基于在线消息传递的主题追踪方法

龚声蓉 叶 芸 刘纯平 季 怡
(苏州大学计算机科学与技术学院 江苏 苏州 215006)

摘 要 主题追踪因可以有效地汇集和组织分散在不同时间、地点的信息,并从主题层次的角度对某个主题相关事件的时效性、动态演化关系等得到比较全面的把握,成为当前数据挖掘领域的重要研究方向.现有基于概率主题模型的主题追踪方法主要以潜在狄利克雷分布(Latent Dirichlet allocation,LDA)模型为基础,采用在线吉布斯采样(Online Gibbs Sampling,OGS)和在线变分贝叶斯(Online Variational Bayesian,OVB)算法进行参数估计.OGS 和 OVB 算法尽管解决了 LDA 模型中使用传统离线近似推理方法所需内存空间的大小随数据集的增长而不断增加,无法训练海量数据集以及数据流数据的问题,但训练的精度和速度均有待提高.该文基于 LDA 模型的改进因子图提出了一种在线消息传递(Online Belief Propagation,OBP)的主题追踪算法.该算法借助因子图中消息传递(Belief Propagation,BP)算法的推理,通过切分海量数据集为段,并用前一段数据集训练后的参数计算当前段的梯度下降,使得主题追踪更加快速和准确.四组大规模文本数据集的实验对比表明,LDA 模型中 OBP 算法在速度和精度上均优越于 OGS 和 OVB 算法,文中也从理论上进一步验证了 OBP 算法的收敛性,并给出了主题追踪的具体应用.

关键词 潜在狄利克雷分布;吉布斯采样;变分贝叶斯;消息传递算法;主题追踪;社交网络;社会计算
中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2015.00249

Topic Tracking Based on Online Belief Propagation

GONG Sheng-Rong YE Yun LIU Chun-Ping JI Yi
(*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)

Abstract Given topics, probabilistic topic model based topic tracking automatically determines the attributions, such as effectiveness, dynamic evolution etc., from topic related events with different place and time stamped documents, which has become a focus research of big data mining. Most of existing topics tracking methods using topic model adopt Online Gibbs sampling (OGS) and Online Variational Bayesian (OVB) to infer the parameters of Latent Dirichlet Allocation (LDA) model. Solving the problems of the growth of memory size with the increasing data sets in traditional offline algorithms of LDA model, OGS and OVB can be used to deal with massive datasets as well as data streaming, but the accuracy and speed of both methods need to be further improved. In this paper, we propose Online Belief Propagation (OBP) based on improved factor graph of LDA model for tracking the changing of topic. Turning to the approximate inference of OBP algorithm on factor graph model, the proposed method first splits the massive data set into different segments, and then computes the gradient descent of the current segment using the results of previous training parameter. This results in topic tracking with the faster and more accurate. Experimental results of four big data sets indicate that OBP algorithm outperforms both OGS and OVB from the accuracy and speed of topic tracking. Moreover, this paper also proved

收稿日期:2013-03-22;最终修改稿收到日期:2014-08-26.本课题得到国家自然科学基金(61272258,61170124,61170020,61301299)资助.**龚声蓉**,男,1966 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、机器学习等.E-mail:shrgong@suda.edu.cn.**叶 芸**,女,1988 年生,硕士,主要研究方向为机器学习等.**刘纯平**(通信作者),女,1971 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为图像与视频处理、模式识别、计算机视觉与机器学习等.E-mail:cpliu@suda.edu.cn.**季 怡**,女,1973 年生,博士,副教授,主要研究方向为计算机视觉、机器学习等.

theoretically the convergence of the proposed OBP algorithm. In addition, the specific application of topic tracking is presented.

Keywords dirichlet allocation; gibbs sampling; variational bayes; belief propagation; topic tracking; social networks; social computing

1 引 言

随着信息技术的高速发展,来源于门户网站、电子商务网站、社交网站、论坛、博客和微博等信息正以指数级的方式增长.搜索引擎虽然可以方便的提供很多信息资源,但却不能有效地发现和管理与某一主题相关的信息.如何从上述海量的文本集中寻找热点话题成为当前信息检索(Information Retrieval,IR)领域的研究关键,而基于概率主题模型的主题追踪方法则可以有效地发现与该主题相关的信息[1].

现有基于概率主题模型的主题追踪方法,利用快速的学习算法从高维稀疏的单词数据中提取低维的主题表示,从而追踪主题的不断变化趋势.常见的主题模型有空间向量模型[2]、潜在语义分析(Latent Semantic Analysis,LSA)[3]模型、概率潜在语义分析(Probabilistic Latent Semantic Analysis,PLSA)[4]模型和潜在狄利克雷分布(LDA)模型[5]等.PLSA模型和LDA模型通过联合概率描述文本单词和主题之间的关系,并且每个概率均有合理的物理解释,从而能够很好地解决文档聚类的问题,是目前最常用的主题模型.应用这两种模型的核心在于学习模型中的参数,而参数的个数对模型的复杂程度有很大的影响.PLSA模型中的参数个数会随训练文本的不断增加而增加,LDA模型对外始终只有两个超参,因此LDA模型更有利于训练海量数据集.

LDA模型学习的关键在于从主题和单词的联合概率中推断出在可观测变量下主题的后验概率分布,但是无法直接通过后验概率分布求解模型中的参数,一般需要采用近似后验推理方法.目前广泛采用的近似推理方法有吉布斯采样(Gibbs Sampling,GS)[6]、变分贝叶斯(Variational Bayesian,VB)[5]和离线消息传递(Belief Propagation,BP)算法[7].

离线GS、VB和BP近似推理算法已经在小规模数据集上取得了应用.文献[5]提出了VB近似推理算法对文本数据分类,在平均包含16 000篇文本的文本集上的实验表明,LDA模型比PLSA模型在分类速度和精度方面有实质性的提高.文献[6]首次提出了GS近似推理算法,并对图像进行分类.在2000幅,每幅大小为5×5像素的图像数据集上的实验验证了GS比VB收敛的速度快且精度高.文献[7]基于LDA模型,在四组文本数据集上对比分析了BP、GS和VB三种近似推理算法.实验表明BP近似推理算法的精度和训练速度均优于GS和VB算法.这些离线算法尽管简单稳定,但收敛速度通常很慢,且需要将整个训练集加载到内存.由于实际中往往处理的是大规模的实时流数据,如博客等流型数据,离线算法因数据集本身不完全以及内存不足而无法处理.

为了克服离线算法在处理这类流数据时的缺陷,将海量数据集切分成若干小段,然后顺序处理每一段数据的在线学习算法成为一个首选.对每个时间段,在线算法只加载一小段数据到内存,并对当前段用梯度下降法[8]估计模型的参数,在当前段训练结束后,将该段数据集移出内存,再加载下一段数据集进行学习.目前在线学习算法已经在主题模型的近似推理[9-14]、目标检测[15-16]、大规模矩阵分解[17-18]、高维数据分析[19]和SVM中的核函数在线学习[20]等众多方面取得了广泛的应用.Canini等人[21]提出了基于LDA模型的增长式吉布斯采样(Incremental Gibbs Sampling,IGS)算法,即在线吉布斯采样(OGS)算法;Hoffman等人[12]提出了在线变分贝叶斯(OVB)算法.这两类在线近似推理方法在分类时比离线算法需要较少的迭代次数就能达到收敛,此外需要的内存空间是固定的,仅与每一小段数据集的大小成比例.但OGS和OVB方法以离线GS和VB算法为基础,而GS算法在近似推理时需要对所有文本中的每个单词训练,VB算法中引入了时间复杂度较高的digamma函数,这就导致OGS和OVB算法的精度和速度都有待提高.

