主题模型中的参数估计方法综述

杜 慧 陈云芳 张 伟
(南京邮电大学物联网学院 南京 210003)

摘 要 主题模型利用快速的机器学习算法从高维稀疏的单词数据中提取出低维的主题表示,实现了对文档单词的聚类。对主题模型中的参数进行估计是该领域的一项重要研究工作。详细描述了概率潜在语义分析模型和潜在狄利克雷模型以及主题模型中基本的参数估计方法,并对模型的困惑度进行了实验比较。

关键词 主题模型,概率潜在语义分析,隐含狄利克雷分布,参数估计
中图法分类号 TP39 **文献标识码** A

Survey for Methods of Parameter Estimation in Topic Models

DU Hui CHEN Yun-fang ZHANG Wei
(Department of the Internet of Things, Nangjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract Topic models extract low-dimensional representation of the topic from high-dimensional sparse data set of word by using fast machine learning algorithms, achieving a word document clustering. It is an important work in this field to study the model parameter estimation. The paper detailed the probabilitic latent semantic analysis model, the latent Dirichlet model and basic methods of parameter estimation in topic model. In addition, the paper gave an experimental analysis of perplexity in topic model.

Keywords Topic models, Probabilitic latent semantic analysis, Latent Dirichlet allocation, Parameter estimation

1 引言

主题模型是在潜在语义分析(Latent Semantic Analysis, LSA)[1]模型的基础上发展起来的,潜在语义分析模型的基本思想是采用统计方法来寻找文档中词与词之间存在的某种潜在的语义结构,可以看成是扩展的向量空间模型。潜在语义模型本质上是考虑词与词在文档中的共现,利用线性代数的方法提取出语义结构。潜在语义分析模型是基于线性代数的推理方法,不是概率模型,而主题模型是一类概率模型,因此严格来讲潜在语义分析模型算不上是主题模型。随着概率统计方法的广泛应用,潜在语义分析从线性代数的分析角度提升到了概率统计的分析角度,从而诞生了概率潜在语义分析(Probabilitic Latent Semantic Analysis, PLSA)[2]模型。主题模型在 2003 年才正式被提出,同时 Blei 等人在概率潜在语义分析的基础上提出 LDA(LatentDirichlet Allocation)[3]模型,用一个服从狄利克雷分布的 K 维潜在随机变量表示文档的主题概率分布,由先验参数随机生成,在一定程度上解决了 PLSA 因参数过多而造成的过拟合问题。

一个主题模型通常包括 5 项内容:主题模型的输入、基本假设、主题模型的表示、参数估计和新样本推断[4]。主题模型中最重要的两组参数是主题-词概率分布和文档-主题概率分布,通过参数估计可以得到具体参数值。PLSA 模型中,各个因素(文档、词和潜在语义空间)之间的概率分布及其求解是最重要的,最常用的近似推理方法是 EM(Expectation Maximization)算法[5]。而 LDA 是一个三层的贝叶斯模型,模型中的参数都被看作随机变量,并且引入了超级参数,使得模型对外只有两个参数。LDA 模型中常见的参数推理方法有吉布斯采样(Gibbs Sampling, GS)算法[6]、变分贝叶斯(Variational Bays, VB)推理算法[7]和消息传递(Belief Propagation, BP)算法[8]等。

本文针对近几年比较热点的主题模型参数估计方法进行归纳总结。首先对主题模型进行简单的概述,在了解生成过程的基础上学习模型的参数估计[10]方法,最后简单介绍了模型评估的方法并选择了一种比较常见的模型评估方法即困惑度来对 PLSA 模型和 LDA 模型进行实验分析。

2 主题模型概述

主题模型[9]是一种用于分析大规模文档的概率模型,其抽取隐藏在文档中的主题信息,然后按照主题信息对文档进行分类管理。主题模型是利用快速的机器学习算法从高维稀疏的单词数据中提取低维的数学表示,进而对文档单词进行聚类。常见的主题模型包括概率潜在语义分析模型、潜在狄利克雷分析模型及其扩展模型。

