK 版本：1.6

(二)数据集

1. 数据源

本文使用由数据堂[8]提供的公开的 GPS 轨迹数据集，其中包括了北京市约 12000 辆出租车在 2012 年 11 月产生了 GPS 数据，记录了出租车在该市的移动轨迹。在我们的实验中，我们采用了时间跨度为 2012 年 11 月 1 日-2012 年 11 月 30 日共 30 天的数据。

2. 数据预处理

为了使数据能合适地用在我们的实验中，我们采用了如下的预处理策略，用以过滤部分轨迹数据。具体操作如下：

1. 字段过滤：我们删除 GPS 状态字段为无效的轨迹数据点，这些轨迹点没有合法的 GPS 数据。

[8] http://www.datatang.com/

2. 时间段过滤：我们删除时间段在 2012 年 11 月 1 日-2012 年 11 月 30 日之外的轨迹数据点，用以去除时间不在实验范围的数据和时间字段有明显错误的数据。
3. 地区过滤：为了对北京市区划分区域，我们只考虑经纬度在东经 116.278 度至东经 116.499 度、北纬 39.834 度至北纬 40 度的轨迹点。由于在北京远郊区范围内 GPS 数据相对稀疏，意义不明显。在本文中，我们只考虑上述范围的数据点，这一范围覆盖了北京市四环以内的区域。

3. 数据集统计量

我们对采用的数据集进行相关的统计量计算，用于更好地理解数据。现将数据统计信息记录如下，见表 5-1：

表 5－1 数据集统计量

Table 5-1 Dataset Statistics

统计量	统计值
轨迹点数	972751200
车辆数	12639
起始时间戳	20121101001447
终止时间戳	20121201001636
数据大小	52.3GB

(三)实验分析

在实验中，我们使用经过预处理的 GPS 轨迹数据集。在区域划分上，我们将北京市经纬度在东经 116.278 度至东经 116.499 度、北纬 39.834 度至北纬 40 度的区域平均划分成 50×50 个矩形区域，其地理范围如图 5-1 所示：

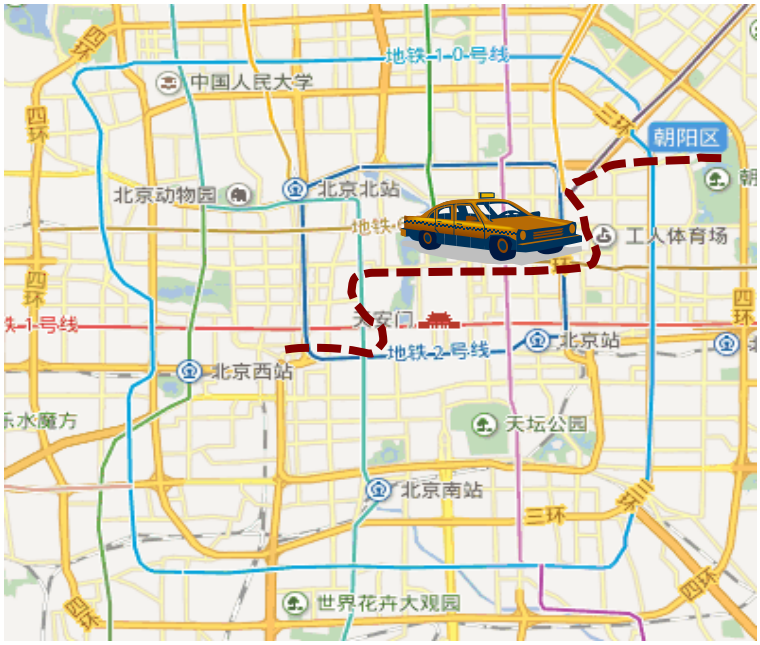

图 5－1 轨迹区域

Figure 5-1 Regions of Trajectories

本文中，我们进行三组实验分析：

i)分析交通轨迹模式的隐含主题，我们选取代表性的主题分析其中的含义及其代表的生活模式。

ii)分析不同类型的区域中的主题分布，从分布中我们比较不同区域模式的区别。

iii)分析不同时间范围（工作日、周末）的隐含主题，从分布中我们比较这些时间范围的模式区别。

1. 模型选择

在 LDA 模型中，需要经过多次实验来调整参数，以取得较好的效果。其中，主题数K需要人为设定，参数的好坏影响了建模的效果。在本文中，我们使用混乱度（Perplexity）[5]作为衡量参数选择的标准。

在自然语言处理领域中，若语料库D中有M篇文档，第i篇文档有N_i个词，则混乱度定义为

$$Perplexity(D) = exp\left\{-\frac{\sum_{i=1}^{M} log\, p(\boldsymbol{w}_i)}{\sum_{i=1}^{M} N_i}\right\}$$

其中$p(\boldsymbol{w}_i)$即为 LDA 模型生成第i篇文档的概率。

我们在K取不同值的情况下比较了混乱度的变化。在实验中，经过参数调整，$K = 20$时效果较好。在以下的实验中，我们统一设参数值$K = 20$。

此外，在 LDA 模型训练和推理过程中，我们设置迭代次数为 1000 次，来得到最终的模型。在实现上，我们利用开源代码库 JGibbLDA[32][33]作为 Gibbs 采样的基本实现，在此基础上扩展为本文提出的版本。