在OGS在线学习算法的基础上,文献[22]和文献[10]提出了基于OGS的在线主题追踪算法.为了提高主题追踪的精度和速度,本文提出基于LDA模型的在线消息传递(OBP)算法,该算法可以将隐藏变量的联合概率分布分解成因子间的乘积,计算其后验边界概率而非后验联合概率,即变量与因子

之间传递的消息，而消息可以通过本地计算并归一化得到. 文献[23]已提出了基于 PLSA 模型的 OBP 算法，由于 PLSA 模型中包含的参数会随训练文本的不断增加而增加，使得模型的复杂度也不断增加. 不同于文献[23]，LDA 模型是将参数看作变量，引入了两个语料库级超级参数，使得模型对外始终只有两个参数. 实验表明 LDA 模型下的 OBP 算法比 PLSA 模型下的 OBP 算法更优越，且 LDA 模型下的 OBP 算法比 OGS 和 OVB 算法更准确和快速.

更具体的说，OBP 算法是把海量数据集切分成若干独立小段. 训练时先随机初始化第 1 段数据集的参数，训练结束后保存训练结果，而从第 2 段到最后一段，OBP 算法将前一段数据集训练的结果作为当前段参数的初始化，然后依次训练每一段数据集. 对于每段数据集，OBP 算法使用随机优化方法稳定学习后的参数，确保 OBP 算法收敛到目标函数的局部最优，最后本文从理论方面进一步验证了 OBP 算法的有效性.

本文第 2 节简单介绍 LDA 模型及现有在线学习算法分析；第 3 节介绍 LDA 模型的因子图表示，给出应用于大规模数据集的 LDA 模型的 OBP 算法，并从理论上证明 OBP 算法的收敛性；第 4 节给出 OBP 算法在 4 个大规模数据集下实验对比及主题追踪的具体应用；最后一节为总结.

2　LDA 及其在线学习算法分析

2.1　LDA 模型概述

LDA 模型是将实际可观测的“文档-词”的高维稀疏空间，通过快速的学习算法降低到低维空间，图 1 给出了 LDA 模型的 3 层概率图表示. LDA 模型由单词、文本和语料库 3 层构成. 其中单词层包括可观测的单词 $w_n(1\leq w_n\leq W)$ 和隐藏主题 $z_n(z_n=k, 1\leq k\leq K)$；文本层包括指定文本所对应的主题分布 θ_d 和指定主题对应单词表的概率分布 ϕ_k 和；语料层包括控制文本层 θ_d 和 ϕ_k 变量的 α 和 β 超参，D 为语料库中总文档数，N 为平均每篇文本的单词数，W 为单词表大小，K 为总主题数. 模型中用到的符号标记说明如表 1 所示. 在每篇文档只有单一主题的假设前提下，展开下面的分析和讨论.

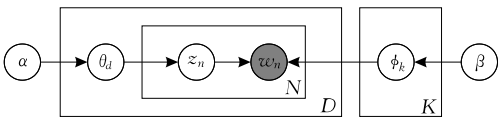


图 1　LDA 模型的概率图表示[5]

表 1　符号标签说明

符号	意义
$1\leq d\leq D$	文本索引
$1\leq w\leq W$	单词表中单词索引
$1\leq k\leq K$	主题索引
$w=\{w,d\}$	词袋
$z=\{z_{w,d}\}$	单词的主题标签
$z_{-w,d}$	文本 d 中除 w 外所属的主题
$z_{w,-d}$	单词 w 除文本 d 外所属的主题
θ_d	文本 d 的因子
ϕ_w	单词 w 的因子
α,β	狄利克雷超参数

LDA 是一个生成模型，即文本可以由多个隐藏主题混合而构成. 基于 LDA 模型，文本生成的过程如下：

(1) 根据先验分布 $\theta_d\sim Dirichlet(\alpha)$，随机选择一个多项式分布 θ_d，其中 $1\leq d\leq D$，确定文本主题分布；

(2) 根据先验分布 $\phi_k\sim Dirichlet(\beta)$，随机选择一个多项式分布 ϕ_k，其中 $1\leq k\leq K$，确定该主题下词表中的单词分布；

(3) 对文本 d 中的每个单词 w，$1\leq w\leq W$：首先根据 $z_j\sim Discrete(\theta_d)$ 选择一个主题 z_j，然后根据 $w_i\sim Discrete(z_j)$ 从被选中主题所对应的单词分布中选择一个单词 w_i，其中 $1\leq i\leq N$.

2.2　OGS 和 OVB 算法分析

LDA 模型的目标是在给定文本数据集 $w=(w_1,\cdots,w_N)$ 的条件下，推断出文本对应的主题分布 θ，主题对应单词表的概率分布 ϕ 和单词所属隐藏主题变量 $z=(z_1,\cdots,z_N)$ 分布. 但后验概率分布 $p(\theta,\phi,z|w)$ 的复杂性使得我们不能直接求解，而是通常采用离线 GS、VB、BP 和在线 OGS、OVB 近似推理算法. 在线算法 OGS 和 OVB 分别以离线 GS 和 VB 算法为基础，而 GS 算法从后验边际概率 $p(z)$ 中，对每个单词 w 采样一个主题标签 z. 理论上而言，多次扫描迭代后 $p(z)$ 会收敛到真实后验概率分布. 由于 GS 算法需对每个单词扫描，当每篇文本中单词数量较大，扫描时间必然增加. 此外，GS 算法收敛速度很慢，实际中需要对文档-词汇矩阵扫描 500～1000 次才会收敛. 因此，GS 算法无法满足对海量数据的处理. 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 VB 算法中下界函数与真实目标函数间存在误差，收敛时精度不如 GS 算法. 因此为了克服这一不足，引入了较复杂的 digamma 函数，但这大大降低了 VB 算法的计算效率，甚至使其收敛速度低于 GS 算法. 鉴于

此，文献[6]提出了离线 BP 算法，并验证了 BP 算法优于离线 GS 和 VB 算法. 针对海量数据训练，直接利用在线 OGS 和 OVB 算法，其精度和速度都有待改善. 本文借助基于离线 GS 和 VB 算法的在线 OGS 和 OVB 算法构建思想，提出了基于离线 BP 的在线 OBP 算法.

3 LDA 模型在线消息传递算法

3.1 LDA 模型的因子图表示

离线消息传递 BP 算法[7]是一种从马尔可夫框架推导出的新颖近似推理方法，为了推理消息传递，Zeng 等人[7]将传统 LDA 模型的概率图表示(图 1)转变为因子图表示(图 2).

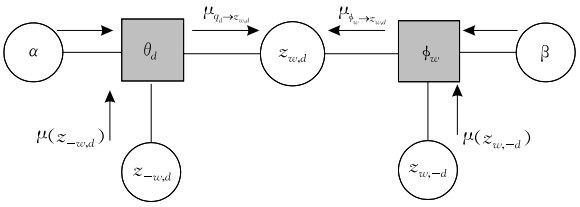


图 2 LDA 模型的因子图[7]

离线 BP 算法将主题模型视为贴标签问题，即为单词表中所有单词索引 $w=\{w,d\}$ 赋予语义标签 $z=\{z_{w,d}\}$. 在无向概率图模型中，马尔可夫模型可以借助邻居系统和团势函数，基于最大化后验估计获得的最大化后验概率指派最佳主题标签. 因此因子图表示的 LDA 模型的离线 BP 算法，首先定义主题标签 $z_{w,d}$ 的邻居系统 $z_{-w,d}$ 和 $z_{w,-d}$，其中 $z_{-w,d}$ 表示除 w 外，文本 d 中所有单词的主题标签，$z_{w,-d}$ 表示除文本 d 外，单词 w 在所有文本中的主题标签；其次设置合适的团势函数，以惩罚或奖励邻居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[7].

