基于动态权重的LDA算法

居亚亚 杨 璐 严建峰
(苏州大学计算机科学与技术学院 江苏 苏州 215006)

摘 要 潜在狄利克雷分布(Latent Dirichlet Allocation,LDA)是一种流行的三层概率主题模型,其实现了文本与文本中的单词在主题层次上的聚类。该模型以词袋(Bag of Words,BOW)模型为假设,所有单词的重要性相同,简化了建模的复杂度,但使得主题分布倾向于高频词,影响了主题模型的语义连贯性。针对此问题,提出了一种基于动态权重的LDA算法,该算法的基本思想是每个单词在建模中具有不同的重要性,在迭代过程中根据单词的主题分布动态生成相应的权重并反作用于主题建模,降低了高频词对建模的影响,提高了关键词的重要性。在4个公开数据集上的实验表明,基于动态权重的LDA算法在主题语义连贯性、文本分类准确率、泛化性能和精度方面比目前流行的LDA推理算法表现得更加优越。

关键词 潜在狄利克雷分布,主题模型,动态权重
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.043

LDA Algorithm Based on Dynamic Weight

JU Ya-ya YANG Lu YAN Jian-feng
(School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China)

Abstract The latent Dirichlet allocation (LDA)is a popular three-layer probability topic model,which implements the clustering of words in document and document at the topic level. This model is based on the Bag of Words(BOW) model,and each word has the same importance. It simplifies the complexity of modeling,but makes the topic distributions tend to high-frequency words,which affects the semantic coherence of the topic model. To achieve this goal,an LDA algorithm based on dynamic weight was proposed. The fundamental idea of the algorithm is that each word has different importance. In the iterative process of modeling,word weights are generated dynamically according to the topic distribution of words and feedback to topic modeling,reducing the influence of high frequency words and improving the role of keywords. Experiments on four public datasets show that the LDA algorithm based on dynamic weight can be superior to the current popular LDA inference algorithms in terms of topic semantic coherence,text classification accuracy,generalization performance and precision.

Keywords Latent dirichlet allocation,Topic model,Dynamic weight

1 引言

随着信息技术的迅速发展,网络中以文本形式呈现的信息增长迅猛。如何有效地挖掘其中隐含的信息,正是人们当前所面临的一大挑战。在此背景下,各种文本挖掘模型被相继提出,包括文档表示模型(Term Frequency-Inverse Document Frequency,TF-IDF)[1]、潜在语义索引模型(Latent Semantic Index,LSI)[2]、概率潜在语义索引模型(probabilistic PLSI)[3-4]和潜在狄利克雷分布(LDA)[5]。其中,LDA是一种处理非结构化文档集合的有效工具,被广泛应用于文本分类[6]、信息检索[7]等任务。参数估计是LDA模型的核心,其中最常用的3种推理算法是变分贝叶斯(Variational Bayes,VB)[5]、吉布斯采样(Gibbs Sampling,GS)[8]和期望最大化算法(Expectation Maximization,EM)[9-10]。这3种推理算法的优化目标差异较大,其中的EM算法是直接优化后验概率以寻找最优拟合数据集的参数,因此在泛化性能和精度上都明显优于VB算法和GS算法[10]。

目前流行的LDA算法在主题建模过程中没有较好地结合相关的语义信息,这严重影响了主题的语义连贯性、可解释性[11]和文本语义表征的准确性[5]。针对这种现象,通常有两种解决方法。1)针对特定的任务,在建模过程中加入合适的外部先验知识[12]。文献[13]提出了一种基于单词共现的熵加权策略以获得解释性更强的主题,但如何有效地获取符合建模的、正确的外部先验知识始终是一大挑战。2)在建模初始化前使用一些统计方法对数据集中的单词进行处理。文献[14]和文献[15]分别使用PMI和TF-IDF作为单词的权重,文献[16]通过单词间共现的关系和主题间相似的关系获得单词权重并将其融入到主题建模中。这些统计方法只是简单地

到稿日期:2018-07-14 返修日期:2018-10-29 本文受国家自然科学基金(61572339,61272449),江苏省科技支撑计划重点项目(BE2014005)资助。

居亚亚(1989-),女,硕士生,主要研究方向为机器学习,E-mail:yayaju@163.com;**杨 璐**(1982-),女,副教授,硕士生导师,主要研究方向为机器学习与软件工程,E-mail:yanglu@suda.edu.cn(通信作者);**严建峰**(1978-),男,副教授,硕士生导师,主要研究方向为机器学习。

对数据集中单词出现的频率进行统计，并没有考虑单词的语义信息对单词重要性的影响。

针对上述问题与挑战，本文研究了传统概率主题模型的语义强化问题，提出了一种基于动态权重的 LDA 算法，基于 EM 算法的框架，在模型迭代的过程中使用语义信息动态地获取单词的权重，使得建模产生的主题具有较高的语义连贯性。实验表明，在互信息指数、分类准确率和混淆度指标上，基于动态权重的 LDA 算法较目前流行的主题模型推理算法表现得更加优越。