2.1 PLSA 模型

在 PLSA 模型中,主题表示为词的一个分布,通过对文档建立概率模型,使文档的似然最大,可以估计得到主题对应的

本文受国家自然科学基金资助项目(61272422)资助。
杜 慧(1991-),女,硕士生,主要研究方向为社会计算及其应用,E-mail:15077879576@163.com;陈云芳(1976-),男,副教授,主要研究方向为社会计算及其应用、信息网络安全、大数据处理和移动应用开发;张 伟(1973-),男,博士,教授,主要研究方向为计算机病毒技术、网络数据行为分析、Web 应用与数据安全等。

分布。图 1 中 d 表示文档，z 表示主题，w 表示单词，M 表示文档集的大小，N 表示文档的长度。假设输入文档集是由 k 个主题生成的，对于每一篇文档，生成过程如下：对于文档中每个词的位置，首先根据分布 $p(z|d)$ 选择这个词的潜在主题，其中 $z=1,2,\cdots,k$ 为随机变量，在为一个词的位置选择了相应的主题后，根据这个词的分布 $p(w|z)$ 生成这个词，从而在文档 d 的这个位置上出现词 w 的概率为：

$$p(w,d)=\sum_z p(w,d,z)=\sum_z p(w|z)*p(z|d)*p(d) \tag{1}$$

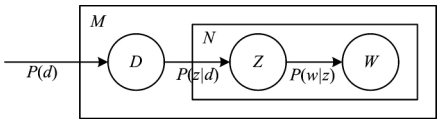


图 1 PLSA 模型

考虑到整个文档集 D 包括所有的文档及其所有的词，其似然为：

$$L(D)=\prod_{d\in D}\prod_{w\in d}\sum_z p(w,d,z)=\prod_{d\in D}\prod_{w\in d}\sum_z p(w|z)*p(z|d)*p(d) \tag{2}$$

通过采用 EM 算法求解带约束的优化问题 $\max L(D)$，可以得到分布 $p(w|z)$ 和 $p(z|d)$。

然而，PLSA 有时会出现过拟合的现象，即一个在训练集上能够比其他假设获得更好拟合的假设在测试集上却不能很好地拟合，为了避免出现过拟合，PLSA 使用一种最大似然估计方法——期望最大化，本文第 3 节将详细介绍该算法。PLSA 会将 doc-topic 这一级的所有变量都作为模型的参数，因此训练参数 $p(z|d)(d\in D)$ 会随着文档集的增加而线形增加，计算也会更加复杂；此外，PLSA 只可以生成所在数据集的文档模型，不能生成新的文档模型。针对以上缺点，Blei 等人在 PLSA 的基础上提出了 LDA。

2.2 LDA 模型

LDA 模型可以看作是 PLSA 的贝叶斯扩展，将 PLSA 中的模型参数即主题在一篇文档中的分布 $p(z|d)(d\in D)$ 作为随机变量，在 LDA 中其由先验参数随机生成。因而其在一定程度解决了 PLSA 参数过多容易造成的过拟合问题，以及训练得到的模型难以在训练集以外的文档上使用的问题。参照图 2，LDA 模型的生成过程如下。

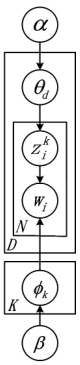


图 2 LDA 三层模型

对于每篇文档 $d\in D$：

(1)从先验参数为 α 的狄利克雷分布中选择一个多项式分布 θ_d，$\theta_d\sim$Dirichlet(α)；

(2)对于文档 d 中的每个词 w_i，从主题分布中选择一个主题 $z_i\sim\theta_d$；

(3)从先验分布为 β 的狄利克雷分布中选择一个多项式分布 ϕ_k，$\phi_k\sim$Dirichlet(β)；

(4)根据选择的主题 z_i，从代表这个主题的词的多项式分布中选择当前的词 $w_i\sim\phi_k$。

与 PLSA 中通过最大似然得到最优参数不同，分布 $p(z|d)(d\in D)$ 即 $\vec{\theta}_d$ 在 LDA 中也是随机变量而不是可以优化的参数。在这种情况下可以通过吉布斯采样或变分推导的方法计算后验概率来估算求解参数。