2. 总体主题分布

首先利用全部的轨迹数据集提取隐含的主题。在全部 20 个主题中，我们在表 5-1 中列出了 4 个主题进行详细分析。其中，受到篇幅限制，对于每个主题，我们只列出主题-词分布概率大小排在前 5 个词，如表 5-1 所示：

我们对其中的隐含主题分析如下：

- Topic 0 显示出的模式是“夜晚车辆的活动减少”，时间大致从 19 时开始，打车行为呈现减少趋势，这反映出晚上人们下班之后回到家，所以打车行为明显减少。
- Topic 1 显示出的模式是“深夜车辆活动少”，时间一般从凌晨开始，上班早高峰之前。在这个时间段，大部分人在睡眠中，但是仍然有少量的打车行为。此外，我们注意到，从 6 时开始热度开始增强。
- Topic 12 显示的模式是明显的晚高峰，这种模式的热度普遍非常高，显示出典型的晚高峰活动频繁的特征。值得说明的是，Topic 0 的 19 时和 Topic 12 的 19 时的热度差异较大，然而，这并不矛盾。这反映出这两个主题会在不同的区域出现，前者较为冷清，后者一般为闹市区。
- Topic 19 显示的模式最明显，是闹市区的模式。其中，时间覆盖了上午到下午，交通热度级达到顶值。这与城市中人们活动行为在地域上的分布明显不均匀有关。

值得说明的是，上表列出的主题并非发现的所有主题。在这些主题中，有的反映的是有关特定时间段的主题，有的反映的是有关特定区域的主题。我们将对这个问题进行进一步分析。

表 5－1 总体主题-词分布

Table 5-1 General Topic-word Distributions

主题	词		主题-词分布概率
	小时	交通热度级	
Topic 0	21	VL	0.098
	22	VL	0.092
	20	VL	0.087
	19	VL	0.076
	23	XL	0.063
Topic 1	5	VL	0.119
	4	VL	0.096
	1	VL	0.093
	6	L	0.078
	2	VL	0.078
Topic 12	19	VH	0.059
	20	VH	0.053
	18	VH	0.051
	16	VH	0.049
	17	VH	0.047
Topic 19	14	XH	0.062
	13	XH	0.061
	11	XH	0.058
	12	XH	0.0575
	10	XH	0.053

3. 不同区域的主题分布

在本节中，我们考虑在不同区域中的主题-词分布，这些分布体现出不同区域交通轨迹和人们生活模式的差别。因为在本实验中，一共有 2500 个区域，我们现在给出案例研究。 对于每一块区域，我们首先计算该区域内概率排在前 3 位的主题，然后分析这些主题表现出的模式含义。

图 5－2 案例分析 I 区域

Figure 5-2 Region of Case Study I

（1）案例分析 I：前门商圈

我们首先考察前门商圈，北京市著名的商业区域，具体位置见图 5-2。对于该区域，我们选出概率前三大的主题，分别列于表 5-2。

通过分析概率分布，我们可以看出，前门商圈位于闹市区的核心位置，其模式特征非常明显。从 Topic 12 和 Topic 19 可以明显地看出中午和晚上两个高峰，在这两个高峰，出粗车活动非常活跃，反映出人们在这一区域活动的密集程度。值得一提的是，从 Topic 16 可以看出，即使到了半夜，人们的热情依然不减，仍然有很多人坐出租车，尤其是 23 时至 0 时的时间范围。

此外，尽管 Topic 12 和 Topic 19 的模式类似，但是其语义完全不同，一为午间的高峰，一为晚高峰，所以被分为两个不同的主题，这体现出本文方法的优越性。

表 5－2 案例分析 I

Table 5-2 Case Study I

主题	文档-主题概率	词		主题-词分布概率
		小时	交通热度级	
Topic 12	0.324	19	VH	0.059
		20	VH	0.053
		18	VH	0.051
		16	VH	0.049
		17	VH	0.047
Topic 16	0.244	3	L	0.061
		2	L	0.055
		4	L	0.052
		23	VH	0.051
		0	VH	0.488
Topic 19	0.157	14	XH	0.062
		13	XH	0.061
		11	XH	0.058
		12	XH	0.0575
		10	XH	0.053

（2）案例分析 II：海淀区大学

案例 II 位于海淀区，这个区域周围大学林立。图 5-3 显示的区域在北京科技大学和北京航空航天大学附近。对于该区域，我们同样选出概率前三大的主题，分别列于表 5-3。

图 5－3 案例分析 II 区域

Figure 5-3 Region of Case Study II

从表 5-3 中，我们重点分析案例 I 和案例 II 区域交通模式的异同点。我们注意到两者都拥有 Topic 12，这说明两区域在晚上的打车行为都有一个明显的高峰。然而，注意到 Topic 14，案例 II 有一个明显的早高峰，体现出在教育区域早上较为繁忙，而商业区上午相对顾客较少。此外，Topic 17 显示出大学在晚上的活动没有商业区活跃，这与 Topic 16 构成鲜明对比。