图 2 中方框表示因子 θ_d 和 ϕ_w，圆圈表示的 $z_{w,d}$ 是两个因子间的连接变量. 由于图 1 和图 2 具有相同的邻居系统、相同的连接隐藏变量以及团势函数，因此从主题模型角度而言，图 2 与图 1 等价.

3.2 LDA 模型的 BP 算法

在 LDA 模型中，BP 算法不直接求后验分布 $p(z|w)$，而是求边缘概率 $p(z_{w,d})$，即消息 $\mu(z_{w,d})$. 消息 $\mu(z_{w,d})$ 等于邻居系统的消息，即

$$\mu(z_{w,d}) \propto \mu_{\theta_d\to z_{w,d}}(z_{w,d}) \times \mu_{\phi_w\to z_{w,d}}(z_{w,d}) \quad (1)$$

其中箭头方向为消息传递方向. 为了描述简单，下面均用 $\mu(z_{\bar{w},d})=\sum_{v\neq w}\mu(z_{v,d})$，$\mu(z_{\cdot,d})=\sum_{w}\mu(z_{w,d})$，$\mu(z_{w,\bar{d}})=\sum_{s\neq d}\mu(z_{w,s})$，$\mu(z_{w,\cdot})=\sum_{d}\mu(z_{w,d})$ 来代替. 因子传递给变量的消息是所有邻居变量传入消息的叠加，即

$$\mu_{\theta_d\to z_{w,d}}(z_{w,d}) \propto f_d\mu(z_{\bar{w},d}=k) \quad (2)$$

$$\mu_{\phi_w\to z_{w,d}}(z_{w,d}) \propto f_w\mu(z_{w,\bar{d}}=k) \quad (3)$$

基于马尔可夫主题平滑先验，本文设因子函数为

$$f_d=\frac{1}{\sum_k[\mu(z_{\bar{w},d}=k)+\alpha]} \quad (4)$$

$$f_w=\frac{1}{\sum_w[\mu(z_{w,\bar{d}}=k)+\beta]} \quad (5)$$

为了便于文本间的可比性，等式(4)用文本 d 所有主题的消息归一化了传入消息. 同理，为了单词间的可比性，等式(5)用单词表中所有单词归一化了传入消息上. 因此消息更新等式可写为

$$\mu(z_{w,d}=k) \propto \frac{\mu(z_{\bar{w},d}=k)+\alpha}{\sum_k[\mu(z_{\bar{w},d}=k)+\alpha]} \times \frac{\mu(z_{w,\bar{d}}=k)+\beta}{\sum_w[\mu(z_{w,\bar{d}}=k)+\beta]} \quad (6)$$

其中 $\sum_k\mu(z_{w,d}=k)$ 实际为 $\sum_{z_{w,d}=k}\mu(z_{w,d}=k)$，为了简洁性，用 $\sum_k\mu(z_{w,d}=k)$ 来表示. 对更新的消息归一化，即 $\sum_k\mu(z_{w,d}=k)=1$. 然后更新参数 θ_d 和 ϕ_w，直到最大循环次数：

$$\theta_d \propto \frac{\mu(z_{\cdot,d}=k)+\alpha}{\sum_k[\mu(z_{\cdot,d}=k)+\alpha]} \quad (7)$$

$$\phi_w \propto \frac{\mu(z_{w,\cdot}=k)+\beta}{\sum_w[\mu(z_{w,\cdot}=k)+\beta]} \quad (8)$$

3.3 LDA 模型在线消息传递算法

鉴于 LDA 模型中离线算法内存空间随数据集大小的增长而不断增加，不能用于处理海量数据集. 仿照 OGS 和 OVB 在线算法构建思路，我们提出了在线 OBP 算法来估计 LDA 模型中的参数. 图 3 给出了在线学习的主要思想. OBP 算法将整个数据集切分成一系列小段，对于第 1 段数据集，OBP 算法和离线 BP 算法相同，从第 2 段到最后一段，OBP 算法先固定前一段的参数 $\phi_w(k)$，然后计算当前段的消息. 当 OBP 算法收敛或达到最大迭代次数时更新参数 $\phi_w(k)$. 根据在线随机优化理论，文中权重函数选用指数函数形式：

$$\rho_t=(\tau_0+t)^{-\kappa} \quad (9)$$

并且当前段及已训练段结果分别设置权重为 ρ_t 和

$1-\rho_t$，其中参数 $\kappa\in(0.5,1]$控制已经处理过的数据集，参数 $\tau_0\geq 0$ 用于减小每段开始迭代时的影响. 图 3 中当 $S=D$ 且 $\kappa=0$ 时，在线消息传递算法即转化为离线消息传递算法.

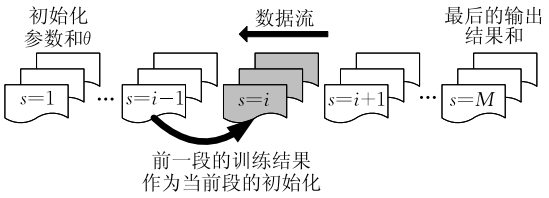


图 3　在线学习算法的流程

在训练中，首先在 0 与 1 之间随机初始化第 1 段参数 $\phi_w(k)$、$\mu_{w,d}(k)$和 $\theta_d(k)$，为了简洁，记 $\mu(z_{w,d}=k)=\mu_{w,d}(k)$训练结束后保存参数 $\phi_w(k)$. 从第 2 段到最后一段，只需随机初始化 $\theta_d(k)$参数，固定参数 $\phi_w(k)$，更新消息直到收敛. 在该算法中，根据 $\theta_d(k)$的差值来判断当前段是否收敛，也可以采用等价的 $\mu_{w,d}$ 的差值来判断，因为在更新参数 $\mu_{w,d}$时是固定参数 $\phi_w(k)$. 基于收敛后的消息，估计参数 $\phi_w(k)^{\text{new}}\propto(\mu_{w,\cdot}(k)+\beta)\Big/\sum_w(\mu_{w,\cdot}(k)+\beta)$. 而对于训练结果参数 $\phi_w(k)$，取当前段和已训练段的权重和：

$$\phi_w(k)=(1-\rho_t)\phi_w(k)+\rho_t\phi_w(k)^{\text{new}} \tag{10}$$

从式(10)可以看出，参数 $\phi_w(k)$是对当前段及所有已经训练段的结果进行权重加和，因此，距离当前段越远的数据段被乘了多重权重因子，对当前数据段的影响也越小；相反距离当前段越近的数据段对当前段的影响也就越大. 所以当处理数据内容信息分布不一致的实时数据流时，也能根据相邻数据段的内容比较快速的给出正确的主题信息.

OBP 算法的时间和空间复杂性相比 OGS 和 OVB 算法而言都是最小的. OBP 算法在每次迭代过程中只计算各个单词间的消息传递，OGS 算法却要对所有文本中的每个单词计算，因此 OBP 算法、OGS 算法和 OVB 算法的时间复杂度分别为 $O(KDW_DT)$、$O(KDN_DT)$和 $O(KDW_DT)$，其中 K 是主题数，D 是当前段的文本数，W_D是单词表大小，而 N_D是每篇文本的单词数，T 是迭代收敛次数. 尽管 OVB 算法和 OBP 算法的时间复杂度相同，但 OVB 算法因引入了非常耗时的 digamma 函数，基于文本的稀疏性 W_D通常远小于 N_D，所以每次迭代时间，OBP 算法少于 OGS 和 OVB 算法. 此外，OGS、OVB 和 OBP 算法的空间复杂度分别是 $O(KN_D+KS)$、$O(KN_D+KS)$和 $O(KW_D+KS)$，其中 S 是每段数据集中的文本数. 因此 OBP 算法相对 OGS 和 OVB 的空间复杂度也是最小的. OBP 算法见算法 1.