本文第 2 节简要介绍了 LDA 模型和推理算法；第 3 节介绍了基于动态权重的 LDA 算法及其图模型和推理过程；第 4 节主要是在公开数据集上将所提算法与目前流行的推理算法进行了实验对比；最后总结全文，并对下一步工作进行展望。

2　相关工作

2.1　潜在狄利克雷分布

LDA[5] 模型是一种无监督的三层贝叶斯概率图模型。不考虑文档之间的顺序，以及文档中单词之间的顺序，如图 1 所示，模型假定整个文本集有 K 个主题，每篇文档 d 可以表示为长度为 K 的主题分布 θ_d，每个主题 k 可以表示为长度为词汇表长度 W 的单词分布 ϕ_k。一篇文档的生成过程如下：

$$\theta_d \sim Dir(\alpha), \phi_k \sim Dir(\beta), z_i \sim \theta_d, x_i \sim \phi_{z_i} \tag{1}$$

其中，假设 θ_d 和 ϕ_k 服从狄利克雷分布（Dir），其超参数分别为 α 和 β。LDA 的建模过程是逆向地通过文本集合生成模型，首先从先验参数为 β 的狄利克雷分布中获取每个主题 k 的分布 ϕ_k，对于一篇文档 d，从先验参数为 α 的狄利克雷分布中获取其主题分布的概率分布 θ_d，然后从 θ_d 中采样出文档 d 中每个单词 t 的主题 z_i，再从主题单词分布 ϕ_{z_i} 中获取 w。重复这样的过程，直到生成所有的文档为止。表 1 列出了本文模型与 LDA 模型相关的一些参数。

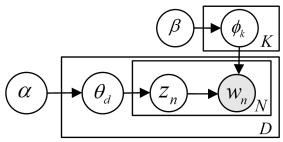


图 1　LDA 图模型

Fig. 1　Graphical model of LDA

表 1　符号定义

Table 1 Definition of notations

符号	意义
$1 \leqslant d \leqslant D$	语料库文本索引
$1 \leqslant w \leqslant W$	词汇表中单词的索引
$1 \leqslant k \leqslant K$	主题索引
$1 \leqslant t \leqslant T$	迭代次数
$x_{w,d}$	索引为 $\{w,d\}$ 的单词词频
x	所有 $x_{w,d}$ 的集合
NNZ	非零元素的个数
$z_{w,d}^k$	文本 d 中所有单词 w 属于主题 k 的个数
z	所有 $z_{w,d}^k$ 的集合
$\theta_d(k)$	文本 d 的主题分布中主题 k 的概率
$\phi_w(k)$	主题 k 的单词分布中单词 w 的概率
$\hat{\theta}_d(k)$	文本 d 的主题分布中主题 k 的概率计数
$\hat{\phi}_w(k)$	主题 k 的单词分布中单词 w 的概率计数
$\mu_{w,d}(k)$	文本 d 中单词 w 属于主题 k 的概率
α, β	狄利克雷分布的超参数

2.2　最大期望算法

主题模型的目标是为每篇文档中的每个单词预测一个主题标签 $z_{w,d}^k$。LDA 根据图模型定义了一个似然函数，并使用推理算法对该似然函数极大化，从而求得每篇文档的主题分布和每个主题的单词分布。EM 算法[9-10] 是一种最大后验（Maximum A Posterior，MAP）推理算法，推理的目标是极大化似然函数中的参数 θ 和 ϕ，如下所示：

$$p(x,\theta,\phi|\alpha,\beta) = \prod_{w,d,i}[p(x_{w,d,i}=1, z_{w,d,i}^k=1|\theta_d(k), \phi_w(k))] \times \prod_d p(\theta_d(k)|\alpha) \times \prod_k p(\phi_w(k)|\beta) \tag{2}$$

其中：

$$p(x_{w,d,i}=1, z_{w,d,i}^k=1|\theta_d(k),\phi_w(k)) = p(x_{w,d,i}=1|z_{w,d,i}^k=1,\phi_w(k)) \times p(z_{w,d,i}^k=1|\theta_d(k)) \tag{3}$$

将式(3)代入式(2)，并根据狄利克雷分布公式得到 EM 算法的极大似然函数，如下所示：

$$L(\theta,\phi) = \log p(x,\theta,\phi|\alpha,\beta) \propto \sum_{w,d,i} x_{w,d,i}[\log \sum_k \mu_{w,d}(k) \frac{\theta_d(k)\phi_w(k)}{\mu_{w,d}(k)}] \times \sum_d \sum_k \log[\theta_d(k)]^{\alpha-1} \times \sum_k \sum_w \log[\phi_w(k)]^{\beta-1} \tag{4}$$