表 5－3 案例分析 II

Table 5-3 Case Study II

主题	文档-主题概率	词		主题-词分布概率
		小时	交通热度级	
Topic 14	0.309	9	VH	0.089
		10	VH	0.081
		11	VH	0.074
		8	VH	0.073
		13	VH	0.072
Topic 12	0.206	19	VH	0.059
		20	VH	0.053
		18	VH	0.051
		16	VH	0.049
		17	VH	0.047
Topic 17	0.190	20	L	0.078
		21	L	0.078
		23	VL	0.076
		19	L	0.076
		22	L	0.065

（3）案例分析 III：居民区

案例 III 位于丰台区，是三环至四环之间的居民区，如图 5-4。该区域前三大的隐含主题见表 5-4。

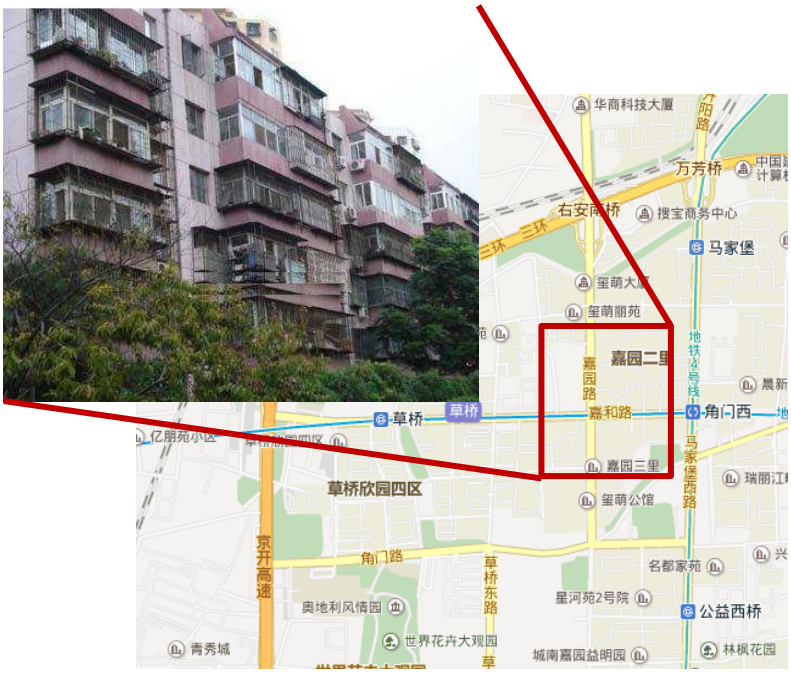

图 5－4 案例分析 III 区域

Figure 5-4 Region of Case Study III

我们重点比较案例 III 与案例 I 和案例 II 之间的区别。从表 5-4 中可以看出，前三个主题：Topic 15、Topic 0 和 Topic 2 有一个相同的特征：交通热度都非常低。这可能由两个理由造成：i）位置比较偏远，位于郊区，和 ii）主要是当地居民，人口流动不剧烈。从这三个主题中，我们仍然能看出语义的不同，这三个主题分别反映出 Topic 15 和 Topic 2 反映出下午的特征，Topic 0 反映出晚上的特征。

此外，我们可以进一步比较 Topic 15 和 Topic 2 的不同。Topic 2 虽然在时间段上与 Topic 15 十分相似，但是 Topic 2 的热度比 Topic 15 稍微高一点，而概率明显偏低。这反映出该地区下午在大部分时间交通很不活跃，而在有的日期比往常稍活跃。

综上所述，从这三个案例中可以看出，本文提出的方法可以有效地挖掘出交通轨迹的隐含主题，这些主题体现的模式实际上反映出人们的生活模式。我们从案例对比中进一步看出，不同区域模式的不同甚至体现出城市功能的不同。通过进一步分析，我们可以找出更多的区域交通模式，比如工业区的模式、远郊区的模式。因为方法相同，本文不再赘述。

表 5－4 案例分析 III

Table 5-4 Case Study III

主题	文档-主题概率	词		主题-词分布概率
		小时	交通热度级	
Topic 15	0.502	14	XL	0.064
		13	XL	0.062
		11	XL	0.061
		15	XL	0.061
		16	XL	0.060
Topic 0	0.322	21	VL	0.098
		22	VL	0.092
		20	VL	0.087
		19	VL	0.076
		23	XL	0.063
Topic 2	0.082	14	VL	0.057
		10	VL	0.056
		15	VL	0.056
		13	VL	0.055
		11	VL	0.054

4. 不同时间范围的主题分布

本实验中，我们分别考虑在工作日（周一至周五）和周末的主题-词分布。我们重点考察位于相同或相似时间段的主题，从中可以发现工作日（周一至周五）和周末交通模式的区别。

在实验中，我们把原轨迹数据集按时间戳分成两个部分：工作日和周末。在这两个子数据集上，我们进行与前述相同的实验，为了使实验更有说服力，我们使用和之前相同的参数配置，学习 LDA 模型。