算法 1.　在线消息传递算法.

输入：ϕ_w，θ_d和 $\mu_{w,d}$

输出：ϕ_w和 θ_d

定义 $\rho_t=(\tau_0+t)^{-\kappa}$

随机初始化并归一化第 1 段参数 ϕ_w和 $\mu_{w,d}$

For $t=0$ to ∞ do

初始化当前段 θ_d

repeat

$$\mu_{w,d}^{t+1}(k)\propto\frac{\mu_{\bar{w},d}^{t}(k)+\alpha}{\sum_k\mu_{\bar{w},d}^{t}(k)+\alpha}\times\frac{\mu_{w,\bar{d}}^{t}(k)+\beta}{\sum_w\mu_{w,\bar{d}}^{t}(k)+\beta},$$

$$\theta_d(k)^{t+1}\propto(\mu_{\cdot,d}^{t+1}(k)+\alpha)\Big/\sum_k(\mu_{\cdot,d}^{t+1}(k)+\alpha)$$

until $\frac{1}{K}\sum_K|\text{changein }\theta_d|<0.000\,01$

计算当前段数据集

$$\phi_w(k)^{\text{new}}\propto(\mu_{w,\cdot}(k)+\beta)\Big/\sum_w(\mu_{w,\cdot}(k)+\beta),$$

$$\phi_w(k)=(1-\rho_t)\phi_w(k)+\rho_t\phi_w(k)^{\text{new}}$$

End for

3.4　LDA 模型的收敛性分析

给定文本数据集 $w=(w_1,\cdots,w_N)$，LDA 模型推断文本对应主题分布 θ、主题对应单词表概率分布 ϕ 和单词所属隐藏主题变量 $z=(z_1,\cdots,z_N)$分布. 算法 1 则可以收敛到一个稳定值，下面给出证明.

LDA 模型的目标函数为

$$\begin{aligned}L(w,\phi,\theta,\mu)&=\prod_d p(\theta_d|\alpha)\prod_k p(\phi_k|\beta)\cdot\\&\quad\prod_w p(\mu_{w,d}|\theta_d)p(w_{d,w}|\mu_{d,w},\phi_k)\\&\propto\sum_d\Big[\log p(\theta_d|\alpha)+\sum_k\log p(\phi_k|\beta)+\\&\quad\sum_w(\log p(\mu_{w,d}|\theta_d)+\log p(w_{d,w}|\mu_{d,w},\phi_k))\Big]\\&=\sum_d\Bigg(\log\Gamma\theta_{d,k}-\log\Gamma\Big(\sum_k\theta_{d,k}\Big)+\\&\quad\sum_k(\alpha-\theta_{d,k})\log\frac{\theta_{d,k}}{\sum_k\theta_{d,k}}\Bigg)+\\&\quad\sum_k\Bigg(\log\Gamma\phi_{k,w}-\log\Gamma\Big(\sum_w\phi_{k,w}\Big)+\\&\quad\sum_w(\beta-\phi_{k,w})\log\frac{\phi_{k,w}}{\sum_w\phi_{k,w}}\Bigg)\Big/D+\\&\quad\sum_w n_{d,w}\sum_k\mu_{d,w,k}\left(\log\frac{\theta_{d,k}}{\sum_k\theta_{d,k}}+\log\frac{\phi_{k,w}}{\sum_w\phi_{k,w}}\right)\\&=\sum_d\ell(n_d,\mu_d,\theta_d,\phi)\end{aligned}\tag{11}$$

因此，用 $\mu(n_d,\phi)$代表 μ_d和 $\theta(n_d,\phi)$代表 θ_d计算 μ 和

θ. 最大化 $L(n,\phi)=\sum_d \ell(n_d,\mu(n_d,\phi),\theta(n_d,\phi),\phi)$ 则可以通过估计参数 ϕ 来实现.

在线 OBP 算法收敛性可用随机自然梯度下降的方法来分析. 在随机最优算法中最优化目标函数一般用梯度估计来完成. 首先定义不断采样文本函数 $s(n)=\frac{1}{D}\sum_{d=1}^{D} I(n=n_d)$, 当 $n=n_d$ 时 $I(n=n_d)=1$, 否则 $I(n=n_d)=0$. 因此似然目标函数可重写为

$$L(s,\phi)=DE_s[\ell(n,\mu(n,\phi),\theta(n,\phi),\phi)\mid\phi] \tag{12}$$

其中 ℓ 定义见等式(11). 给定 ϕ, 等式(11)的最大化, 可通过 $n_t\sim s$ 不断采样观测样本, $\mu_t=\mu(n_t,\phi)$, $\theta_t=\theta(n_t,\phi)$ 来实现, 因此更新 ϕ 参数为

$$\phi\leftarrow\phi+\rho_t D\nabla_\phi\ell(n_t,\mu_t,\theta_t,\phi) \tag{13}$$

其中权重 $\rho_t=(\tau_0+t)^{-\kappa}$. 对每篇文本 n_t, 固定参数 ϕ, 将 μ_t 和 θ_t 参数均看作随机变量, 则有 $E_s[D\nabla_\phi\ell(n_t,\mu_t,\theta_t,\phi)\mid\phi]=\nabla_\phi\sum_d\ell(n_d,\mu_d,\theta_d,\phi)$. 当 $\sum_{t=0}^{\infty}\rho_t=\infty$, 且 $\sum_{t=0}^{\infty}\rho_t^2<\infty$, 参数 ϕ 收敛, 并且梯度 $\nabla_\phi\sum_d\ell(n_d,\mu_d,\theta_d,\phi)$ 会收敛到 0[7], 因此 ϕ 将会收敛到某个稳定值. 等式(13)中只用了一阶梯度. 若对梯度乘以一个合适的正定矩阵 $\boldsymbol{H}$ 的逆, 可加速随机梯度算法, 常用正定矩阵 $\boldsymbol{H}$ 是目标函数的哈森矩阵[7].

$$\begin{aligned}\frac{\partial\ell(n_t,\mu_t,\theta_t,\phi)}{\partial\phi_{k,w}}&=\sum_{v=1}^{W}\frac{\partial\log\left(\frac{\phi_{k,v}}{\sum_v\phi_{k,v}}\right)}{\partial\phi_{k,w}}\cdot\\&\quad(-\phi_{k,w}/D+\beta/D+n_{t,v}\mu_{t,v,k})\\&=\sum_{v=1}^{W}-\frac{\partial^2\log\phi_k}{\partial\phi_{k,v}\partial\phi_{k,w}}\cdot\\&\quad(-\phi_{k,w}/D+\beta/D+n_{t,v}\mu_{t,v,k})\cdot\\&\quad\left(-\frac{\partial^2\log\phi_k}{\partial\phi_{k,v}\partial\phi_{k,w}}\right)^{-1}\frac{\partial\ell(n_t,\mu_t,\theta_t,\phi)}{\partial\phi_{k,w}}\\&=-\phi_{k,w}/D+\beta/D+n_{t,v}\mu_{t,v,k}\end{aligned} \tag{14}$$

对等式(13)乘以 $\rho_t D$ 再加上 ϕ, 便得到算法 1 中参数 ϕ 的更新等式.