2.3 LDA 扩展模型

目前，关于主题模型的研究工作一部分集中在对 LDA 模型的扩展上，有的对参数进行扩展，有的在模型中引入上下文的信息，有的面向特定的任务对模型进行扩展。对参数进行扩展的模型有相关主题模型(Correlated Topic Model，CTM)[11]，该模型从对数正态分布中采样出主题的概率分布，先验参数中包含一个描述主题相关性的协方差矩阵；Blei 等人提出的动态主题模型(Dynamic Topic Models，DTM)[12]认为主题会随着时间变化。从图 3 中可以看出 2 个超参数都会随时间变化，并且依赖于前一时刻。

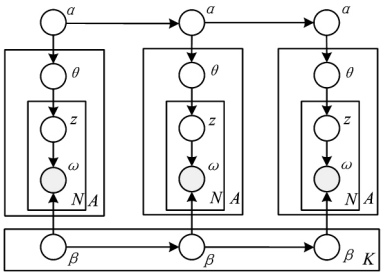


图 3 动态主题模型

作者主题模型(Author Topic Model，ATM)[13]是面向特定任务的扩展模型，该模型在生成过程中会随机选择一个作者，根据作者的主题概率分布生成词，重复这个过程直到生成一篇文档。LDA 的扩展模型十分丰富，这里不一一列出。就目前的研究结果来看，主题模型的扩展大部分集中在面向特定任务上，参数扩展和引入上下文的信息的扩展相对较少，主要原因是参数扩展和引入上下文的信息需要对主题模型的整体进行修改，可以入手的研究点相对较少。

3 参数估计方法

3.1 EM 算法

由于各种原因，现实数据库中经常会存在不同程度的数据丢失现象，已不能用参数之间的独立性进行计算。对于不完整的数据集，常采用的方法有 EM 算法、Gibbs 抽样方法等。本节主要介绍 EM 算法在主题模型中的应用。之所以称之为 EM 算法，是因为算法的每一次迭代由一个期望步(E-step)和极大步(M-step)构成。EM 算法[5]是由 Dempster 等人提出的在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型[14]依赖于无法观测的隐藏变量(Latent Variable)。该算法通过不断迭代修改模型参数直到达到局部最优点，即每次都用现有的模型推断隐藏变量的后验概率分布，然后对参数重新估计得到一个新的模型，如此反复直到满足终止条件。由于 EM 算法不能保证全局最优解，因此有时需要变换参数的初始值，或者选择较多的迭代次数，才能得到较为理想的参数估计值。

在 PLSA 模型中 $p(z_k|d_i)$ 和 $p(w_j|z_k)$ 分别对应了两组多项分布，对这两组分布的参数进行估计时，常用的估计算法是 EM 算法。具体步骤如下。

E-step：计算隐藏主题 z 的后验概率。

$$P(z_k|d_i,w_j)=\frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{k}P(w_j|z_l)P(z_l|d_i)} \tag{3}$$

M-step：对后验概率期望最大化。

$$P(w_j|z_k)=\frac{\sum_{i=1}^{N}n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N}n(d_i,w_m)P(z_k|d_i,w_m)} \tag{4}$$

$$P(z_k|d_i)=\frac{\sum_{j=1}^{M}n(d_i,w_j)P(z_k|d_i,w_j)}{n(d_i)} \tag{5}$$

其中，$P(d_i)$ 表示单词出现在文档 d_i 中的概率，$P(z_k|d_i)$ 表示文档 d_i 中出现主题 z_k 下单词的概率词的概率，$P(w_j|z_k)$ 表示给定主题 z_k 出现单词 w_j 的概率。

EM 算法的特点是简单稳定，特别是每一次迭代能保证观察数据对数的后验似然是单调不减的，这一点可以利用 Jensen 不等式的性质来证明。但 EM 算法需要反复迭代，计算量大，收敛速度慢。

3.2　变分贝叶斯推理算法

变分贝叶斯方法[7] 由 Matthew J. Beal 提出，它是用于贝叶斯估计和机器学习领域中近似计算复杂积分的技术。它主要应用于复杂的统计模型中，这种模型一般包括 3 类变量：观测变量、未知参数和隐变量。在贝叶斯推理中，参数和隐变量统称为不可观测变量。变分贝叶斯方法主要有两个目的：