由于不同数据集上学习到的主题不相同，没有办法进行直接比较。我们考虑相似的主题进行比较，详细结果见表 5-5。

表 5－5 主题比较

Table 5-5 Comparison of Topics

对比组	时间范围	主题	词		主题-词分布概率
			小时	交通热度级	
1	工作日	Topic 2	9	VH	0.086
			10	VH	0.076
			8	VH	0.075
			11	VH	0.072
			12	VH	0.068
	周末	Topic 12	12	VH	0.054
			13	VH	0.054
			11	VH	0.050
			14	VH	0.049
			10	VH	0.049
2	工作日	Topic 18	20	H	0.054
			19	H	0.052
			21	H	0.049
			18	H	0.047
			17	H	0.045
	周末	Topic 18	23	VH	0.033
			16	VH	0.032
			17	VH	0.030
			22	VH	0.029
			0	VH	0.027

在表 5-5 中，我们进行了两组对比。在对比组 1 中，我们选择了两个时间范围中热度级最大的组进行对比。这两个主题在交通热度级上相似，但是时间的分布有区别。在工作日，时间从 8 时开始，然而在周末，时间从 10 时开始。这反映出工作日人们

外出工作较早，而在周末由于不需要工作，时间相对较晚。对比组 1 整体体现了工作日与周末在上午的不同模式。

在对比组 2 中，我们考虑的时间是晚上。工作日的 Topic 18 和周末的 Topic 18 在时间段和交通热度级上都比较相似，它们都体现晚上的活动。然而，工作日比起周末热度更低，而且结束的时间更早。

综上所述，本文提出的方法能分析不同时间范围内整体的区域交通模式。通过比较其异同点，可以发现交通轨迹背后反映出人们的生活模式区别。

六、 总结和展望

本文提出了一个基于隐性狄利克雷分配的区域交通模式挖掘算法。本文从 GPS 轨迹数据的建模出发，提出了交通热度级等概念，在此基础上将不同区域的轨迹数据及表示为“文档”和“词”，并给出了相应的生成算法在 MapReduce 框架上的实现方法。本文利用 Gibbs 采样算法，迭代地对 LDA 模型进行学习和推理。通过求解 LDA 模型，本文提出挖掘区域交通轨迹下的隐含主题，这些主题显示了交通轨迹的内在语义模式，反映出人们的生活模式。为了证明本文方法的正确性和有效性，在真实的大数据集下进行实验，分别从挖掘隐含主题、不同区域的主题分析、不同时间范围的主题分析等三个方面进行说明。

尽管本文提出的方法较好地解决了区域交通模式挖掘的问题，然而，还有一些工作有待未来完成。原有的基于经纬度的区域划分方法可以改进了基于路网的，更加自然地反映出城市的实际情况[34]。其他概率主题模型，例如作者主题模型也可能对本文提出的算法有改进作用。此外，可以利用 POI 等信息来帮助自动分析隐含主题的语义信息[35][36]，以更好地从数据角度理解我们的城市和人们的生活。

参考文献

[1] Yuan J, Zheng Y, Zhang L, et al. Where to find my next passenger[C]//Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011: 109-118.

[2] Luo W, Tan H, Chen L, et al. Finding time period-based most frequent path in big trajectory data[C]//Proceedings of the 2013 international conference on Management of data. ACM, 2013: 713-724.

[3] Pan B, Zheng Y, Wilkie D, et al. Crowd sensing of traffic anomalies based on human mobility and social media[C]//Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2013: 334-343.

[4] Zheng K, Zheng Y, Yuan N J, et al. On discovery of gathering patterns from trajectories[C]//Data Engineering (ICDE), 2013 IEEE 29th International Conference on. IEEE, 2013: 242-253.

[5] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3: 993-1022.

[6] Porteous I, Newman D, Ihler A, et al. Fast collapsed gibbs sampling for latent dirichlet allocation[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 569-577.

[7] Gidofalvi G, Pedersen T B. Cab-sharing: An effective, door-to-door, on-demand transportation service[C]//Proceedings of the 6th European Congress on Intelligent Transport Systems and Services, 18-20 Jun, 2007, Aalborg, Denmark. ERTICO, 2007: 8.

[8] Sumpter N, Bulpitt A. Learning spatio-temporal patterns for predicting object behaviour[J]. Image and Vision Computing, 2000, 18(9): 697-704.

[9] Li Z, Han J, Ji M, et al. Movemine: Mining moving object data for discovery of animal movement patterns[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(4): 37.

[10] Life and Motion of Socio-Economic Units: GISDATA[M]. CRC Press, 2003.

[11] Iwase S, Saito H. Tracking soccer players based on homography among multiple views[C]//Visual Communications and Image Processing 2003. International Society for Optics and Photonics, 2003: 283-292.

[12] Tang L A, Zheng Y, Yuan J, et al. On discovery of traveling companions from streaming trajectories[C]//Data Engineering (ICDE), 2012 IEEE 28th International Conference on. IEEE, 2012: 186-197.

[13] Pang L X, Chawla S, Liu W, et al. On mining anomalous patterns in road traffic streams[M]//Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2011: 237-251.

[14] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. Indexing by latent semantic analysis[J]. JASIS, 1990, 41(6): 391-407.

[15] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 50-57.

[16] Klema V, Laub A J. The singular value decomposition: Its computation and some applications[J].

Automatic Control, IEEE Transactions on, 1980, 25(2): 164-176.