4 实验结果与分析

实验采用 4 组海量数据集: 美国政坛领域关于政治博客 blog[24]、邮件 enron[25]、新闻摘要 nytimes[25] 和摘要 pubmed[25]. 4 个数据集大小如表 2 所示, 其中 D 为数据集总文本数, W 为数据集对应单词表的总单词数. 在训练前先打乱重排数据集, $Train$ 为训练文档数, $Test$ 为测试文档数. 所有实验迭代次数为 500, 实际迭代次数以模型收敛为止, 主题数均为 10~50, 步长为 10. 在 CPU 为两个 6 核、频率为 3.46 GHz 和内存 128 GB 的 Sun Fire X4270 M2 服务器下用 MATLAB[26] 和 MEX C++ 获得实验.

表 2 4 个数据集的大小

	blog	enron	nytimes	pubmed
D	5177	39861	300000	8200000
W	33574	28102	102660	141043
$Train$	4500	36000	250000	8000000
$Test$	677	3861	50000	200000

为验证 LDA 模型下 OBP 算法的高效性和准确性, 在 4 个数据集上比较了 OBP 和 OGS 及 OVB 算法的混淆度[5] 和训练消耗时间, OBP 算法在 PLSA 模型和 LDA 模型上的实验对比, 且给出了 nytimes 数据集上主题随训练数据集变化的演变图.

4.1 评估学习参数

在线学习算法的权重函数[12] 中引入了 3 个学习参数、控制已训练数据段被遗忘的缓慢程度 $\kappa\in(0.5,1]$ 参数、用于降低每段数据集起始迭代结果影响的常数 τ_0 ($\tau_0\geq0$), 和限制切分后每段数据集文本数参数 S. LDA 模型在线算法训练结果与 3 个学习参数的有效性密切相关, 通常对于 S 值的选取是在内存容量范围内越大越好; 若 $S=D$, 则在线算法等价于传统的离线算法. 表 3 给出了 enron 数据集上最佳参数值的选取和测试集预测混淆度值. 从表 3 中不同参数值组的实验分析可以看出, 当 $\kappa=0.6$ 和 $\tau_0=1$ 时, 测试混淆度值最小. 为了更好的对比 LDA 模型的不同近似推理算法, 必须在相同参数条件下进行, 为此固定 $\kappa=0.6$ 对不同参数值组实验, 对参数的选取进行了对比(表 4). 从表 4 中可以看出, 当 τ_0 越小, S 越大, 对应的预测混淆度值也越小. 因此, 本文所有实验均选用 $\kappa=0.6$ 和 $\tau_0=1$, 而 S 值的选取则根据具体数据集的大小确定.

表 3 enron 语料库不同参数对比

参数	κ	τ_0	S	混淆度
组 1	0.9	1024	4	5554.1
组 2	0.8	1024	16	5490.6
组 3	0.7	1024	64	3817.9
组 4	0.6	256	256	2936.2
组 5	0.6	64	1024	2353.6
组 6	0.6	1	4000	2021.5
组 7	0.6	1	10000	1967.0

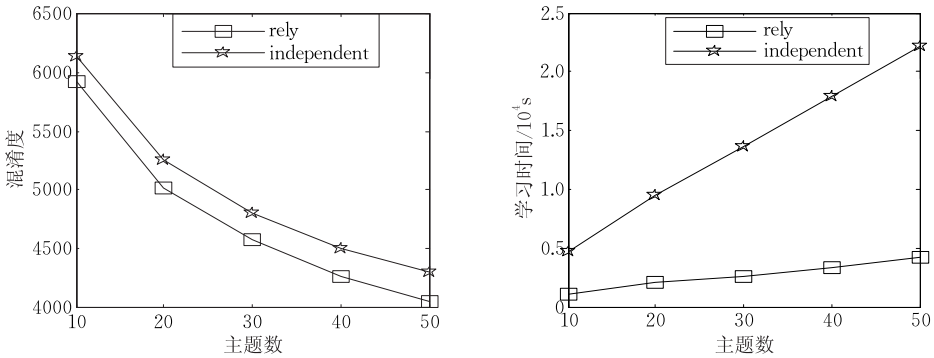
表 4　enron 语料库不同参数对比

参数	κ	τ_0	S	混淆度
组 1	0.6	1024	256	3045.2
组 2	0.6	256	256	2936.2
组 3	0.6	256	1024	2472.7
组 4	0.6	64	1024	2353.6
组 5	0.6	64	4000	2151.0
组 6	0.6	1	4000	2021.5
组 7	0.6	1	10000	1967.0

4.2　算法自身性能分析

在假设海量数据分段后，后段权重训练依赖前段训练结果的前提下，采用提出的 OBP 算法对海量数据进行训练. 为验证这一假设的准确性，采用混淆度和训练时间两个评价指标，在 nytimes 数据集上分别进行了权重依赖 rely 和完全独立 independent 的实验，实验结果如图 4 所示. 其中混淆度是评价用训练数据集训练所得到的结果来预测测试集的一个客观指标，值越小表明对未知测试集的预测能力越好.

从图 4 中可以看出，完全独立实验不仅模型预测集的混淆度值大，而且训练也更加耗时，若后段依赖前段训练结果，模型收敛所需迭代次数减少，相对耗时少. 实验表明，OBP 算法对海量数据集的分段假设在训练 LDA 模型上是可行的.



(a) 混淆度　　(b) 训练耗时

图 4　海量数据集前段训练结果对后段的影响

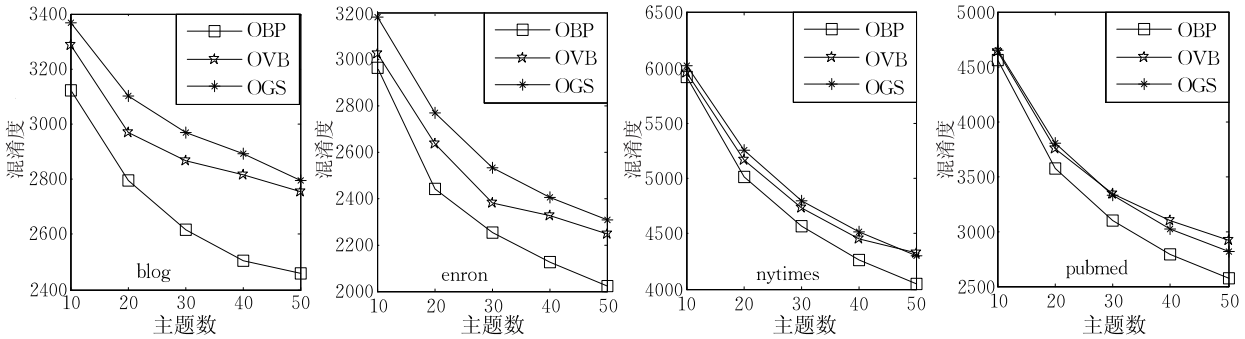
4.3　LDA 模型不同算法的对比分析

下面给出 LDA 模型下，OBP 与 OGS 及 OVB 算法在 4 个数据集 blog，enron，nytimes 和 pubmed 上混淆度和训练耗时的对比分析.

图 5 给出了 OBP、OGS 及 OVB 算法的混淆度对比分析结果. 在 4 个大规模数据集上，OBP 算法的混淆度均低于 OGS 和 OVB 算法，这说明相对于 OGS 和 OVB 算法，用 OBP 算法训练 LDA 模型，具有更好的预测性能. 图 6 也给出了 3 种在线算法训练时间的对比分析. 由于 digamma 函数的引入，OVB 算法训练非常耗时，图 6 中所给 OVB 算法的时间是其真实时间的 0.3 倍. OBP 算法相比 OGS 和 OVB 算法都要快速. 因此，在 LDA 模型中，OBP 算法相对于 OGS 和 OVB 算法更加的高效和准确.