其中，$\mu_{w,d}(k)$ 表示文本 d 中单词 w 属于主题 k 的概率，在 EM 算法中，$\mu_{w,d}(k)$ 是一个隐变量且满足 $\sum_k \mu_{w,d}(k)=1$，$\mu_{w,d}(k) \geqslant 0$。由于 $\sum_{w,d,i}[x_{w,d,i}=1] = \sum_{w,d} x_{w,d}$，因此可以消除式(4)中的单词编号 i，并使用杰森不等式（Jensen's Inequality）将式(4)转化为：

$$L(\theta,\phi) \geqslant L(\mu,\theta,\phi) = \sum_{w,d} \sum_k x_{w,d}\mu_{w,d}(k)[\log \sum_k \frac{\theta_d(k)\phi_w(k)}{\mu_{w,d}(k)}] \times \sum_d \sum_k \log[\theta_d(k)]^{\alpha-1} \times \sum_k \sum_w \log[\phi_w(k)]^{\beta-1} \tag{5}$$

对式(5)进行求导后，可得 EM 算法的 EM 框架。EM 算法迭代一次语料库共分为两步，依次为 E 步骤（E-step）和 M 步骤（M-step），算法流程如算法 1 所示。其中，E-step 更新的是隐变量 $\mu_{w,d}(k)$：

$$\mu_{w,d}(k) \propto \frac{[\hat{\theta}_d(k)+\alpha-1][\hat{\phi}_w(k)+\beta-1]}{\sum_w [\hat{\phi}_w(k)+\beta-1]} \tag{6}$$

而 M-step 则利用 E-step 中得到的 $\mu_{w,d}(k)$ 更新 $\theta_d(k)$ 和 $\phi_w(k)$ 的概率计数 $\hat{\theta}_d(k)$ 和 $\hat{\phi}_w(k)$：

$$\hat{\theta}_d(k) = \sum_w x_{w,d}\mu_{w,d}(k) \tag{7}$$

$$\hat{\phi}_w(k) = \sum_d x_{w,d}\mu_{w,d}(k) \tag{8}$$

算法 1　EM 算法

输入：x，K，T，α，β

输出：θ_d，ϕ_k

1. 随机地为每个单词 $x_{w,d}$ 分配主题，初始化和标准化 $\mu_{w,d}^1(k)$，并初始化 $\hat{\theta}_d(k)$ 和 $\hat{\phi}_w(k)$；
2. for t=1 to T //迭代循环，T 为循环次数
3. 　$\hat{\theta}_d^t(k) \leftarrow 0$，$\hat{\phi}_w^t(k) \leftarrow 0$
4. 　for $x_{w,d}$ in x：//遍历语料库中的每个单词
5. 　　for k in K //分别对每个主题进行更新
6. 　　使用式(6)更新 $\mu_{w,d}^t(k)$，使用式(7)和式(8)更新 $\hat{\theta}_d^t(k)$ 和 $\hat{\phi}_w^t(k)$；

7. $\hat{\theta}_d(k) \leftarrow \hat{\theta}_d^t(k), \hat{\phi}_w(k) \leftarrow \hat{\phi}_w^t(k)$

8. //更新概率分布 $\theta_d(k)$ 和 $\phi_w(k)$

$$\theta_d(k) \leftarrow \frac{\hat{\theta}_d(k)+\alpha-1}{\sum_k \hat{\theta}_d(k)+\alpha-1}, \phi_w(k) \leftarrow \frac{\hat{\phi}_w(k)+\beta-1}{\sum_w \hat{\phi}_w(k)+\beta-1}$$

由式(7)和式(8)可知，$\hat{\theta}_d(k)$ 和 $\hat{\phi}_w(k)$ 受单词词频 $x_{w,d}$ 的影响较大，因此在主题建模的迭代过程中，高频词比关键词或低频词更重要。主题的单词分布 ϕ_k 中高频词的概率值较高，这会使得高频词更加均匀地分布在几乎所有的主题上，严重影响了 LDA 模型产生的主题的语义连贯性和文档语义表征的准确性。针对以上问题，本文在 EM 框架下，提出了一种基于动态权重的 LDA 算法。

3 基于动态权重的 LDA 模型

传统 LDA 模型采用的是基于词袋(Bag of Words)模型的表示方法，将文本信息转化为易于建模的数学信息，将每一篇文档视为一个词频向量。这种方法使得语料库中的每个单词都具有同等的重要性，简化了建模的复杂度。由 Zipf's[17] 定律可知，只有极少数的词经常被使用，而绝大多数词很少被使用，因此语料库中的高频词被采样的概率远比关键词或低频词高，这将严重影响了 LDA 的建模效果。本文提出的 dwEM(Dynamic Weight Expectation Maximization)模型并不是同等地看待每个单词，而是以权重的形式来衡量每个单词在建模过程中的重要性，主要是通过动态地获得单词的权重来优化 LDA 模型的采样过程。这种方法并不需要从外部数据集中引入先验知识，可以避免不合理的先验知识对建模产生的误导，同时也没有使用一些统计方法来对单词的词频进行前期处理，避免了丢失文档中单词的语义信息。