[17] Zheng Y, Li Q, Chen Y, et al. Understanding mobility based on GPS data[C]//Proceedings of the 10th international conference on Ubiquitous computing. ACM, 2008: 312-321.

[18] Yin J, Chai X, Yang Q. High-level goal recognition in a wireless LAN[C]//Proceedings of the national conference on artificial intelligence. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004: 578-584.

[19] Liao L, Fox D, Kautz H. Extracting places and activities from gps traces using hierarchical conditional random fields[J]. The International Journal of Robotics Research, 2007, 26(1): 119-134.

[20] Eagle N, Pentland A. Reality mining: sensing complex social systems[J]. Personal and ubiquitous computing, 2006, 10(4): 255-268.

[21] Eagle N, Pentland A S. Eigenbehaviors: Identifying structure in routine[J]. Behavioral Ecology and Sociobiology, 2009, 63(7): 1057-1066.

[22] Farrahi K, Gatica-Perez D. Discovering routines from large-scale human locations using probabilistic topic models[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(1): 3.

[23] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004: 487-494.

[24] Kesten H, Morse N. A property of the multinomial distribution[J]. The Annals of Mathematical Statistics, 1959: 120-127.

[25] Fabius J. Two characterizations of the Dirichlet distribution[J]. The Annals of Statistics, 1973: 583-587.

[26] Lee P M. Bayesian statistics: an introduction[M]. John Wiley & Sons, 2012.

[27] Bishop C M. Pattern recognition and machine learning[M]. New York: springer, 2006.

[28] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[29] Gilks W R. Markov chain monte carlo[M]. John Wiley & Sons, Ltd, 2005.

[30] Casella G, George E I. Explaining the Gibbs sampler[J]. The American Statistician, 1992, 46(3): 167-174.

[31] Wei X, Croft W B. LDA-based document models for ad-hoc retrieval[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 178-185.

[32] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]//Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 91-100.

[33] Bíró I, Szabó J, Benczúr A A. Latent dirichlet allocation in web spam filtering[C]//Proceedings of the 4th international workshop on Adversarial information retrieval on the web. ACM, 2008: 29-32.

[34] Yuan J, Zheng Y, Zhang C, et al. An interactive-voting based map matching algorithm[C]//Proceedings of the 2010 Eleventh International Conference on Mobile Data Management.

IEEE Computer Society, 2010: 43-52.

[35] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 186-194.

[36] Zheng Y, Liu Y, Yuan J, et al. Urban computing with taxicabs[C]//Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011: 89-98.

附录

(一)概率与统计

1. 多项式分布

多项式分布[24]（multinomial distribution）可以用来建模投掷一个有K面的骰子的结果。设K维向量$\boldsymbol{x} = (x_1, x_2, \cdots, x_K)$，其中$x_j$是骰子的第$j$面朝上的次数，则$\boldsymbol{x}$的概率质量函数（probability mass function）为

$$p(\boldsymbol{x}|n, \boldsymbol{\theta}) = \binom{n}{x_1 \cdots x_K} \prod_{i=1}^{K} {\theta_j}^{x_j}$$

其中，$\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_K)$是模型参数，$\theta_j$是骰子的第$j$面朝上的概率。$\binom{n}{x_1 \cdots x_K}$是多项式系数，定义为

$$\binom{n}{x_1 \cdots x_K} = \frac{n!}{x_1! \cdots x_K!}$$

n为投掷的总数，即

$$n = \sum_{i=1}^{K} x_i$$

2. 狄利克雷分布

狄利克雷分布（Dirichlet distribution）[25]是连续的多变量的概率分布，其参数为$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K)$。狄利克雷分布的概率密度函数（probability density function）为

$$f(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^{K} {\theta_i}^{\alpha_i - 1}$$

其中，$\Delta(\boldsymbol{\alpha})$为归一化值，表示为：

$$\Delta(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$$

$\Gamma(x)$是 Gamma 函数，即为

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

对于正整数，Gamma 函数有如下性质：

$$\Gamma(n) = (n-1)!$$

如果$\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$，对于$\boldsymbol{\theta}$中的每一项$\theta_i$，其期望为

$$E(\theta_i) = \int_0^1 \theta_i \cdot Dirichlet(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\theta_i = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\Gamma(\alpha_i)} \cdot \frac{\Gamma(\alpha_i + 1)}{\Gamma(\sum_{i=1}^{K} \alpha_i + 1)} = \frac{\alpha_i}{\sum_{i=1}^{K} \alpha_i}$$

所以，对于$\boldsymbol{\theta}$，期望为：

$$E(\boldsymbol{\theta}) = \left(\frac{\alpha_1}{\sum_{i=1}^{k}\alpha_i}, \frac{\alpha_2}{\sum_{i=1}^{k}\alpha_i}, \cdots, \frac{\alpha_k}{\sum_{i=1}^{k}\alpha_i}\right)$$

3. 狄利克雷-多项式共轭

在贝叶斯统计学中，共轭先验（conjugate prior）是一个重要的概念。对于数据$\boldsymbol{x}$和参数$\boldsymbol{\theta}$，如果后验分布（posterior distribution）$p(\boldsymbol{\theta}|\boldsymbol{x})$和先验分布（prior distribution）$p(\boldsymbol{\theta})$为同一概率分布族，先验分布是似然函数（likelihood function）$p(\boldsymbol{x}|\boldsymbol{\theta})$的共轭先验[26]。根据贝叶斯定理，后验分布、先验分布和似然函数的关系如下所示：