(1)近似不可观测变量的后验概率，以便通过这些变量作出统计推断。

(2)对一个特定的模型，给出观测变量的边缘似然函数的下界，主要用于模型的选择，若模型的边缘似然值越高，则模型对数据的拟合程度越好。

变分贝叶斯是用来估计后验分布的方法，变分贝叶斯推理过程为：当遇到无法直接计算后验分布的情况时，可以设置一个函数来近似它，将该函数设为 Q，其后验分布设为 P，接下来的目的就是使 P 和 Q 之间的差最小，因此在使用变分推断前，需要选取 Q 的分布形式，在 LDA[3] 中选取：

$$q(\theta,z|\gamma,\emptyset)=q(\theta|\gamma)\prod_{n=i}^{N}q(z_n|\emptyset_n) \tag{6}$$

选取了近似函数 Q 之后，可以使用 KL 距离(Kullback-Leibler divergence)来计算 P 和 Q 之间的差，经过推导可以得到：

$$\emptyset_{ni}\propto\beta_{iw_n}\exp\{E_q[\log(\theta_i)|\gamma]\} \tag{7}$$

$$\gamma_i=\alpha_i+\sum_{n=1}^{N}\emptyset_{ni} \tag{8}$$

最后，进行 EM 迭代，直到 α 和 β 收敛。

变分贝叶斯方法是通过在传统贝叶斯推理与 EM 迭代估计算法的基础上引入变分近似理论而提出的。变分贝叶斯通过最大化变分参数的对数边缘似然函数的下界来求解模型的参数，利用统计物理学中均值域理论[15]，将多变量的联合概率分布近似为各变量边缘概率分布的乘积，使得对多变量的联合估计方便地转化为对这些变量边缘分布的迭代估计，计算复杂度显著降低，运算效率得到提高。VB 算法利用一个可以分解且方便优化的近似下界函数逼近后验概率函数，由于下界函数与目标函数之间存在误差，收敛精度不高，为了弥补这一缺陷，引入了 digamma 函数，其也降低了 VB 算法的计算效率。文献[5]中，在未引入 digamma 函数的情况下数据集的训练时间仅为引入 digamma 函数的情况下的 3/10。因此可以认为 digamma 函数的引入导致了计算效率的降低。

3.3　Gibbs 抽样算法

Gibbs 抽样方法是由 Heckerman 提出的一种用于不完备数据参数学习的方法。Gibbs 抽样方法是 MCMC 算法[14] 的一个特例。该算法的运行方式是每次选取概率向量的一个维度，给定其他维度的变量值，采样当前维度的值。不断迭代，直到收敛输出待估计的参数。

Gibbs 抽样算法在 LDA 模型中的抽样过程如下：首先从文本集合中抽取一个词标记，在其它所有词标记和主题给定的条件下，选定的词分配给一个主题的概率为 $P(z_i=j|z_{-i},w_{di},\alpha,\beta)$。然后从中抽取一个主题 z_i 来取代当前词的主题，不断循环这个过程，最终会收敛于一个不变点。具体的计算公式如下：

$$P(z_i=j|z_{-i},w_{di},\alpha,\beta)\propto\frac{n_{-i,j}^{(w)}+\beta_{i,j}}{\sum_{u=1}^{v}(n_{-i,j}^{(w)}+\beta_{u,j})}\times\frac{n_{-i,j}^{(d)}+\alpha_{d,j}}{\sum_{k=1}^{k}(n_{-i,k}^{(d)}+\alpha_{d,k})} \tag{9}$$

其中，$n_{-i,j}^{(w)}$ 表示单词 w 被分配给主题 j 而没有包含当前主题 i 的次数，$n_{-i,j}^{(d)}$ 表示在文档 d 中分配给主题 j 的词元而没有包含当前主题 i 的次数。最终，可以得到：

$$\varphi_j^{(w)}=\frac{n_j^{(w)}+\beta_{w,j}}{\sum_{u}^{v}(n_{u,j}^{(w)}+\beta_{u,j})} \tag{10}$$

$$\theta_j^d=\frac{n_j^{(d)}+\alpha_{w,j}}{\sum_{k=1}^{k}(n_{j,k}^{(d)}+\alpha_{d,k})} \tag{11}$$

