文章编号：2095-2163(2021)09-0200-06 中图分类号：TP31 文献标志码：A

分布式潜在狄利克雷分配研究综述

过云燕，李建中
（哈尔滨工业大学 海量数据计算研究中心，哈尔滨 150001）

摘 要：作为主题模型中最重要的机器学习模型，潜在狄利克雷分配问题在包括自然语言处理和信息检索等各领域展现出不可替代的地位。求解潜在狄利克雷分配问题主要采用变分推断和马尔科夫链蒙特卡洛两类算法。目前，数据的增长速度早已远超硬件能力的增长速度，因此在大数据时代，分布式平台的使用成为大数据训练的主流解决方案。利用分布式系统加速对潜在狄利克雷的训练和推断，成为相关研究领域的热门问题。本文对分布式潜在狄利克雷分配算法的相关工作进行分类整理和评估，对未来该领域的研究方向具有引导作用。

关键词：分布式系统；潜在狄利克雷分配；变分推断；马尔科夫链蒙特卡洛

A survey of distributed latent Dirichlet allocation

GUO Yunyan, LI Jianzhong
(Massive Data Computing Research Center, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 As the most important machine learning model in topic modeling, latent Dirichlet allocation (LDA) shows its irreplaceable position in many fields, such as natural language processing and information retrieval. Variational inference and Monte Carlo Markov chain are two main methods to solve LDA problems. At present, the growth rate of data has far exceeded the growth rate of hardware capabilities. Therefore, in the era of big data, the use of distributed platforms has become the mainstream solution for big data training. Using distributed systems to accelerate the training and inference process of latent Dirichlet allocation has become a hot issue in related research fields. This survey classifies and evaluates the related work of distributed LDA algorithms, and it will inspire the future research direction in this field.

【Key words】 distributed systems; latent Dirichlet allocation; variational inference; Monte Carlo Markov chain

0 引 言

主题模型(Topic Modeling) 是一类重要的机器学习(Machine Learning) 算法，可以从已有的文档集合中发现潜在的混合主题，这是其成为用来推断知识的无监督学习工具[1]。潜在狄利克雷分配(Latent Dirichlet Allocation)作为最重要的主题模型的问题设定，在过去十年的实践中已经显示其不可或缺的作用[2]。其假设每个潜在主题是由字典中多个单词混合组成的，而每个文档是由多个主题混合组成的，基于这两个假设，可以通过生成模型生成文档集。

潜在狄利克雷分配已被广泛应用于机器学习的多个不同领域，包括信息检索、文本分析、数据可视化、在线广告、推荐系统和网络分析[3]。这些领域不断收集待处理的数据，引发了大数据时代的到来，目前数据的增长速度早已远超硬件能力的增长速度，因此分布式平台的使用成为大数据训练的主流解决方案。现有相关综述汇总的潜在狄利克雷分配的算法研究，停留在解决应用数学问题的阶段。分布式潜在狄利克雷算法的设计包含系统工程问题，如设计数据和模型的划分与聚合、算法复杂性、通信计算平衡性等许多方面，需要对相关研究依据不同方法进行分类综述。

潜在狄利克雷分配可以通过两大类算法来求解，分别是马尔科夫链蒙特卡洛(MCMC)和变分推断(VI)。MCMC 算法对一个马尔可夫链进行采样，并用链上的样本来逼近后验概率的分布，该样本是渐近精确的。VI 算法试图找到一个分布族中能够最小化估计后验概率和精确后验概率之间的 KL 距离的分布，将原来的推理问题转化为优化问题。使用 MCMC 或者 VI 在大数据上进行训练时，每轮对模型的更新都需要对整个数据集进行完整的读写和推断，虽然这种全批策略适合于较小的数据集，但是

基金项目：国家自然科学基金(61832003)。

作者简介：过云燕(1991-)，女，博士研究生，主要研究方向：机器学习；李建中(1950-)，男，教授，博士生导师，主要研究方向：大数据计算理论、海量数据管理。

收稿日期：2021-03-26

由于每轮迭代太耗时，性能随着数据集大小的增长而显著降低。因此，如何利用分布式系统实现快速训练和推断，成为当前重要研究方向。

尽管利用随机变分推断算法解决小型和静态数据集上的问题已被广泛和深入的研究。但在实际情况下，数据集通常非常庞大，并且是以流的形式收集的。在现实世界中，在海量流数据上运行机器学习算法，面临 3 个挑战：模型演化、数据动荡和实时推断。一系列相关研究阐述了如何分别应对这些新的挑战。