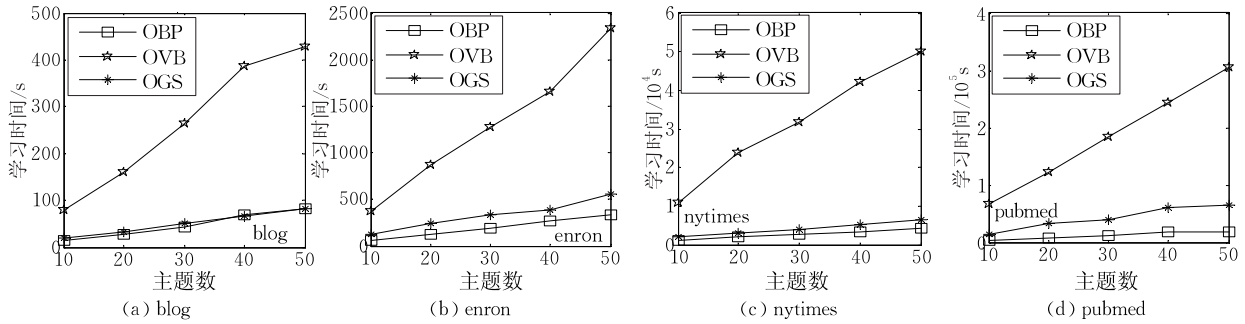
4.4　OBP 算法在 LDA 模型与 PLSA 模型上的对比分析

文献[23]中已经提出了基于 PLSA 模型因子图的 OBP 算法，并给出了算法的具体实现过程以及算法收敛条件. 从主题模型的角度，PLSA 模型中参数的个数会随训练文本数的增加而不断增加，从而导致在训练海量数据时模型中参数的个数较为庞大. LDA 模型是在 PLSA 模型的基础上提出来的，是一个层次贝叶斯模型，模型中将参数看作随机变量，并且引入了控制参数的参数，即语料库级超参，



(a) blog　(b) enron　(c) nytimes　(d) pubmed

图 5　主题上混淆度的对比

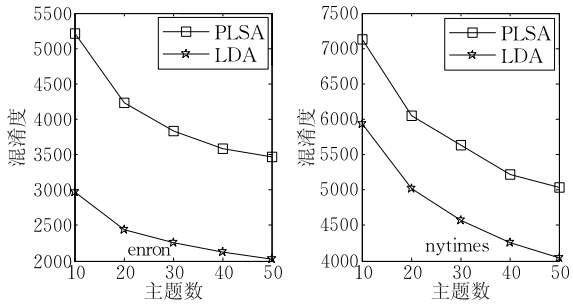


(a) blog (b) enron (c) nytimes (d) pubmed

图 6 主题上训练耗时的对比

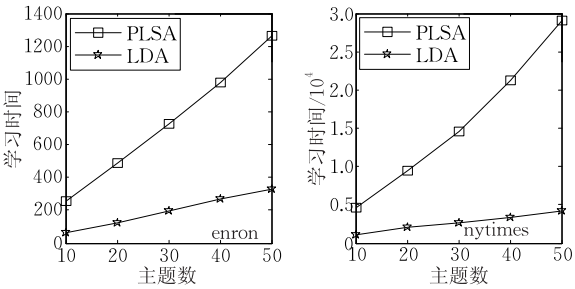
因此 LDA 模型对外表现出的参数始终只有两个超参,有效地减少了模型中参数个数. 图 7 和图 8 分别给出了 enron 和 nytimes 数据集在 PLSA 和 LDA 模型上训练混淆度和训练耗时对比.

从图 7 和图 8 可以明显的看出,在相同训练数据集上,LDA 模型不仅训练的混淆度值低于 PLSA 模型,而且训练所消耗的时间也远少于 PLSA 模型. 这就从实验上进一步验证了 LDA 模型下的 OBP 算法比 PLSA 模型的 OBP 算法预测的更加精确.



(a) enron (b) nytimes

图 7 主题上混淆度的对比



(a) enron (b) nytimes

图 8 主题上训练耗时的对比

4.5 主题追踪

主题追踪的目标是针对不断增长的数据流,追踪某个给定主题随时间的不断变化. 基本思路是,根据给定的训练文本,采用主题模型的近似推理算法对训练文本进行学习,得到每篇文本属于各个主题的概率以及各个主题对应单词表的概率分布. 当新的流数据到来时,按照主题模型已经训练的结果对新数据进行预测,一方面预测新数据中每篇文本属于各主题的概率值;另一方面同时更新各主题所对应单词表的概率分布.

为了更准确给出主题追踪信息,首先验证 OBP 算法在 LDA 模型下预测测试集的准确性. 图 9 给出了在 nytimes 数据集上,OBP 算法和 OGS 及 OVB 算法对测试数据集预测的混淆度对比分析图,实验选取 nytimes 数据集的前 25 000 篇文本作为训练样本集,后 50 000 篇文本作为测试样本集. 训练前先将训练集切分成 10 段视为数据流. 由于目前给定标签的语料库都是小规模的数据集,而且仅有很少的语料库会给定标签,本文采用的 4 组数据集均未给定标签,主题个数的最佳取值未知,所以文中在进行主题追踪时,主题个数选定为 $K=50$. 图 9 给出了 nytimes 数据集在 OGS,OVB 和 OBP 三种数据集上训练数据流数据对应的预测混淆度值对比,其中横坐标表示当前已经训练过的总文本数,纵坐标是当前训练样本对应的测试混淆度. 从图 9 中可看出,在整个训练过程中,OBP 算法的混淆度值均小于 OGS 和 OVB 算法,而且在最终收敛时,OBP 算法的混淆度值也远小于 OGS 和 OVB 算法,这说明在同一模型和相同训练样本集下,OBP 算法对未知测试集的预测能力最好. 此外,OBP 算法随训练样本数的不断增加,混淆度值降低越迅速,表明收敛速度

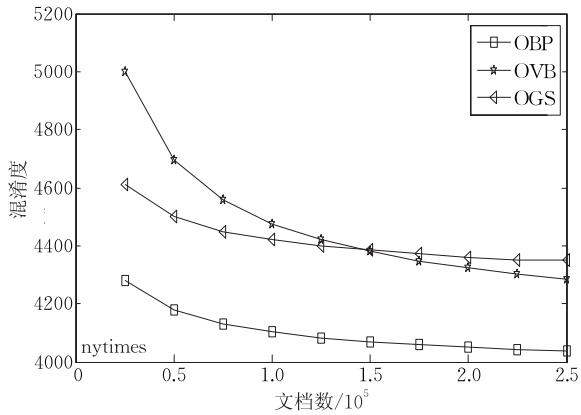


图 9 nytimes 数据集在OBP与OGS及OVB算法的混淆度对比

也更快.实验表明,基于 LDA 模型,OBP 算法预测准确性随处理数据流数据的增加而不断上升,即在追踪某个主题时,能更准确的给出与该主题相关的信息.

由于数据获取的限制,图 9 给出的实验是将海量数据切分后视为数据流处理,而并不是真的数据流,对同一个数据集,其内部可能是服从一致分布,但对流数据,可能其服从的分布会随时间不断变化.本文的在线算法是对当前段和已经得到的结果取权重叠加,而权重均是(0,1)之间的小数,所以对任意段数据,对其影响最大的是其相邻段,而距当前段很远的数据段由于经过若干段权重的相乘,影响就很小.所以,若处理的数据存在不一致分布,则不一致的前几段训练结果可能不准确,但是若干段之后又能准确预测.