凭借以上两个公式，可以对 φ 和 θ 进行评估。

Gibbs 抽样算法在精度和收敛速度上比 VB 算法好。在实际应用中，Gibbs 抽样算法需要扫描海量文档集中的每个单词，当训练文档的单词数目过大时，Gibbs 抽样算法的效率会大大降低。

3.4　BP 算法

因子图模型可以将主题模型的问题抽象成一个标签问题。它的目的就是给文本-词矩阵里观测到的非零元素分配语义主题标签。其基本思想来源于 collapsed Gibbs 抽样算法。在 LDA 中运用 BP 算法进行参数估计时，需要将 LDA 模型图转换为因子图，图 4 是 LDA 模型的因子图表示，从中可见消息的传递方向，首先，定义主题标签 $z_{w,d}^k$ 的邻域系统 $z_{-w,d}^k$ 和 $z_{w,-d}^k$；接着，根据消息的传递方向设置因子函数，用来表示奖励或者惩罚邻域系统中不同的局部标签，从而实现主题模型的 3 个本质假设：共现、平滑和聚集。图 4 中的方框表示因子，圆圈表示连接因子间的变量，因子 θ_d 连接的是同一个文本不同词索引的主题标签，而因子 $\emptyset_d$ 连接的是词索引相同但文本不同的主题标签。

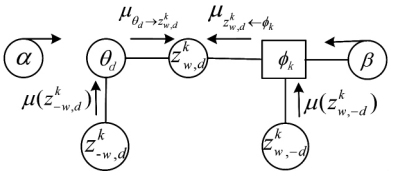


图 4　LDA 模型因子图

图 2 中 LDA 的联合概率可以表示为：

$$P(x,z,\theta,\phi|\alpha,\beta)\propto\prod_{d=1}^{D}P(\theta_d|\alpha)\prod_{k=1}^{K}P(\phi_k|\beta)\prod_{d=1}^{D}\prod_{i=1}^{N_d}P(z_i^k|\theta_d)P(w_i|z_i^k,\phi_k) \tag{12}$$

基于狄利克雷多项共轭关系，整合出式(10)中的多项式参数$\{\theta_d,\emptyset_d\}$，因此 LDA 的联合可以表示为式(13)：

$$\begin{aligned}&P(x,z|\alpha,\beta)\\&=P(x|z,\beta)P(z|\alpha)\\&\propto\prod_{d=1}^{D}\prod_{k=1}^{K}\frac{\Gamma(\sum_{w=1}^{W}x_{w,d}z_{w,d}^{k}+\alpha)}{\Gamma[\sum_{k=1}^{K}(\sum_{w=1}^{W}x_{w,d}z_{w,d}^{k}+\alpha)]}\times\\&\prod_{w=1}^{W}\prod_{k=1}^{K}\frac{\Gamma(\sum_{d=1}^{D}x_{w,d}z_{w,d}^{k}+\beta)}{\Gamma[\sum_{w=1}^{W}(\sum_{d=1}^{D}x_{w,d}z_{w,d}^{k}+\beta)]}\end{aligned}\tag{13}$$

该算法用来计算条件后验概率 $p(z_{w,d}^{k}=1,x_{w,d}\mid z_{-w,-d}^{k},x_{-w,-d})$，又称为消息 $\mu(z_{w,d}^{k})$，结合 BP 算法和因子图，运用贝叶斯算法可以得到：

$$\begin{aligned}&\mu(z_{w,d}^{k})\\&\propto\frac{p(z_{\cdot,d}^{k},k,x_{\cdot,d})}{p(z_{-w,d}^{k},x_{-w,d})}\times\frac{p(z_{w,\cdot}^{k},x_{w,\cdot})}{p(z_{w,-d}^{k},x_{w,-d})}\\&\propto\frac{(\sum_{-w}x_{-w,d}z_{-w,d}^{k}+\alpha)}{\sum_{k=1}^{K}(\sum_{-w}x_{-w,d}z_{-w,d}^{k}+\alpha)}\times\frac{\sum_{-d}x_{w,-d}z_{w,-d}^{k}+\beta}{\sum_{w=1}^{W}(\sum_{-d}x_{w,-d}z_{w,-d}^{k}+\beta)}\end{aligned}\tag{14}$$