本文旨在将当前针对分布式系统中高效运行潜在狄利克雷分配算法这一研究方向的现有成果进行归纳整理。

1　潜在狄利克雷分配问题

潜在狄利克雷分配（LDA）是针对文档的概率生成模型。给定输入文档集合 D 和主题数量 K，LDA 旨在将每个文档 d 表示为主题的混合分布，并将每个主题建模为字典 V 上的混合分布。此外，LDA 还推断每个文档中每个单词的主题分布。

LDA 假设了主题和文档的生成过程。β 代表 K 个主题中词的混合比例，是一个大小为 $K \times V$ 的矩阵。其中，第 k 行对应表示了主题 K 上词的分布，$\beta_k \sim Dirichlet(\eta)$，表示每一个 β_k 是从拥有对称参数 η 的狄利克雷分布中抽取到的。θ 是文档的主题混合比例，是一个大小为 $D \times K$ 的矩阵。第 d 行的 θ_d 对应表示文档 d 的主题混合比例，$\theta_d \sim Dirichlet(\alpha)$，表示每一个 θ_d 是从拥有对称参数 α 的狄利克雷分布中抽取到。在生成文档 d 中的第 n 个单词时，首先根据该文档的主题分布 θ_d，抽取一个主题作为该词对应的主题，主题的编号 $z_{dn} \sim Multinomial(\theta_d)$。再根据该主题的词的分布 $\beta_{z_{dn}}$，抽取了这个词 w_{dn}，$w_{dn} \sim Multinomial(\beta_{z_{dn}})$。整个文档可以通过重复多次后生成，而整个语料库可以在重复多次抽取文档后生成，变量之间的关系在图 1 中用水平箭头表示。

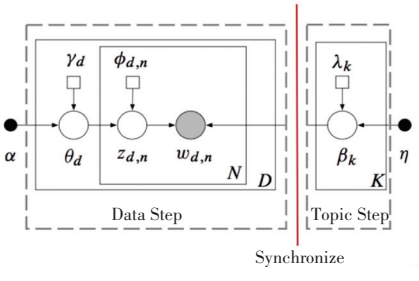

图 1　潜在狄利克雷分配的生成图模型

Fig. 1　Generative graphical model of latent Dirichlet allocation

2　变分推断及其分布式算法

2.1　变分推断

求解 LDA 问题的目标是对已给定的文档集合 w 检验后验条件概率分布 $p(\beta,\theta,z|w)$。由于后验分布的精确推断是难以解决的，Blei 等人建议使用变分推断（VI）来找到近似的后验条件概率分布[4]。

VI 引入了一个简单的可以被完全因子化的分布 $q(\beta,\theta,z|\lambda,\gamma,\varphi)$。引入新添加的变分参数：$\lambda$，$\gamma$，$\varphi$ 后，q 成为了精确后验条件概率分布的新的近似表示。在图 1 中，水平箭头之间的依赖关系复杂，而新填入的变分参数，带来的纵向箭头代表的依赖关系之间相互独立，替换了原来复杂的依赖关系。

$q(\beta_k \mid \lambda_k)$ 是主题 k 对应的变分概率分布，是主题级别的变分概率分布，$q(\beta_k \mid \lambda_k) \sim Dirichlet(\lambda_k)$。不需要直接推断真实的分布 $p(\beta_k \mid w,\eta)$，VI 选择了一个分布族 $q(\beta_k \mid \lambda_k)$，选择其中与之距离最近的一个，作为其近似分布。类似地，文档级别的变分参数 γ_d 和 φ_d，$q(\theta_d \mid \gamma_d) \sim Dirichlet(\gamma_d)$，并且 $q(z_{dn} \mid \varphi_{dn}) \sim Multinomial(\varphi_{dn})$。