dwEM 模型中文档的生成过程与传统的 LDA 模型中文档的生成过程相似，两者的不同在于基于动态权重的 LDA 模型在采样过程中动态地引入了权重信息。图 2 是基于动态权重的 LDA 模型的图模型。基于动态权重的 LDA 模型的主要思想是在模型迭代过程中动态地获取单词的语义信息，以不断更新语料库中单词的权重，并将权重融入到模型的下一次迭代中。权重的获取方法为：使用建模过程中生成的主题单词分布将每个单词表示成具有主题语义信息的向量，将单词的主题向量与噪音向量(式(10))之间的相似度作为每个单词的权重，相似度越高表示单词在建模中越不重要，相似度越低表示单词在建模中越重要。

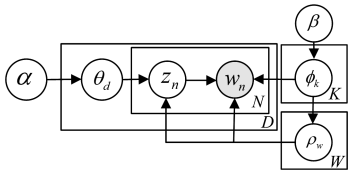


图 2 dwEM 图模型

Fig. 2 Graphical model of dwEM

3.1 dwEM 的实现

在 dwEM 中，利用在迭代过程中生成的主题单词分布 $\phi_w(k)$ 将单词 w 向量化为式(9)，$\boldsymbol{w}_{\text{topic}}$ 表示单词 w 的主题向量，它是一个归一化的 K 维向量，K 为主题个数。因为高频词在主题建模过程中几乎占据了所有的主题，所以其主题向量分布相对于关键词或低频词来说更倾向于均匀分布，本文将这个均匀分布表示成式(10)，其表示单词 w 在所有主题 k 上都有相同的主题概率 $1/K$，即每个主题都是同等重要的。这个均匀向量并没有实际的意义，被称为噪音向量。通过计算单词的主题向量与噪音向量之间的相似度，可以有效地区分出语料库中每个单词的重要性，并将其作用到模型的迭代过程中来对建模进行优化。

$$\boldsymbol{w}_{\text{topic}} = \frac{1}{\sum_{t=1}^{K} \phi_w(k)}[\phi_w(1), \phi_w(2), \cdots, \phi_w(K)] \tag{9}$$

$$\boldsymbol{w}_{\text{iniform}} = \frac{1}{K}[1,1,\cdots,1] \tag{10}$$

$$D_{\text{KL}}(P \parallel Q) = \sum_i P(i)\frac{P(i)}{Q(i)} \tag{11}$$

$$D_{\text{JS}}(P \parallel Q) = \frac{1}{2}[D_{\text{KL}}(P \parallel M) + D_{\text{KL}}(M \parallel P)] \tag{12}$$

其中，$M=\frac{1}{2}(P+Q)$。

相似度使用 KL 散度(Kullback-Leibler Divergence)[18] 来衡量。KL 散度描述的是在同一空间中两个概率分布之间的差异性，当两个分布相等时，其 KL 散度为 0，当两分布之间的差异性增大时，其 KL 散度也增大。概率分布 P 和 Q 的 KL 散度被表示为式(11)，由于 KL 散度是非对称的，不能表示真实的距离度量关系，因此本文使用 KL 散度的对称形式即 JS 散度(Jenson Shannon Divergence)，如式(12)所示。在主题建模过程中，将动态权重表示为单词的主题向量与噪音向量之间的 JS 散度，高频词与噪音向量之间的分布更加相似，即具有较小的 JS 散度，因此在迭代过程中赋予其较小的权重，以降低其在主题建模中的重要性；关键词和低频词与噪音向量之间的分布差异较大，因此在迭代过程中赋予其较大的权重，以增加其在主题建模中的重要性。式(13)表示单词 w 在迭代过程中的权重 ρ_w，将 JS 散度 δ_w 进行标准化，使得语料库中所有单词的总数不变。因此，在 dwEM 算法的迭代过程中，$\hat{\theta}_d(k)$ 和 $\hat{\phi}_w(k)$ 的更新公式分别如式(14)和式(15)所示：

$$\rho_w = \frac{\sum_{w,d} x_{w,d}}{\sum_{w,d} \delta_w * x_{w,d}}, \delta_w = D_{\text{JS}}(\boldsymbol{w}_{\text{topic}} \parallel \boldsymbol{w}_{\text{uniform}}) \tag{13}$$

$$\hat{\theta}_d(k) = \sum_w \rho_w x_{w,d} \mu_{w,d}(k) \tag{14}$$

$$\hat{\phi}_w(k) = \sum_d \rho_w x_{w,d} \mu_{w,d}(k) \tag{15}$$

3.2 dwEM 算法的流程

基于动态权重的 LDA 算法的具体流程如算法 2 所示。

算法 2 基于动态权重的 LDA 算法(dwEM 算法)