但是我们只知道邻居消息而不知道精确的主题分配情况，因此用相关的信息取代主题配置，从而式(13)可以表示为：

$$\mu(z_{w,d}^{k})\propto\frac{\mu(z_{-w,d}^{k})+\alpha}{\sum_{k}[\mu(z_{-w,d}^{k})+\alpha]}\times\frac{\mu(z_{w,-d}^{k})+\beta}{\sum_{w}[\mu(z_{w,-d}^{k})+\beta]}\tag{15}$$

这里运用 EM(Expectation-Maximization)算法来估计多项式参数 θ_d 和 $\emptyset_d$，直到最大循环次数。

$$\theta_d(k)=\frac{\mu(z_{\cdot,d}^{k})+\alpha)}{\sum_{w}[\mu z_{\cdot,d}^{k})+\alpha]}\tag{16}$$

$$\phi_w(k)=\frac{\mu(z_{w,\cdot}^{k})+\beta}{\sum_{w}[\mu(z_{w,\cdot}^{k})+\beta]}\tag{17}$$

BP 算法是相对于以上几种算法来说更加快速和高精度的学习算法。文献[8]表明 BP 算法在训练速度和精度上均优越于 GS 和 VB 算法。

4 主题模型的评估

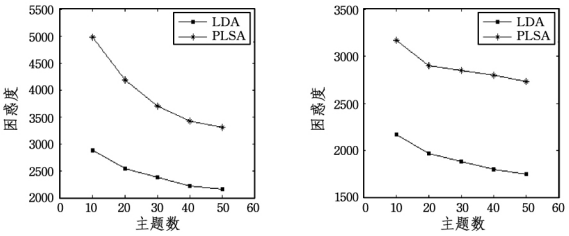
针对上述提出的各种主题模型，我们需要判断他们的“好坏”，如何对主题模型进行评估一直受到人们的关注。目前，对模型的评估方法主要两大类：1)实用方法，通过模型在实际应用中的表现结果来评价；该方法比较直观，但缺乏针对性，模型间的比较存在客观性的问题。2)理论方法，目前常用的是困惑度(Perplexity)，其衡量模型建模能力的好坏。通常，困惑度计算公式如下：

$$Perplexity(D)=\exp\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\}\tag{18}$$

其中，M 是文本的数量，N_d 是文本 d 中词汇的综述，$p(w_d)$ 表示文本概率。一般认为困惑度越小，模型的建模效果越好。

另外，还可以通过一些间接的任务来评估模型的好坏，如文档的相似度、主题间的相似度等。针对不同的任务，每个模型的优点可能不一样，且比较具有相对片面性，因此通过间接任务来评估模型的方法只适合在特定任务的情形下。

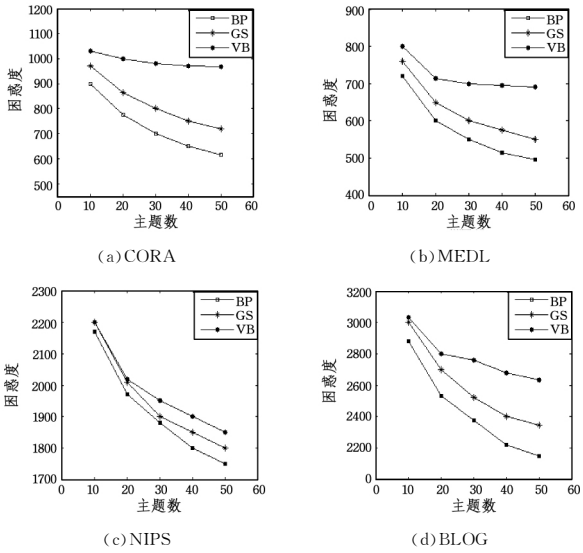
图 5 是不考虑数据集大小的影响，在两个数据集上对 LDA 模型和 PLSA 模型的困惑度进行比较的结果。NIPS 主要是神经信息处理系统领域的文章；BLOG 主要包含了美国 2008 年的政治博客集合，博客没有作者的信息。所有实验一直迭代，直到模型收敛为止，主题数均为 10～50，步长为 10，在 Win7 下用 MATLAB 和 MEX C++进行实验。