VI 的目标是最小化实际分布与近似分布之间的 *KL* 距离，这等价于最大化 *ELBO*。*ELBO* 可以通过 3 个变分参数来表示自变量，优化 *ELBO* 的过程就是选择最佳的 λ,γ,φ，使之最大化的过程如下：

$$ELBO = \sum_d \{E_q[\log p(\beta,\theta_d,z_d,w_d \mid \alpha,\eta)] - E_q[\log q(\beta,\theta_d,z_d \mid \lambda,\gamma_d,\varphi_d)]\}$$

2.2　随机变分推断

尽管变分推断在模型质量和收敛速度两方面优于马尔科夫链蒙特卡洛，但在处理大规模数据集时仍然暴露出一些不足之处。使用变分推断进行训练时，每轮对模型的更新都需要对整个数据集进行完整的读取和推断，称作全批量处理（full-batch），这种全批量策略仅适合于较小的数据集。由于每轮迭代太耗时，算法性能随着数据集的增长而显著降低。

随机变分推断算法，在计算海量数据时具有很好的扩展性[5]。随机变分推断的主要思想是在每轮迭代时，不利用整体数据集，而是通过采集数据点作为样本，在样本上求得有噪声的梯度，用以对整个数据集的梯度进行无偏的估计。Online-LDA 使用随机自然梯度上升的优化方式，每轮迭代中，都在 D 中采样生成小样本集合，用来优化主题级别参数 λ，

而文档级别的参数(γ,ϕ) 依然采用坐标上升法进行优化。

主题级别参数和文档级别参数之间有着紧密的关系,互相依赖。λ 包含了来自数据集合中文档级别参数的信息,其质量直接决定了对文档推断结果的质量。当使用已经收敛到较优值的 λ 推断文档级别的参数时,才能得到更加合理的推断结果。

Online-LDA 与 VI-LDA 相比,其优势来源于更频繁地更新 λ。 频繁地利用具有正反馈含义的文档级别参数,会使 λ 收敛地更快,反之会收敛更慢或最终发散,无法收敛。基于随机优化的思想,Online-LDA 可以快速处理小批量(mini-batch)文档,以此来近似估计整体数据集的信息。虽然小批量文档包含一定的数据集信息,但是会存在噪声。所以小批量的大小有一个折中的选择,过小的小批量具有较少的语料信息,较大的噪声,导致模型收敛过慢甚至发散;而过大的小批量减缓了更新的频率,产生较弱的正反馈,收敛速率较低。在每一次迭代中采集小批量的文档作为样本,而不是单一文档作为样本时,Online-LDA 可以在单机环境中快速收敛,比基于变分推断的潜在狄利克雷分配算法收敛快数倍。

2.3 分布式变分推断类算法的工作

VI-LDA 是解决潜在狄利克雷分配问题的第一个变分推断算法[2]。在每轮迭代中,其都会为每个文档和文档中的每个词推断其参数。通过划分文档,可以对 VI-LDA 算法中文档级别计算过程实现并行计算[6],在共享内存的多处理器和多核系统中可以实现这一思想;在 Map-Reduce 框架中扩展 VI-LDA 的算法为 Mr.LDA[7]。在分布式计算框架结构中,Mr.LDA 设计 Mappers 划分文档级别的计算,Reducer 划分主题级别的计算。在共享内存的系统中,并行实现 VI-LDA 算法可以使用块矩阵划分策略[8],其避免了内存访问的冲突,减少了多线程之间的互锁时间,提供了基于剩余资源的动态调度。然而,随着数据集大小的不断增加,上述 3 个工作的内存需求也在增大。Online-LDA 用随机优化方法解决了这个问题;DoLDA 在 MapReduce 框架中提出了一种具有文档级任务划分的分布式 Online-LDA,利用了有限内存和并行计算的优势[9];Spark-MLlib 库选择使用 DoLDA 作为求解潜在狄利克雷分配问题的主要算法,并且对其算法的实现进行了系列优化,提供了当前性能最优的分布式 Online-LDA。Online-LDA 理论上同样可以在参数服务器上实现,但目前对其研究仍是空白。