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')\,d\boldsymbol{\theta}'}$$

如果数据$\boldsymbol{x}$满足多项式分布，参数为$\boldsymbol{\theta}$，$\boldsymbol{\theta}$满足狄利克雷分布，参数为$\boldsymbol{\alpha}$，即

$$\boldsymbol{x} \sim Multinomial(\boldsymbol{\theta})$$

$$\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$$

我们计算后验概率：

$$p(\theta_i|\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)} \cdot \prod_{i=1}^{K}{\theta_i}^{\alpha_i - 1}$$

$$p(x_i|n, \boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{K}x_i!} \cdot {\theta_i}^{x_i}$$

$$p(\theta_i|\boldsymbol{x}) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i + x_i)}{\Gamma(n + \sum_{i=1}^{K}\alpha_i)} \cdot \prod_{i=1}^{K}{\theta_i}^{\alpha_i + x_i - 1}$$

我们容易发现，如果

$$p(\boldsymbol{\theta}) = Dirichlet(\boldsymbol{\alpha})$$

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = Multinomial(\boldsymbol{\theta})$$

那么

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = Dirichlet(\boldsymbol{\alpha} + \boldsymbol{x})$$

因为先验概率和后验概率的分布都为狄利克雷分布，所以狄利克雷分布是多项式分布的共轭先验[26]。

4. 参数估计

我们可以利用共轭先验的知识简单地从数据集中估计参数的值。

设数据集$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$是从参数为$\boldsymbol{\theta}$的分布中独立采样得到，则$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$独立同分布。在贝叶斯统计中，我们认为参数$\boldsymbol{\theta}$不是一个常数，而是服从一个分布，其参数为$\boldsymbol{\alpha}$[26]。我们考虑后验分布：

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta})\prod_{i=1}^{n} p(\boldsymbol{x_i}|\boldsymbol{\theta})$$

在狄利克雷-多项式共轭中，若$p(\boldsymbol{x_i}|\boldsymbol{\theta})$服从参数为$\boldsymbol{\theta}$的多项式分布，$\boldsymbol{\theta}$服从参数为$\boldsymbol{\alpha}$的狄利克雷分布，所以$p(\boldsymbol{\theta}|D)$服从参数为$\boldsymbol{\alpha}+\boldsymbol{x}$的狄利克雷分布。由于狄利克雷分布的期望可推之：

$$E(\boldsymbol{\alpha}+\boldsymbol{x}) = \left(\frac{\alpha_1+x_1}{\sum_{i=1}^{k}\alpha_i+x_i}, \frac{\alpha_2+x_2}{\sum_{i=1}^{k}\alpha_i+x_i}, \cdots, \frac{\alpha_k+x_k}{\sum_{i=1}^{k}\alpha_i+x_i}\right)$$

所以，

$$\widehat{\boldsymbol{\theta}} = \left(\frac{\alpha_1+x_1}{\sum_{i=1}^{k}\alpha_i+x_i}, \frac{\alpha_2+x_2}{\sum_{i=1}^{k}\alpha_i+x_i}, \cdots, \frac{\alpha_k+x_k}{\sum_{i=1}^{k}\alpha_i+x_i}\right)$$

上述结论可用于本文模型的求解。

(二)Gibbs 采样与 LDA

1. Gibbs 采样推导

根据马尔可夫链的收敛定理，不可约的、非周期的马尔可夫链会收敛到平稳分布[29]。我们设在时间i的概率分布为$\pi_i(x)$，平稳分布为$\pi(x)$，若该马尔可夫链到第n步收敛，则X_n、X_{n+1}、X_{n+2}等都是同分布的(即采样自$\mathrm{p}(\mathbf{x})$)的样本。这种方法的关键在于构建状态转移矩阵。

根据马尔可夫链具有的细致平稳条件（detailed balance condition），如果对于任意i和j，如果有

$$\pi(i)P_{ij} = \pi(j)P_{ji}$$

则$\pi(x)$为马尔可夫链的平稳分布。P_{ij}为从状态i转移到状态j的概率。

对于二维随机变量的联合分布率$P(\mathrm{X},\mathrm{Y})$，考虑

$$p(x_1,y_1)p(y_2|x_1) = p(x_1)p(y_1|x_1)p(y_2|x_1)$$
$$p(x_1,y_2)p(y_1|x_1) = p(x_1)p(y_2|x_1)p(y_1|x_1)$$

我们知道

$$p(x_1,y_1)p(y_2|x_1) = p(x_1,y_2)p(y_1|x_1)$$

所以，对于细致平稳条件

$$p(X)Q(X\to Y) = p(Y)Q(Y\to X)$$

我们可以把上述条件概率当成状态转移概率，即X状态为(x_1,y_1)，Y状态为(x_1,y_2)，$p(y_2|x_1)$为从X状态转移到Y状态的概率，$p(y_1|x_1)$为Y状态转移到X状态的概率。