(a)BLOG (b)NIPS

图 5 LDA 和 PLSA 的困惑度比较

实验的参数估计算法均采用 BP 算法。从图 5 中可以看出，在相同的训练数据集上，LDA 模型训练的困惑度低于 PLSA 模型。在对模型和参数估计算法进行评估时，还可以通过比较训练耗时等来比较他们的性能。EM 算法是一种处理含有缺失数据的最大似然估计问题的专门的方法，通过两个交替进行的步骤，算法将收敛到一个局部最优解。在实际应用中，无法保证算法能够找到最大似然估计的全局最优解；当数据量很大时，EM 算法需要反复迭代，计算量相当大，收敛较慢。Gibbs 算法的优势在于简单，缺点是一次只能更新一个变量，效率较低。VB 近似推理算法优化的是目标函数的下界函数，理论上下界函数与原始的目标函数之间存在误差，这就导致 VB 算法的精度不高，因此引入了 digamma 函数，但这同时降低了 VB 算法的计算效率。因此，VB 算法比 GS 的学习精度略差。对海量文档进行处理时，VB 算法的收敛速度要快于 GS 算法。BP 算法在训练速度和精度上均比 GS 和 VB 算法优越，BP 算法每次迭代将重新计算和更新所有的消息，然而多数情况下，只有少数信息在相连的两次更新中有较大变化，而大多数的消息更新几乎不改变计算结果，因此会造成计算资源的浪费。文献[8]基于 4 组文本数据集在 LDA 模型上对 BP，GS 和 VB 3 种算法进行了实验比较。



(a)CORA (b)MEDL

(c)NIPS (d)BLOG

图 6 4 个数据集上的预测性困惑度比较

图 6 分别示出了在 CORA，MEDL，NIPS 和 BLOG 4 个数据集上的预测性困惑度比较结果，CORA 主要包括机器学习研究领域的文章摘要；MEDL 主要包括生物医学领域的文

(下转第 47 页)

[18] WANG Y F，PETRINA S，FENG F. VILLAGE-Virtual immersive language learning and gaming environment：immersion and presence [J]. British Journal of Educational Technology，2015，12(11)：1-20.

[19] 王扬，郭晨，章晓明. 现代仿真技术[M]. 北京：国防工业出版，2010：58.

[20] HWANG G H，LEE C Y，HWANG H L，et al. Using augmented reality to assist an interactive multi-language learning system in an elementary school[C]// Workshop Proceedings of the 21st International Conference on Computers in Education. Indonesia，2013：404-411.

[21] KNUTZEN B，KENNEDY D. The global classroom project：learning a second language in a virtual environment[J]. Electronic Journal of E-learning，2012，10(1)：90-106.

[22] CHENG J L，CHEN C. The crossroads of English language learners，task-based instruction，and 3D multi-user virtual learning in Second Life[J]. Computers & Education，2016，102(25)：152-171.

（上接第 32 页）

章摘要：NIPS 主要是神经信息处理系统领域的文章；BLOG 主要包含了美国 2008 年的政治博客集合，博客没有作者的信息。实验过程中设置固定的超参数 $\alpha=\beta=0.01$。

从图 6 可以看出，BP 算法的困惑度相对于 GS 和 VB 算法较低，也就说明 BP 算法训练 LDA 模型时具有更好的预测性能。4 个数据集包含的总文本大小分别是 2410，2317，1740，5177。从图 5 可以观察到，数据集越大，在同等条件下预测性困惑度也会越大。