3 马尔科夫链蒙特卡洛及其分布式算法

3.1 马尔科夫链蒙特卡洛

除了变分推断以外,潜在狄利克雷分配可以通过另一大类方法来求解,即马尔科夫链蒙特卡洛(MonteCarloMarkovChain)。马尔科夫链蒙特卡洛对一个马尔可夫链进行采样,并用链上的样本来逼近后验概率的分布,该样本是渐近精确的。而变分推断试图找到一个分布族中能够最小化估计后验概率和精确后验概率之间的 KL-距离的分布,这将原来的推理问题转化为优化问题。

马尔科夫链蒙特卡洛与变分推断相比有两个缺点:

(1)虽然马尔科夫链蒙特卡洛和变分推断在理论上可以渐近地达到相似优秀的损失函数,但在经验上马尔科夫链蒙特卡洛的模型质量并不令人满意[10]。因为潜在狄利克雷分配假设一个词属于多个主题,是一个混合模型,但是马尔科夫链蒙特卡洛技术通常将分配仅仅集中在单一主题上[11],因此其在海量数据集和复杂模型上表现不佳,变分推断已被证明更适合于混合模型[4]。

(2)最近的研究工作中多次表明,马尔科夫链蒙特卡洛收敛得比变分推断慢[12]。例如,在 PublicMedicine 数据集上,马尔科夫链蒙特卡洛的收敛需要对整体数据集进行成百上千轮训练,而变分推断只需要数十轮访问整体数据集来迭代优化模型。因此,马尔科夫链蒙特卡洛的整体性能比变分推断慢了一个数量级,特别是在较大的数据集上劣势更为明显[4]。

3.2 分布式马尔科夫链蒙特卡洛算法的工作

用吉布斯采样的方法,可以解决潜在狄利克雷分配问题,但采样器运行代价过高[13]。在此基础上,另一些工作提出各类采样器设计方法以简化计算;利用分布偏斜的特性,仅计算每个采样计算主题概率的一小部分[14];利用稀疏结构来实现较低的采样算法复杂性[15];减少了子采样操作的内部迭代所花费的时间[16];上述方案的混合策略,成为当前运行最快的采样器[12]。另一方面,在异步系统与参数服务器中对文档级别的计数矩阵进行划分,同样是比较流行的方案。YahooLDA 建立了一个基于原则和利用潜在狄利克雷分配固有稀疏性的系统[17];可以对文件进行分区,并允许每个采样器与分布式系统中的中央服务器不断的通信,这里每个采样器将

差分发送给中央服务器，并从其接收最新的全局变量值[18]；可以采用非对称结构来降低通信代价，并采用偏斜的划分策略来平衡不同服务器的负载[12]。另一些工作在系统中同时并行划分文档级计数矩阵和分区级计数矩阵，由于调度复杂和分区数量庞大，Peacock 的吞吐量通常低于仅仅划分数据的系统[19]；F+LDA 提出了一种游牧分配方案，并利用芬威克树进行抽样[8]；在 LightLDA 中进一步改进其分配方案的同时，使用别名表优化采样[20]；同时可以采用随机版本吉布斯采样算法求解潜在狄利克雷分配[21]；可以在 GPU 计算环境下实现吉布斯采样解决潜在狄利克雷分配问题[22]。

3.3　分布式混合类算法的工作

混合类算法的主要思想是在主题级别推理中使用变分推断的方法，在文档级别推理中使用马尔科夫链蒙特卡洛的方法。CVB[23]、CVB0[24] 和稀疏 SVB[25] 用吉布斯采样来推断文档级别变量的分布，但是在变分推断的设定中，条件概率的计算代价十分昂贵，需要数轮的采样。ESCA 存储文档级别变量分布的充分统计量(sufficient statistics)，使得内存占用较小[10]。随机 CVB、稀疏 SVB 和 ESCA 是混合类算法的随机版本，该类方法在面向主题级别的参数时，同样可以设计出 Online 版本的算法，与 Online-LDA 类似。同时，文档级别的划分方式也被用来加快算法的速度。最小化在 GPU 上的冲突，可以实现并行 CVB[26]。分布式 CVB 采用异步通信策略[27]，分布式 ESCA 采用同步通信策略[10]。