为了验证在 LDA 模型下 OBP 算法能应用于主题追踪,表 5 给出了模型在 enron 语料库前 9 个数据段训练后对应主题的变化,由于空间的限制,表 5 中仅给出了第 10,20,30 和 40 个主题在第 1,第 5 和第 9 段数据集训练之后所包含的单词.实验是选取 enron 语料库前 36000 篇文本,并将其均分成 9 段数据集视做数据流处理,且模型训练时主题的个数选定为 50.根据表 5 中列出的主题在训练若干段数据流之后的不断变化,可以看出每个主题所包含的单词会随时间不断的变化,但始终是围绕当前主题的主旨,如第 10 个主题,在训练完第 1 个数据段后包含 office,interview 等与工作相关的词,训练第 5 段和第 9 段后的主题包含的单词也主要是围绕 business 浮动.因此 LDA 模型下的 OBP 算法可以用于处理数据流数据,并追踪主题的不断变化.

表 5　OBP 算法在 enron 数据流上的主题变化

第 10 个主题:business	
1:	ena office group london process interview sally analyst role
5:	team group process office ena sally business support forward business_unit
9:	group team office process business support ena management program forward
第 20 个主题:employee	
1:	bass eric dad respond weekend lum think dpr Friday dinner
5:	bass employees dpr respond eric larry weekend Friday think dad
9:	respond employees Friday bass eric dpr weekend think open floor
第 30 个主题:account	
1:	account access password user online page oasis statement portfolio visit
5:	page account access service user customer online password site statement
9:	access account page customer service user online site statement schwab
第 40 个主题:program	
1:	agreement attached cost section plan loan program provide copy number
5:	agreement plan provide employees attached issues section cost program required
9:	agreement plan provide issues proposal cost program direct issue section

在处理流数据时,LDA 模型对每个主题会随流数据训练样本的变化而变化.实验中的训练数据是将 nytimes 数据集切分成 10 段视为数据流,每段 25000篇文本.图 10 和图 11 给出了 OBP 算法的 LDA 模型训练流数据时,50 个主题的演变图.图 10 给出了前 25 个主题演变图,图 11 给出了后 25 个主题演变图.

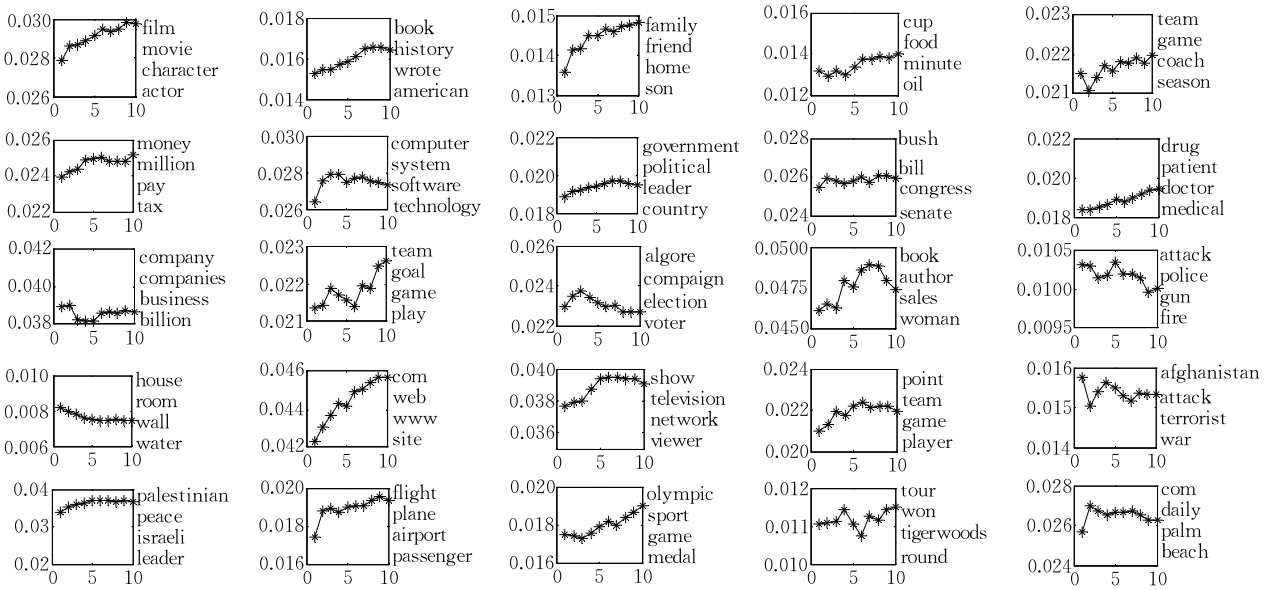


图 10　nytimes 数据集在 OBP 算法上前 25 个主题演变过程(横坐标为当前训练集所处理的训练段数;纵坐标为各主题在当前数据段所对应的概率值;右侧为当前主题中概率值较大的部分单词)

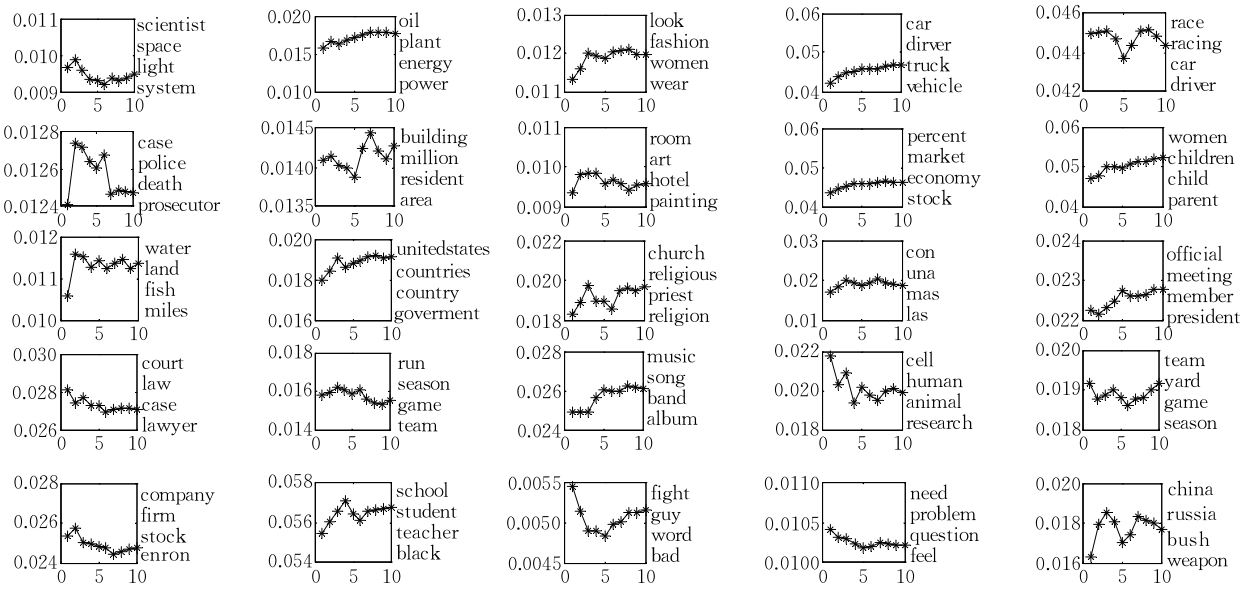


图 11 nytimes 数据集在 OBP 算法上后 25 个主题的演变过程(横坐标为当前训练集所处理的训练段数;纵坐标为各主题在当前数据段所对应的概率值;右侧为当前主题中概率值较大的部分单词)

从图 10 和图 11 可以看出,主题随训练数据集的增加而不断变化,如图 10 中的第 1 个主题演变图,其概率值随训练文本的增多而不断变大,表明该主题得到的关注度正在持续上升;图 10 的第 13 个主题对应的概率值随训练文本的增加而先变大后变小,表明该主题被研究的热度不断下降;图 10 的第 9 个主题随训练文本的增加,概率值变化比较缓慢,表明该主题被关注的程度几乎保持不变. 每个主题演变图反映了主题随训练样本不断增加的变化,结合所有主题图还可以挖掘出哪些主题是当前的热门话题. 根据主题演变图能够直接反映出各个主题随时间的变化趋势,追踪与各主题相关的数据信息,所以搜索引擎等可以借助该模型通过对历史数据的不断训练,从而给用户提供更加准确的搜索结果.