输入：x，K，T，α，β

输出：θ_d，ϕ_k

1. 随机地为每个单词 $x_{w,d}$ 分配主题，初始化以及标准化 $\mu_{w,d}^1(k)$，并初始化 $\hat{\theta}_d(k)$ 和 $\hat{\phi}_w(k)$；

2. 初始化 $\rho_w=[1,1,\cdots,1]$，$\mathbf{w}_{\text{iniform}}=\frac{1}{K}[1,1,\cdots,1]$；

3. for t=1 to T //迭代循环，T 为循环次数

4. $\hat{\theta}_d^t(k) \leftarrow 0, \hat{\phi}_w^t(k) \leftarrow 0$ //初始化概率计数

5. if t<bound

6. 执行 EM 算法

7. else

8. for $x_{w,d}$ in x //遍历语料库中的每个单词

9.　　for k in K //分别对每个主题进行更新

10.　　使用式(6)更新 $\mu_{w,d}^{t}(k)$，使用式(14)和式(15)更新 $\hat{\theta}_{d}^{t}(k)$和$\hat{\phi}_{w}^{t}(k)$；

11.　$\hat{\theta}_{d}(k)\leftarrow\hat{\theta}_{d}^{t}(k)$，$\hat{\phi}_{w}(k)\leftarrow\hat{\phi}_{w}^{t}(k)$

12.　if t% interval==0 do

13.　　使用式(9)获得单词的主题向量 $\mathbf{w}_{\text{topic}}$；

14.　　使用式(13)计算单词的权重 ρ_w；

15.　//更新概率分布 $\theta_d(k)$和 $\phi_w(k)$

$$\theta_d(k)\leftarrow\frac{\hat{\theta}_d(k)+\alpha-1}{\sum_k\hat{\theta}_d(k)+\alpha-1},\phi_w(k)\leftarrow\frac{\hat{\phi}_w(k)+\beta-1}{\sum_w\hat{\phi}_w(k)+\beta-1}$$

其中，*bound* 参数表示迭代下限，*interval* 表示动态更新单词权重的迭代间隔次数。当迭代次数 $1\leqslant t\leqslant bound$ 时，运行 EM 算法实现模型的初步收敛并获得主题单词分布；当 $bound\leqslant t\leqslant T$ 时，运行 dwEM 算法，且每次间隔为 *interval* 时动态地获取单词的权重并将其反作用于主题建模的下一次迭代过程中。

4　实验分析

4.1　实验环境和数据集

本实验是在单机多核服务器上进行的，该服务器由 2 个 CPU 组成，每个 CPU 有 8 个核，内存为 140 GB。

实验在 4 个公开数据集(Cora，WebKB，Reuters R8(R8)，20 Newsgroups (20 NG))中进行，文献[8]详细介绍了相关数据集，这 4 个数据集的相关描述如表 2 所列。

表 2　数据集

Table 2　Datasets

数据集	*D*	*W*	*NNZ*	*Category*
Cora	2410	2961	103699	7
WebKB	4168	7764	202995	4
R8	7674	22931	322973	8
20 NG	18821	92800	1549945	20

表 2 概括统计了实验所使用的 4 个数据集，其中 *D* 为文档个数，*W* 为单词表长度，*NNZ* 为数据集中非零元素的个数，*Category* 为数据集中文本的类别数目。在进行实验之前，先对数据集进行预处理，主要包括去除标准的停用词、出现次数小于 3 的单词，并对单词进行了词干化等。

在主题模型的研究和应用中，先验参数的选取对主题的建模产生了一定的影响，但由于先验参数的研究不是本文的重点，同时为了简化实验并保证算法比较的公平性，本实验直接参考文献[5]的参数设置，将所有算法的先验参数分别设置为 $\alpha=50/K$，$\beta=0.01$，其中 *K* 为主题个数。总迭代次数 $T=1000$，权重更新下界 $bound=20$，更新间隔 $interval=20$。

4.2　评价标准

本文对主题模型的建模能力进行了评估，主要使用的是模型产生的两个分布，即文档-主题分布和主题-单词分布。采用主题模型通用领域的性能评价指标即点互信息指数(Pointwise Mutual Information，PMI)[19-20]、分类准确率(Accuracy)[5]和混淆度(Perplexity)[6-21]来评价算法的性能。

点互信息是衡量主题语义连贯性的常用评价指标。将建模产生的每个主题中概率最高的 *N* 个单词的相关性作为 *PMI* 值，*PMI* 值越高，则该主题的语义连贯性越强。其中主题 *k* 的互信息指数的计算公式为：

$$PMI(k,W^k)=\frac{2}{N(N-1)}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\log\frac{Q(w_i^k,w_j^k)+\varepsilon}{Q(w_i^k)Q(w_j^k)}\tag{16}$$