4　面向流数据处理的相关工作

4.1　数据贝叶斯

流变分贝叶斯(Streaming Variational Bayes)代表了学习流数据中潜在变量的最新方法[28]。流变分贝叶斯引入了贝叶斯更新，以避免在每轮迭代中大量访问历史数据：给定已经训练过的模型作为先验知识，以及新生成的、具有相同时间戳的文档作为观测数据，就可以使用流变分贝叶斯算法计算出近似的后验知识；同时，后验知识对应的矩阵成为当前最新模型。但是，用随机变分推断求解潜在狄利克雷分配问题时，需要假定主题分布和数据分布是相对稳定的，这意味着其没有考虑数据动荡的情况，以及如何保持动荡数据中的主题一致性。当新生成数据实例上的主题分布与拥有上一个时间戳的数据实例上的主题分布明显不同时，利用流变分贝叶斯所获得的先验知识更接近于被随机初始化。在这种情况下，每一轮的训练将不得不建立在非常模糊的草图主题和一批无法预知其主题分布的数据实例上。因此，流变分贝叶斯无法保持主题一致性，并且等待训练和推断的延迟也非常高。

4.2　通用流数据处理平台

为了在大量的流数据上进行实时推理，一些分布式流处理系统将工作负载分为两个阶段：训练阶段和推理阶段[29]。在这种框架中，训练阶段充当了预处理步骤，延迟仅包括推断时间，因此可以通过分布式计算来进一步加速推断。该框架的主要局限性在于推理结果完全取决于从历史数据中提取的主题，而忽略了数据流上存在主题演变的事实。为了获得高质量的实时推理，主题演变意味着需要在数据流上进行终身学习(持续学习)。因此，现有的分布式流处理系统中的潜在狄利克雷分配算法，不再是最优解。

4.3　动态主题模型

另一方面，包括动态主题模型(DynamicTM)在内的一些已有工作，研究了如何在带有时间戳的小型静态历史数据集上挖掘不断发展的主题的问题[30-31]。与经典的潜在狄利克雷分配相比，其通过修改潜在变量之间的因果关系，设计了衍生的主题模型公式。类比经典潜在狄利克雷分配算法的求解推导过程，针对动态主题模型推导类似的更新函数，用来训练动态主题模型。但是，这些算法无法轻量级的部署在真实的流数据场景中，因为处理整个数据集的训练过程是非常耗时的[31]，并且当数据增加时，该类算法对存储的需求非常昂贵[30]。

4.4　动态数据集

在动荡的数据流上，新数据实例上的主题分布与从整个流中采样获得的数据实例上的主题分布是不同的。在这种情况下，已有研究针对流数据上的潜在狄利克雷分配提出了种群变异贝叶斯(PP-LDA)[32]和信任区域(TR-LDA)[33]，以应对数据动荡的情况。PP-LDA 定义了流数据的总体后验，需要存储所有见到过的数据在所有特征上的分布，在实践中，当数据量和特征数量增加时，这个记录的大小会成比例增长；TR-LDA 使用信任区域优化思想更新模型，该算法要求从整个历史数据中无偏采集样本数据实例来迭代优化，因此采样代价会随着数据流的增长而增加。尽管 PP-LDA 和 TR-LDA 具有一定的理论保证，但其都需要昂贵的存储空间来存储大量历史记录，并且每轮采样都需要昂贵的读写代价。因此都不适合应用在大规模流数据集上。

5 分布式机器学习平台

5.1 参数服务器

通常可以将参数服务器视为一种架构,该架构使用分布式内存来维护机器学习中的模型,并支持保持节点间一致性的灵活的通信控制,提供了诸如 pull 和 push 之类的原语,使用户能够使用自定义的一致性控制器如:BSP、SSP 和 ASP 同步或异步更新模型的一部分。自提出参数服务器概念以来,其灵活和卓越的性能使其变得非常流行,例如:Li 等提出在参数服务器上执行小批量随机梯度下降算法(mini-batchSGD)[18],Petuum 是使用参数服务器架构的通用机器学习系统[34];也有使用参数服务器实现求解潜在狄利克雷分配算法的工作,例如:YahooLDA 在参数服务器之间划分数据对应的矩阵;LightLDA 使用参数服务器划分并存储部分主题矩阵。上述系统均属于基于马尔科夫链蒙特卡洛思想求解潜在狄利克雷分配的算法。