鉴于本文主题模型上的主题定义为单词表上的概率分布,即每个主题是由单词表中所有单词的不同排列构成,而图 10 和图 11 的每个主题右侧列出了该主题对应概率值较大的部分单词,所以对每个主题会存在相同的单词.

5 结 论

本文基于 LDA 模型的因子图提出了在线消息传递(OBP)算法,通过实验验证了 OBP 算法比 OGS 和 OVB 算法有显著提高,且 LDA 模型下的 OBP 算法优越于 PLSA 模型下的 OBP 算法,并将其应用到主题追踪上,获取了更准确的信息.

参 考 文 献

[1] Zeng Jia, Yan Jian-Feng, Gong Sheng-Rong. Advances in topic models for complex document network data. Chinese Journal of Computers, 2012, 35(12): 2431-2445(in Chinese)
(曾嘉, 严建峰, 龚声蓉. 复杂文本网中的主题建模进展. 计算机学报, 2012, 35(12): 2431-2445)

[2] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communications Association for Computing Machinery, 1975, 18(11): 613-620

[3] Landauer T K, Foltz P W, Laham F. An introduction to latent semantic analysis. Discourse Processes, 1998, 25: 259-284

[4] Hofmann T. Probabilistic latent semantic analysis//Proceedings of the 15th Conference on Uncertainty in Artificial (UAI'99). Stockholm, Sweden, 1999: 289-296

[5] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[6] Griffiths T L, Steyvers M. Finding scientific topics. Proceedings of the National Academy Sciences, 2004, 101(Suppl. 1): 5228-5235

[7] Zeng Jia, Cheung W K-W, Liu Ji-Ming. Learning topic models by belief propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(5): 1121-1134

[8] Bottou L. Online learning and stochastic approximations// Saad D ed. On-Line Learning in Neural Networks. New York, USA: Cambridge University Press, 2009: 9-42

[9] Xu J, Ye G, Wang Y, et al. Online learning for PLSA-based visual recognition//Proceedings of the 10th Asian Conference on Computer Vision. Queenstown, New Zealand, 2010, Part II: 95-108

[10] AlSumait L, Barbara D, Domeniconi C. On-line LDA: Adaptive topic models for mining text stream with applications to topic detection and tracking//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 3-12

[11] Yao Li-Min, Mimno D, McCallum A. Efficient methods for topic model inference on streaming document collections//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09). Paris, France, 2009: 937-946

[12] Hoffman M, Blei D M, Bach F. Online learning for latent dirichlet allocation. Advances in Neural Information Processing Systems, 2010, 23: 856-864

[13] Wang Chong, Palsley J, Blei D M. Online variational inference for the hierarchical dirichlet process//Proceeding of the 14th International Conference on Artificial Intelligence and Statistics. Ft Lauderdale, Florida, USA, 2011, 15: 752-760

[14] Banerjee A, Basu S. Topic models over text streams: A study of batch and online unsupervised learning//Proceedings of the 7th Society for Industrial and Applied Mathematics International Conference on Data Mining. Minneapolis, Minnesota, USA, 2007: 437-442

[15] Nair V, Clark J J. An unsupervised, online learning framework for moving object detection//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA, 2004, 2: II-317-II-324

[16] Pham M-T, Cham T-J. Online learning asymmetric boosted classifiers for object detection//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[17] Shai S-S, Singer Y, Ng A Y. Online and batch learning of pseudo-metrics//Proceedings of the 21st International Conference on Machine Learning. Banff, Alberta, Canada, 2004: 743-750

[18] Mairal J, Bach F, Ponce J, Spairo G. Online learning for matrix factorization and sparse coding. The Journal of Machine Learning Research, 2010, 11(3):19-60

[19] Vijayakumar S, D'Souza A, Schaal S. Incremental online learning in high dimension. Neural Computation, 2005, 17(12): 2602-2634

[20] Kivinen J, Smola A J, Williamson R C. Online learning with kernel. IEEE Transactions on Signal Processing, 2004, 52(8): 2165-2176

[21] Canini R K, Lei S, Griffiths T L. Online inference of topics with latent dirichlet allocation//Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS). Clearwater Beach, Florida, USA, 2009: 65-72

[22] Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06). New York, USA, 2006: 424-433

[23] Ye Yun, Gong Shen-Rong, Liu Chun-Ping, et al. Online belief propagation algorithm for probabilistic latent semantic analysis. Frontiers of Computer Science, 2013, 7(4): 526-535

[24] Eisenstein J, Xing E. The CMU 2008 political blog corpus. Carnegie Mellon University, Pittsburgh, PA, USA: Technical Report CMU-ML-10-101, 2010

[25] Porteous I, Newman D, Ihler A, et al. Fast collapsed Gibbs sampling for latent dirichlet allocation//Prodeedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 569-577

[26] Zeng Jia. A topic modeling toolbox using belief propagation. Journal of Machine Learning Research, 2012, 13: 2233-2236

GONG Sheng-Rong, born in 1966, Ph. D., professor, Ph. D. supervisor. His research interests include computer vision, machine learning etc.

YE Yun, born in 1988, M. S. Her research interest is machine learning.

LIU Chun-Ping, born in 1971, Ph. D., professor. Her research interests include image and video process, pattern recognition, computer vision and machine learning etc.

JI Yi, born in 1973, Ph. D., associate professor. Her research interests are computer vision and machine learning.

Background

Topic model can be used to summarize and classify documents, such as portal website, e-commerce, the social network, BBS, blog, microblog and so on. Compared with other extraction methods of document, topic models can obtain the more accurate result of classification, meanwhile, can predict and track more important and promising topics from the big data sets. Recently, online learning is regarded as an important algorithm for dealing with massive data set in text mining, computer vision, etc.

This paper proposes Online Belief Propagation (OBP) algorithm based on improved factor graphic of Latent Dirichlet Allocation (LDA) model to obtain more accurate results of topic tracking. LDA model is the classic topic model, which based on probabilistic Latent Semantic analysis

(PLSA). Existing LDA model with OGS or OVB algorithms cannot efficiently deal with massive dataset and result in the problem of low precision and speed. Moreover offline BP algorithm outperforms the other offline algorithms, such as GS and VB. To further improve the precision and speed of online algorithms using LDA models, this paper proposes OBP algorithm using message propagation. The experimental results of four public big data sets prove that proposed OBP is better than other online algorithms, which produces more precision information of topic tracking from the related information.

This work is supported by National Natural Science Foundation of China (Grant Nos. 61272258, 61170124, 61170020, 61301299). The significance of this paper depends on two aspects. First, this paper reviews the current online algorithms for topic models, and proposes high-precision online OBP algorithm using improved factor graphical LDA model. Second, the proposed algorithm provides a more accurate topic tracking, which may inspire some related machine learning applications.

Our research group focuses on the method and application of pattern recognition, machine learning, image and video process and computer vision. Several papers are published. For example, we published "Advances in Topic Models for Complex Document Network Data" in Chinese Journal of Computers. We also published "Online Belief Propagation Algorithm for Probability Latent Semantic Analysis" in Frontiers of Computer Science. This paper summarizes and extends our current research, which gives a solid foundation for the future research and applications in text/image/video classification and topic tracking.