其中，$Q(w)$表示语料库中包含单词 w 的文档数目，$Q(w_i^k,w_j^k)$表示包含单词$\{w_i,w_j\}$的文档数目，$W^k=(w_1^k,\cdots,w_N^k)$为主题 *k* 中概率最大的 *N* 个单词的列表，ε 是用来避免对数为 0 的一个小的正整数，本文设置 $N=10$，$\varepsilon=1$。

分类准确度是衡量文档语义表征能力的常用指标。将建模产生的主题作为文档的特征进行分类，本文以 6∶4 的比例随机将数据集划分为训练集和测试集，使用支持向量机(SVM)实现文本分类，分别进行 10 次实验，将其平均值作为准确率。经过实验验证，其他分类器的分类效果与其一致。分类准确率的计算公式为：

$$Acuuracy=\frac{1}{|C|}\sum_{i\in C}\frac{T_i}{D_i}\tag{17}$$

其中，$|C|$表示文本类别的数目，D_i 表示类别 *i* 中的文本数目，T_i 表示类别 *i* 中被分类正确的文本数目。

混淆度是一种信息理论的测量方法，常被用于统计语言模型中来衡量建模能力的好坏。其通过衡量单词在建模产生的文档主题分布 θ_d 和主题单词 ϕ_k 分布下的概率似然大小来评价建模的效果，越低的混淆度表示越好的泛化性能。混淆度的计算公式为：

$$Perp=\exp\left\{-\frac{\sum_{w,d}x_{w,d}\log[\sum_k\theta_d(k)\phi_w(k)]}{\sum_{w,d}x_{w,d}}\right\}\tag{18}$$

4.3　实验对比分析

4.3.1　语义连贯性分析

图 3 展示了在 4 个数据集上，dwEM 算法与目前流行的 3 种推理算法(VB[5]，GS[8]，EM[9-10])在不同主题数 *K* 下的 *PMI* 值，其中主题数 $K=\{20,40,60,\cdots,200\}$。从实验结果中可以看出，本文提出的 dwEM 算法的 *PMI* 值总体较高，表明其抽取出的主题具有较强的语义连贯性，使得主题下概率较高的词之间的相关性更强。

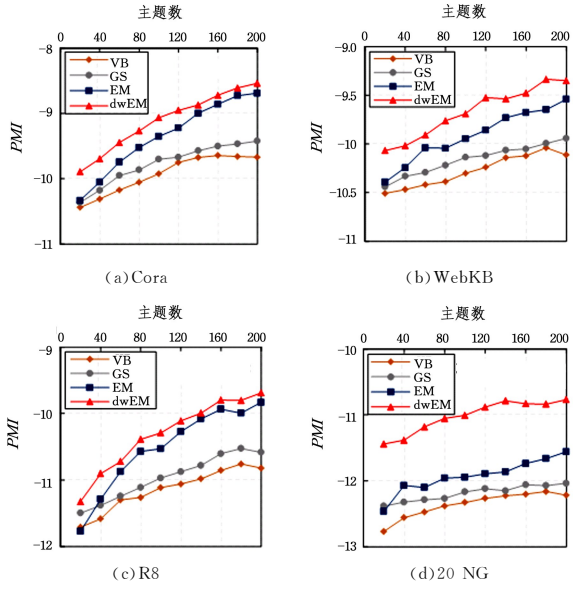


图 3　不同主题数下各算法 *PMI* 值的比较

Fig. 3　Comparison of *PMI* of different algorithms with different topic numbers

目前，在泛化性能与精度方面，EM 算法是最优的主题推理算法[5]，相比 EM 算法，dwEM 算法在数据集 Cora，

WebKB,R8 和 20 NG 上的 *PMI* 值分别平均提高了 2.6%,2.5%,2.0%和 7.4%。这说明在迭代过程中,dwEM 算法将单词的语义信息以权重的形式融合到主题建模的过程中,能够有效地提高主题的语义连贯性。

4.3.2 主题词效果的展示

为了更加直观地展现主题的语义连贯性,本文对模型产生的主题单词分布进行了分析,通过选取主题中概率值最高的 5 个单词来展现主题的可读性[22]。由于篇幅有限,本文仅展示了 GS 算法和 dwEM 算法在 20 NG 数据集上的部分主题的代表词和其中与主题无关的高频词(用粗体标记),其中 $K=50$。如表 3 所列,GS 算法获得的主题的可读性较差,从主题 2 中并不能看出主题所表达的主旨;此外,不难发现几乎所有的主题下都包含高频词,如"make""use"等。而本文提出的 dwEM 表现良好,出现高频词的情况明显减少,它将高频词替换为更符合主题语义的关键词汇,如将主题 1 中的高频词"use"替换为关键词汇"linux",则明显增强了主题的可读性和语义连贯性。

