LDA

丁兆云[1,2]，王晖[1]

(1 国防科学技术大学信息系统与管理学院 长沙 410073, 2 国防科学技术大学信息系统与管理学院国防科技重点实验室 长沙 410073)

:Blei 提出的 LDA 模型通过对主题反复抽样产生文本中的每个词，而对产生的每个词在文本中的位置没有做抽样。本文在传统的 LDA 模型基础上，抽样每个词出现位置的概率分布，提出了词位置相关的 LDA 模型（PLDA）。同时针对不同的位置项定义不同的词贡献度，结合词-位置概率分布以及合适的词贡献度修正主题-词的概率分布，PLDA 在一定程度上提高了主题-词可解释精度。实验说明了通过定义不同位置项合适的词贡献度，在 NIPS 数据集上，PLDA 能够提高主题-词可解释平均精度。

:LDA；概率主题模型；词位置；词贡献度；词干扰

: TP393.08

0

概率主题模型（Probabilistic Topic Models）近年来得到非常广泛应用，包括在文本分割[1,2]，文本过滤[3]，文本分类[4]，主题分析[5]等领域。概率主题模型是从潜在语义索引 LSI（Latent Semantic Index）发展而来，通过定义一种概率产生式规则来模拟文本生成过程。概率主题模型基本观点是：文档是主题的混合，主题是词空间上的概率分布。

潜在语义索引 LSI 也称潜在语义分析 LSA（Latent Semantic Analysis），是 1988 年 S.T. Dumais 等人提出了一种新的信息检索代数模型[6]，用于知识获取和展示的计算理论和方法，它使用统计计算的方法对大量文本集进行分析，从而提取出词与词之间潜在的语义结构，并用这种潜在的语义结构，来表示词和文本，达到消除词之间的相关性和简化文本向量实现降维的目的。潜在语义分析的基本观点是：把高维的向量空间模型（VSM）表示中的文档映射到低维的潜在语义空间中。这个映射是通过对项/文档矩阵的奇异值分解 SVD（Singular Value Decomposition）来实现的。

在 LSA 基础上，哈夫曼在 1999 年引进了 PLSI（Probabilistic Latent Semantic Index）模型，也叫 aspect 模型[7]。该模型如图 1 所示，其生成公式为：

$$p(d,w_n)=p(d)\sum_z p(w_n\mid z)p(z\mid d) \tag{1}$$

PLSI 首先根据特定的文档 d，根据 $p(z\mid d)$ 选择其主题 z，然后根据 $p(w_n\mid z)$ 生成文档中的词 w。PLSI 模型对文档中主题的混合比例没有做任何假设，使得模型中的主题混合比例与特定文档相关，因此缺乏处理新文档的自然方法，待估参数的数量随着文档数量的增多线性增长，模型过度拟合。

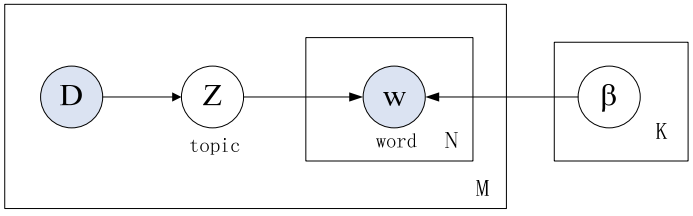

Fig. 1 The aspect model

图 1 aspect 模型

针对这些问题，Blei 等[8]在 2003 年提出的 LDA（Latent Dirichlet Allocation），在 PLSI 的基础上，用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文档的主题混合比例，模拟文档的产生过程。由于该模型将主题混合权重 θ 视为 k 维参数的潜在随机变量，而非与训练数据直接联系的个体参数集合，克服了 PLSI 模

型的不足。LDA 对主题的混合权重 θ 进了 Dirichlet 先验，用一个超参数 α 来产生参数 θ，该模型如图 2 所示。

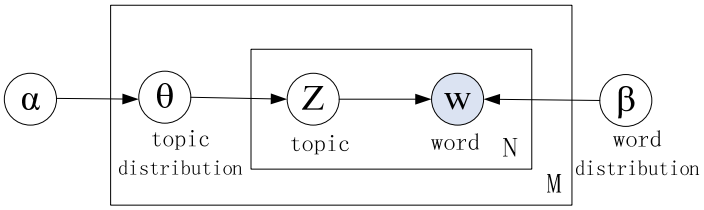

Fig. 2 The original LDA

图 2 朴素 LDA 模型

Blei 提出的 LDA 模型通过对主题反复抽样产生文本中的每个词，而对产生的每个词在文本中的位置没有做抽样。本文在传统的 LDA 模型基础上，对词抽样产生每个词出现位置的概率分布，结合主题-词和词-位置的概率分布修正主题-词的概率分布，从而在一定程度上提高了主题-词可解释精度。