本节主要介绍了 EM，VB 和 BP 这 3 种基本的参数估计算法。实际应用中，由于文本数量巨大以及文本流的时序特征，有些研究开始关注 LDA 的快速推理算法[16]、在线学习[17]、文本流的推理算法[18-19]、分布式学习[20-22]。这些研究将会很好地提高 LDA 模型求解的效率，同时也更适应现在大数据时代下信息飞速增长的趋势。

结束语　本文介绍了主题模型发展过程中几种比较经典的参数估计方法，比较了不同算法的优缺点和适用范围。在研究过程中发现，尽管主题模型看似成熟，参数估计的学习仍然有很多问题值得进一步探索：1）可以针对特定的任务改进参数估计方法，例如针对一段时间内变化不明显的消息文本，是否可以尝试用残余信念传播算法代替 BP 算法；2）近年来，大数据的方向比较热门，可以关注如何处理海量实时信息，将其应用到特定的场景。

参 考 文 献

[1] DEERWESTER S C，DUMAIS S T，LANDAUER T K，et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science，1990，41(6)：391-407.

[2] HOFMANN T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International SIGIR Conference. New York：ACM Press，1999：50-57.

[3] BLEI D，NG A，JORDAN M. Latent Dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022.

[4] 徐戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报，2011(8)：1423-1436.

[5] DEMPSTER A P，LAIRD N M，RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society，1997，B39(1)：1-38.

[6] GRIFFITHS T L，STEYVERS M. Finding scientific topics [J]. Proceedings of the National academy of Sciences of the United States of America，2004，101(Suppl 1)：5228-5235.

[7] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research，2003，3：993-1022.

[8] ZEHG J，CHEUNG W K W，Liu J M. Learning topic models by belief propagation [C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012.

[9] 赵鑫，李晓明. 主题模型在文本挖掘中的应用：PKU-CS-NCIS-TR2011XX[R]. Technical Report，2011.

[10] HEINRICH G. Parameter estimation for text analysis[R]. Technical Report，2008.

[11] BLEI D M，LAFFERTY J D. Correlated topic models[C]// Advances in Neural Information Processing Systems 18. Cambridge，MA：MIT Press，2006.

[12] BLEI D，LAFFERTY J. Dynamic topic models[C]// Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh，Pennsylvania，USA，2006：113-120.

[13] ROSEN-ZVI M，GRIFFITHS T，STEYVERS M，et al. The Auther-Topic Model for Author-Topic Model for Authors and Documents[C]// Proc. 20th Conf. Uncertainty in Artificial Intelligence. 2004：487-494.

[14] BISHOP C M. Pattern Recognition and Machine Learning[M]. New York，USA：Springer，2006.

[15] 赖裕平，周亚建，丁洪伟，等. 混合逆狄利克雷分布的变分学习及应用[J]. 电子学报，2014，42(7)：1435-1440.

[16] PORTEOUS I，NNWMAN D，IHLER A，et al. Fast collapsed gibbs sampling for latent dirichlet allocation[C]// Proceeding of The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'08). New York，NY，USA，2008：569-577.

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

[18] YAO L M，MIMNO D，MCCALLUM A. Efficient methods for topic model inference on streaming document collections[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'09). New York，NY，USA，ACM，2009：937-946.

[19] BANERJEE A，BASU S. Topic Models over Text Streams：A Study of Batch and Online Unsupervised Learning[C]// SDM. SIAM，2007.

[20] NEWMAN D，ASUNCION A，SMYTH P，et al. Distributed Inference for Latent Dirichlet Allocation[C]// Conference on Neural Information Processing Systems. 2007：1-6.

[21] LIU Z Y，ZHANG Y Z，CHANG E Y，et al. Plda+：Parallel latent dirichlet allocation with data placement and pipeline processing[J]. ACM Trans. Intell. Syst. Tschnol.，2011，2(3)：26

[22] YAN F，XU N Y，QI Y. Parallel inference for latent dirichlet allocation on graphics processing units[C]// NIPS. 2009：2134-2142.

[23] ASUNCION A，SMYTH P，WELLING M. Asynchronous distributed learning of topic models[C]// NIPS. 2008：81-88.