根据上述分析，我们可以得出二维随机变量的联合分布P(X, Y)的 Gibbs 采样算法，如下所示：

算法 7-1：二维随机变量分布的 Gibbs 采样算法

输入：条件分布率$p(Y|X)$，开始输出样本的时间n

输出：$p(X,Y)$的样本

步骤：

1. 随机初始化X和Y，即$X_0 = x_0$，$Y_0 = y_0$
2. 对于时间$\mathrm{t} = 1,2,3,\cdots$，循环对X和Y进行采样
   2.1. 采样$y_{t+1}{\sim}p(y|x_t)$
   2.2. 采样$x_{t+1}{\sim}p(x|y_{t+1})$
   2.3. 当$t \geq n$，输出(x_{t+1}, y_{t+1})

我们可以把二维的情况推广至n维。对于n维随机变量的联合分布率$P(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n)$，其 Gibbs 采样算法如下所示：

算法 7-2：n维随机变量分布的 Gibbs 采样算法

输入：条件分布率$p(X_i|X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$，开始输出样本的时间$m$

输出：$p(X_1, X_2, \cdots, X_n)$的样本

步骤：

1. 随机初始化$\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$，即$X_1 = x_1$，$\mathrm{X}_2 = x_2$，……，$\mathrm{X}_n = x_n$
2. 对于时间$\mathrm{t} = 1,2,3,\cdots$，循环对$\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$进行采样[9]
   2.1. 采样$x_1{}^{(t+1)}{\sim}p\left(x_1\middle|x_2{}^{(t)}, x_3{}^{(t)}, \cdots, x_n{}^{(t)}\right)$
   2.2. 采样$x_2{}^{(t+1)}{\sim}p\left(x_2\middle|x_1{}^{(t+1)}, x_3{}^{(t)}, \cdots, x_n{}^{(t)}\right)$
   2.3. 采样$x_3{}^{(t+1)}{\sim}p\left(x_3\middle|x_1{}^{(t+1)}, x_2{}^{(t+1)}, \cdots, x_n{}^{(t)}\right)$
   …
   2.i. 采样$x_i{}^{(t+1)}{\sim}p\left(x_i\middle|x_1{}^{(t+1)}, x_2{}^{(t+1)}, \cdots, x_{i-1}{}^{(t+1)}, x_{i+1}{}^{(t)}, \cdots, x_n{}^{(t)}\right)$

[9] 以下算法变量的上标含义为时间，不是指数。

…

2.n. 采样$x_n^{(t+1)}\sim p\left(x_n\middle|x_1^{(t+1)},x_2^{(t+1)},\cdots,x_{n-1}^{(t+1)}\right)$

2.(n+1). 当$t\geq m$，输出$(x_1,x_2,\cdots,x_n)$

详细的分析可以参考 Giibs 采样的参考文献[30]。

2. LDA 的 Gibbs 采样推导

利用n维随机变量的 Gibbs 采样法，在$\boldsymbol{w}$已知的情况下，依次对z_i进行采样，所以我们需要计算$p(z_i=k|\boldsymbol{z}_{\neg i},\boldsymbol{w})$的分布。利用狄利克雷-多项式共轭的性质，我们知道$\boldsymbol{\theta_m}$和$\boldsymbol{\varphi_k}$的后验分布也是狄利克雷分布[26]：

$$p(\boldsymbol{\theta_m}|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})=Dirichlet\left(\boldsymbol{\theta_m}\middle|\boldsymbol{n}_{\boldsymbol{m},\neg i}+\boldsymbol{\alpha}\right)$$

$$p(\boldsymbol{\varphi_k}|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})=Dirichlet\left(\boldsymbol{\varphi_k}\middle|\boldsymbol{n}_{\boldsymbol{k},\neg i}+\boldsymbol{\beta}\right)$$

其中，$\boldsymbol{n}_{\boldsymbol{m},\neg i}$是除了$i$之外第$m$篇文档分别被指定到各个主题的词的计数，$\boldsymbol{n}_{\boldsymbol{k},\neg i}$是除了$i$之外第$k$个主题中各个词的计数。$p(z_i=k|\boldsymbol{z}_{\neg i},\boldsymbol{w})$的推导过程如下所示：