5.2 MapReduce

Spark 原生机器学习库 MLlib,基于变分推断实现 VI-LDA 算法[35]。为了在大数据集上实现快速收敛,MLlib 提供了基于随机变分推断实现的 OnlineLDA 算法,优于 VI-LDA。同时,Kaoudi 构建了一个基于代价的优化器,以针对给定的工作量选择最佳的计划[36];Anderson 将 MPI 集成到 Spark 中,并将工作负载转移到 MPI 环境中[37]。将数据从 Spark 传输到 MPI 环境,使用高性能 MPI 二进制文件进行计算,最后将结果复制回分布式文件系统中以供进一步使用。

5.3 计算和通信的平衡

为了平衡计算和通信,有以下几方面工作。首先,在给定分布式工作任务的情况下确定要使用的计算节点数量,使用过多的计算节点会增加通信代价,而使用过少的计算节点会增加每个节点的计算代价。按照这一思路,McSherry 认为,分布式计算的性能至少应优于单节点上最优实现的性能[38];Huang 使用尽可能少的机器来确保性能和效率[39]。其次,有许多工作建议通过尽可能多地执行本地计算来降低通信代价。例如,Grape 是当前最优的分布式图形处理系统,其试图在一台机器上进行尽可能多的计算,并减少分布式图形处理中的迭代次数[40];Gaia 是使用参数服务器的地理分布式机器学习系统,试图尽可能使用局域网间通信而非广域网间通信来降低通信代价[41];也有一些工作通过对工作负载进行划分,以实现更好的负载平衡,从而降低通信代价[42]。

6 结束语

本文首先介绍了狄利克雷分配问题,然后简要阐述了求解该问题的两个主流方法:变分推断和马尔科夫链蒙特卡洛。并且,围绕两个方法的经典单机算法,从数据划分、任务划分等思路,介绍将其转化为分布式算法的相关工作。综合基于变分推断、基于马尔科夫蒙特卡洛、基于两者混合算法的并行工作,可以得出如下结论:第一,针对马尔科夫链蒙特卡洛类算法的研究丰富,针对变分推断算法的研究仍为起步阶段,而从理论和实践两方面表明,变分推断算法优于马尔科夫链蒙特卡洛类算法,更值得被关注和进一步研究。第二,现有研究多集中于共享内存的并行算法设计,而针对不共享内存的分布式计算环境,如何提升基于随机思想的变分推断算法效率问题,具有重要的研究意义。

紧接着,面对流数据处理带来的新挑战,本文汇总了分别应对这些挑战的代表性算法。现有研究开始设计解决方案,以同时应对实时推断、主题演变、数据动荡 3 方面挑战[43]。

最后,从分布式系统角度,对系统框架设计方案、计算与通信平衡设计方案进行了更深入的讨论。将更多计算任务本地化,是减少通信代价的重要思想。参数服务器与 Spark 相比,通信模式具有优势。利用上述两点特性,提升 Spark 的性能,是值得研究的方向[44]。

综上所述,分布式潜在狄利克雷分配算法的训练效率和推断效果均有提高,但与此同时依然存在着更大的提升空间。

参考文献

[1] SMOLA A, NARAYANAMURTHY S. An architecture for parallel topic models[J]. PVLDB, 2010,3(1-2): 703-710.

[2] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. JMLR, 2003, 3(1): 993-1022.

[3] CHEN J, LI K, ZHU J, et al. Warplda: a cache efficient o (1) algorithm for latent dirichlet allocation [J]. PVLDB, 2016, 9 (10): 744-755.

[4] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Variational inference: A review for statisticians[J]. Journal of the American Statistical Association, 2017, 112(518): 859-877.

[5] HOFFMAN M, BACH F R, BLEI D M. Online learning for latent dirichlet allocation[C] // NeurIPS. 2010: 856-864.

[6] NALLAPATI R, COHEN W, LAFFERTY J. Parallelized variational EM for latent Dirichlet allocation: An experimental evaluation of