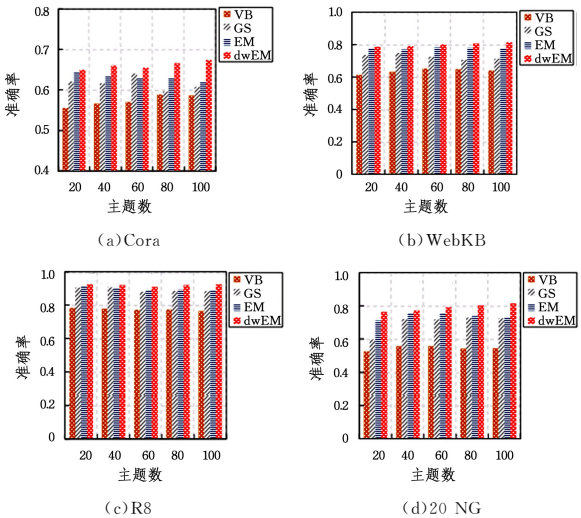
表 3 主题代表词汇的展示

Table 3 Display of representation vocabularies in topics

ID			GS		
1	windows	doc	**use**	microsoft	**make**
2	**people**	mean	point	thing	wrong
3	drug	**use**	**make**	vitamin	kidney
4	gun	**people**	public	control	**make**
5	food	eat	brain	effect	**use**
ID			dwEM		
1	windows	doc	linux	microsoft	vista
2	circuit	wire	audio	**use**	voltage
3	drug	harvard	vitamin	kidney	liver
4	gun	people	weapon	kill	crime
5	food	eat	brain	taste	drink

4.3.3 文本分类效果的分析

文本分类任务是一种评价主题整体建模能力的有效方法,文本分类准确率越高,表示通过主题模型获得的主题的特征表达能力越强。图 4 给出了 4 种 LDA 算法在不同主题 K 下的文本分类准确率。



(a)Cora (b)WebKB

(c)R8 (d)20 NG

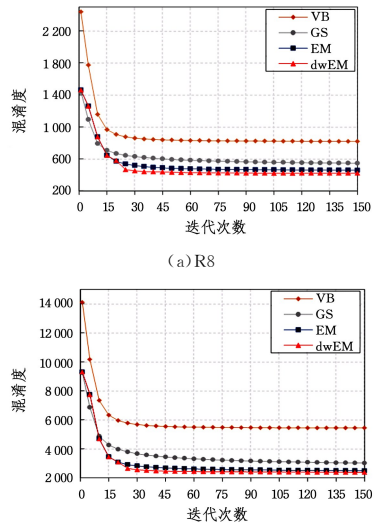
图 4 不同主题数下各算法分类准确率的比较

Fig. 4 Comparison of classification accuracy of different algorithms with different topic numbers

由图 4 可知,在 4 个数据集上,本文提出的 dwEM 模型的文本分类准确率均最高:在主题 $K=100$ 的情况下,EM 算法和 dwEM 算法的分类准确率在 Cora 数据集上分别为 0.619 和 0.667,在 WebKB 数据集上分别为 0.775 和 0.807,在 R8 数据集上分别为 0.887 和 0.925,在 20 NG 数据集上分别为 0.738 和 0.8141。这说明基于动态权重的 dwEM 主题模型在文档语义表征方面具有较强的刻画能力。

4.3.4 算法收敛性的分析

收敛性是一种评价模型训练速度的常用指标,图 5 给出了 4 种 LDA 算法在数据集 R8 和 20 NG 上的混淆度随迭代次数增加的变化情况。由图 5 可知,相比于其他 3 种算法,dwEM 算法在最终混淆度方面存在着明显的优势,即具有较高的泛化性能和精度,且收敛速度也是最快的。当迭代次数小于 20 时,EM 算法和 dwEM 算法有相同的混淆度;当迭代次数大于 20 时,dwEM 算法随着迭代次数的增加,逐渐趋向于收敛状态并获得了较低的混淆度,其主要原因是 dwEM 在建模过程中动态地引入了有效的单词语义信息来引导建模,加快了模型收敛的速度,提高了模型的泛化性能和精度。



(a)R8

(b)20 NG

图 5 各算法混淆度随迭代次数的变化情况

Fig. 5 Perplexity of different algorithms with the number of iteration

4.3.5 算法复杂度的分析

表 4 对 4 种 LDA 算法在数据集 R8 和 20 NG 上的运行时间进行了比较。当主题数 $K=100$ 和 $K=200$ 时,迭代次数统一设置为 $T=500$,这足以使得算法在各个数据集上收敛。

表 4 LDA 算法在两个数据集上的运行时间

Table 4 Runtime of LDA algorithms on two datasets

(单位:s)

数据集 / 算法	R8		20 NG	
	$K=100$	$K=200$	$K=100$	$K=200$
VB	19079	39424	40067	77506
GS	405	791	2043	3643
EM	543	1064	2554	5026
dwEM	592	1104	2703	5408

在时间复杂度方面,从表 4 的实验结果可以看出,变分贝叶斯算法的运行时间最长,这是因为该算法在每次的迭代过程中都花费了大量时间计算 Digamma 函数。dwEM 算法的运行时间比吉布斯采样和期望最大化算法略长,其原因是在