$$\begin{aligned}
p(z_i=k|\boldsymbol{z}_{\neg i},\boldsymbol{w}) &\propto p(z_i=k,w_i=t|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})\\
&=\iint p(z_i=k,w_i=t,\boldsymbol{\theta_m},\boldsymbol{\varphi_k}|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})\,d\boldsymbol{\theta_m}\,d\boldsymbol{\varphi_k}\\
&=\iint p(z_i=k|\boldsymbol{\theta_m},\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})p(w_i=t,\boldsymbol{\varphi_k}|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})\,d\boldsymbol{\theta_m}\,d\boldsymbol{\varphi_k}\\
&=\iint p(z_i=k|\boldsymbol{\theta_m})p(\boldsymbol{\theta_m}|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})p(w_i=t|\boldsymbol{\varphi_k})p(\boldsymbol{\varphi_k}|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})\,d\boldsymbol{\theta_m}\,d\boldsymbol{\varphi_k}\\
&=\int p(z_i=k|\boldsymbol{\theta_m})p(\boldsymbol{\theta_m}|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})d\boldsymbol{\theta_m}\\
&\cdot\int p(w_i=t|\boldsymbol{\varphi_k})p(\boldsymbol{\varphi_k}|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})\,d\boldsymbol{\varphi_k}\\
&=\int p(z_i=k|\boldsymbol{\theta_m})\,Dirichlet\left(\boldsymbol{\theta_m}\middle|\boldsymbol{n}_{\boldsymbol{m},\neg i}+\boldsymbol{\alpha}\right)d\boldsymbol{\theta_m}\\
&\cdot\int p(w_i=t|\boldsymbol{\varphi_k})Dirichlet\left(\boldsymbol{\varphi_k}\middle|\boldsymbol{n}_{\boldsymbol{k},\neg i}+\boldsymbol{\beta}\right)d\boldsymbol{\varphi_k}\\
&=\int \theta_{m,k}Dirichlet\left(\boldsymbol{\theta_m}\middle|\boldsymbol{n}_{\boldsymbol{m},\neg i}+\boldsymbol{\alpha}\right)d\boldsymbol{\theta_m}\\
&\cdot\int \varphi_{k,t}Dirichlet\left(\boldsymbol{\varphi_k}\middle|\boldsymbol{n}_{\boldsymbol{k},\neg i}+\boldsymbol{\beta}\right)d\boldsymbol{\varphi_k}=E\left(\theta_{m,k}\right)\cdot E\left(\varphi_{k,t}\right)=\hat{\theta}_{m,k}\cdot\hat{\varphi}_{k,t}
\end{aligned}$$

所以，我们可以得到

$$p(z_i=k|\boldsymbol{z}_{\neg i},\boldsymbol{w})\propto\hat{\theta}_{m,k}\cdot\hat{\varphi}_{k,t}$$

根据模型参数的贝叶斯估计量：

$$\hat{\theta}_{m,k} = \frac{n_{m,k} + \alpha_k}{\sum_{i=1}^{K}(n_{m,i} + \alpha_i)}$$

$$\hat{\varphi}_{k,t} = \frac{n_{k,t} + \beta_t}{\sum_{n=1}^{V}(n_{k,n} + \beta_n)}$$

我们可以得到：

$$p(z_i = k|\boldsymbol{z}_{\neg \boldsymbol{i}}, \boldsymbol{w}) \propto \frac{n_{m,k} + \alpha_k}{\sum_{i=1}^{K}(n_{m,i} + \alpha_i)} \cdot \frac{n_{k,t} + \beta_t}{\sum_{n=1}^{V}(n_{k,n} + \beta_n)}$$

上式即为 LDA 的 Gibbs 采样公式。

致谢

华东师范大学软件学院学习的四年时间，是我终生难忘的四年。在这四年里，我学到了很多东西，我相信这些东西能使我受益终身。值此论文完成之际，我的心中感慨良多，我能体会到辛勤劳动的喜悦，同时让我感到它与老师同学的帮助和支持是分不开的。我想借此机会向他们由衷地致以感谢。

首先感谢我的论文指导老师何晓丰研究员。他对我的谆谆教诲使我进入了数据挖掘和机器学习等领域。他渊博的知识、勤恳的作风和一丝不苟的态度是我学习的楷模。他给我了参与各种学习和讨论的机会，在项目中给我自由发挥的空间，使我收获了知识，提高了能力。

感谢周傲英教授。周老师开阔的学术视野、敏锐的思维和严谨的学风深深地使我感动。他对我的鼓励、信任和支持是我不断努力的动力。

感谢数据科学与工程研究院和软件学院的其他老师给我的帮助和支持。钱卫宁教授在项目中给了我很多了指导和建议，宫学庆教授在 LBS 方向给了我很多启示和帮助。此外，王晓玲教授、王长波教授、金澈清教授、张蓉副教授、周敏奇副教授、高明副教授、张召副教授等也给了我很多关心和支持。感谢数据科学与工程研究院为我提供优越的学习和科研氛围，能让我在良好的环境下尽自己所能，也让我懂得“以云水趣看成败，以木石心图将来”。

除了感谢帮助过我的各位老师外，还必须感谢数据科学与工程研究院和软件学院的各位学长学姐，还有日日夜夜奋战在实验室同学们、战友们和朋友们。他们在平时学习生活中给了我帮助和照顾。我们一起讨论问题，一起看论文，一起完成项目，这使我收获良多。他们是程文亮、丁铖、段小艺、樊艳、郭心语、胡颢继、康强强、纪文迪、李金洋、李叶、梁翊涛、刘辉平、刘骁、马海欣、马建松、宋乐怡、苏永浩、孙凯、王燕华、夏帆、肖冰、闫季鸿、于程程、张弛、章群燕等。他们在我的大学生涯中在不同方面都给予我极大的帮助。在毕业的前夕，我的感激之情难以用文字描述，故仅以姓名拼音为序，在此一一列举。

最后，感谢华东师范大学软件学院对我的培养，我想对其他所有给我上过课、帮助过我的老师们和辅导员张炜帆老师和李恒超老师，以及支持、鼓励我的同学致以谢意。感谢养育我的父母和家人，使我能在各方面不断成长。

谢谢！