迭代过程中，该算法需要动态地获取每个单词的权重，并反作用到模型迭代过程中。

在算法空间复杂度方面，dwEM 算法和 EM 算法都需要保存 $\mu_{w,d}(k)$，此外还需要保存文档单词矩阵、文档主题矩阵和主题单词矩阵，在此基础上，dwEM 算法还需要保存所有单词的权重，因此 dwEM 算法的空间复杂度度为 $O(D*W+K*(NNZ+D+W)+W)$，而 EM 算法的空间复杂度为 $O(D*W+K*(NNZ+D+W))$。变分贝叶斯算法和吉布斯采样都不需要 $\mu_{w,d}(k)$，因此其空间复杂度都为 $O(D*W+K*(D+W))$。

结束语　本文首先提出了传统 LDA 算法中存在主题语义连贯性较差的问题，然后在 EM 算法的框架下提出了一种基于动态权重的 LDA 算法，并将所提算法与目前流行的变分贝叶斯、吉布斯采样和期望最大化主题模型推理算法进行了实验对比。通过实验验证，本文提出的基于动态权重的 LDA 模型能够产生语义连贯性更强的主题，同时在文本分类准确率、收敛性、泛化能力和精度方面都有显著提高。但是，dwEM 算法由于需要在迭代过程中对权重进行处理和保存，因此在时间复杂度和空间复杂度方面还有待提高；同时，其在实际应用中的性能也是一个有待研究的问题，这正是我们下一步的研究方向。

参 考 文 献

[1] SALTON G, MCGILL M J. Introduction to Modern Information Retrieval [M]. New York: McGraw-Hill, 1983: 239-240.

[2] DEERWESTER S. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science & Technology, 1990, 41(6): 391-407.

[3] HOFMANN T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: IEEE Press, 1999: 50-57.

[4] HOFFMAN T. Unsupervised learning by probabilistic latent semantic indexing [J]. Sigir Audit Reports, 1999, 40(22): 28-31.

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

[6] LI X, OUYANG J, ZHOU X. Labelset topic model for multi-label document classification [J]. Journal of Intelligent Information Systems, 2016, 46(1): 83-97.

[7] WU M S. Modeling query-document dependencies with topic language models for information retrieval [J]. Information Sciences, 2015, 312(C): 1-12.

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

[9] LIU X, ZENG J, YANG X, et al. Scalable Parallel EM Algorithms for Latent Dirichlet Allocation in Multi-Core Systems[C]// Proceedings of the 24th International Conference on World Wide Web. Florence, Italy: ACM, 2015: 669-679.

[10] ZHANG J, ZENG J, YUAN M, et al. LDA Revisited: Entropy, Prior and Convergence [C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2016: 1763-1772.

[11] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing Semantic Coherence in Topic Models[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 262-272.

[12] PETTERSON J, SMOLA A, CAETANO T, et al. Word features for Latent Dirichlet Allocation[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc., 2010: 1921-1929.

[13] LI X, ZHANG A, LI C, et al. Exploring coherent topics by topic modeling with term weighting [J]. Information Processing & Management, 2018, 54(6): 1345-1358.

[14] CHEW P A, CHEW P A. Term weighting schemes for Latent Dirichlet Allocation [C]// Human Language Technologies: the 2010 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 465-473.

[15] NEWMAN D, KARIMI S, CAVEDON L. External evaluation of topic models [C]// Australasian Document Computing Symposium (ADCS). Sydney, Australia: University of Sydney, 2009: 1-8.

[16] SHAMS M, BARAANI-DASTJERDI A. Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction [J]. Expert Systems with Applications, 2017, 80(C): 136-146.

[17] GEORGE K. Human behavior and the principle of least effort: An introduction to human ecology [M]. Boston: Addison-Wesley Press, 1949: 180-183.

[18] LIN J. Divergence measures based on the Shannon entropy [J]. IEEE Transactions on Information Theory, 1991, 37(1): 145-151.

[19] WU X, ZENG J, YAN J, et al. Finding Better Topics: Features, Priors and Constraints[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. New York: Springer, 2014: 296-310.

[20] NEWMAN D, LAU J H, GRIESER K, et al. Automatic evaluation of topic coherence[C]// The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California: Association for Computational Linguistics, 2010: 100-108.

[21] CHANG D Y, YAN J F, YANG L, et al. Sliding-window Based Topic Modeling [J]. Computer Science, 2016, 43(12): 101-107. (in Chinese)
常东亚，严建峰，杨璐，等. 基于滑动窗口的主题模型[J]. 计算机科学，2016，43(12)：101-107.

[22] GAO Y, YANG L, LIU X S, et al. Study of Semantic Understanding by LDA [J]. Computer Science, 2015, 42(8): 279-282. (in Chinese)
高阳，杨璐，刘晓升，等. LDA 语义理解研究[J]. 计算机科学，2015，42(8)：279-282.