为了解释主题-词分布的语义信息，Chang 等[9]提出通过词干扰方法（Word intrusion）解释主题-词分布的语义信息。词干扰方法首先随机选择一个主题，然后从该主题中选择最可能代表该主题的五个词作为样本词，同时选择一个干扰词，最后依靠是否能够很好的识别该干扰词来衡量主题-词可解释精度。样本词选择方法对主题中词概率值大小进行排序，取概率值大小排序靠前的 Top-k 词汇作为样本词汇。仅依靠主题中概率值大小排序靠前的 Top-k 词汇作为样本词汇将导致主题特征词被一些常用背景词汇淹没，将影响结果的可解释精度。本文考虑文档中不同位置项词汇对主题不同的贡献度，通过抽样每个词汇在不同位置的概率分布，结合主题-词分布与词-位置分布修正主题-词的概率分布，则能够在一定程度上减少常用背景词汇对主题特征词的影响，提高结果的的可解释精度。

另外 Chang 等[9]提出的词干扰方法没有考虑样本词空间大小影响主题-词分布可解释精度，本文提出了样本空间大小相关词干扰方法。

1

1.1 LDA

LDA 是一个三层贝叶斯概率模型，包含词、主题和文档三层结构。LDA 将每个文档表示为一个主题混合，每个主题是固定词表上的一个多项式分布。LDA 假设文档由主题混合产生，同时每个主题是在固定词表上的一个多项式分布；这些主题被集合中的所有文档共享；每个文档有一个特定的主题混合比例(Topic Proportion)，从 Dirichlet 分布中抽样产生，用一个超参数 α 来产生参数 θ。

Dirichlet 分布的公式为：

$$Dir(\alpha_1,...,\alpha_T)=\frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)}\prod_{j=1}^{T}\theta_j^{\alpha_j-1} \tag{2}$$

Dirichlet 分布和多项式分布是共轭分布。若 k 维随机向量 θ ~ Dirichlet 分布，则 θ 的 k 个分量 $\theta_1,\theta_2,...,\theta_k$ 都是连续非负值，且 $\theta_1+\theta_2+...+\theta_k=1$。

LDA 模型假设语料 D 中的每一篇文本生成过程如算法 1 所示。

算法 1. LDA 文本生成过程

①选择 $N \sim Possion(\xi)$，这里 N 代表文档长度

②选择 $\theta \sim Dir(\alpha)$，这里 θ 是列向量，代表的是主题发生的概率，α 是 Dirichlet 分布的参数

③对 N 个单词中的每一个词 w_n：

(1)选择主题 $z_n \sim Multinomial(\theta)$

(2) 根据概率分布 $p(w_n \mid z_n;\beta)$ 选择一个词 w_n，其中 p 为主题 z_n 条件下的多项式分布

在 Blei 的原始 LDA 文献[8]中提出的模型如图 2 所示，其只对文档-主题的混合参数 θ 加上了 Dirichlet 先验，而没有对主题-文档概率分布进行任何先验假设，采用的是 Mean Field Variational 推理算法。T.L.Giffiths 在文献[10]中对主题-词概率分布 ϕ 也加上了 Dirichlet 先验，然后基于多项式分布和 Dirichlet 分布的共轭特性，提出了 Gibbs 算法进行推理，该模型如图 3 所示。

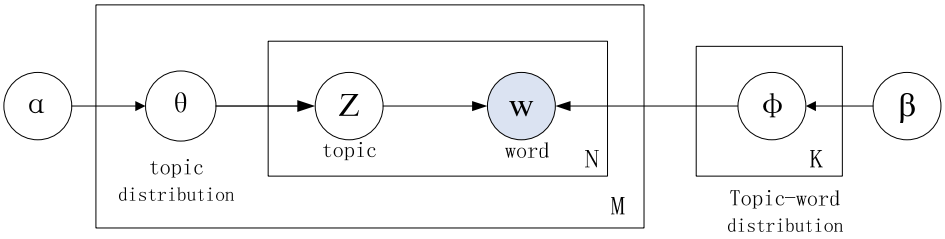

Fig. 3 The LDA model

图 3 LDA 模型

1.2 Gibbs

在 LDA 模型中为了获取词汇的概率分布，Gibbs 抽样没有将 ϕ 和 θ 作为参数直接计算，而是考虑词汇对于主题的后验概率 $p(z \mid w)$，利用抽样间接求得 ϕ 和 θ 的值。基本的 LDA 模型，仅仅需要对主题的词汇分配，也就是变量 z_i 进行抽样。Blei 的原始 LDA 并没有对 ϕ 加上 Dirichlet 先验，只是对 θ 加上 Dirichlet 先验，为了充分利用共轭概率分布的特性，便于使用 Gibbs 算法进行推理，文献[10]对 ϕ 加上 Dirichlet 先验进行扩展：

$$\begin{aligned} &w_i \mid z_i, \phi^{(z_i)} \sim Dirichlet\ (\phi^{(z_i)}), \phi \sim Dirichlet\ (\beta) \\ &z_i \mid \theta^{(d_i)} \sim Dirichlet\ (\theta^{d_i}), \theta \sim Dirichlet\ (\alpha) \end{aligned} \tag{3}$$

这里的β可以理解为，在见到语料库的任何词汇之前，从主题抽样获得的词汇出现频数，而α可以理解为，在见到任何文档文字之前，主题被抽样的频数。尽管β和α的具体取值会影响到主题及词汇被利用的程度，但不同的主题被利用的方式几乎没有变化，不同的词汇被利用的方式也基本相同，因此可以假定对称的 Dirichlet 分布，即所有的β取相同的值，所有的α取相同的值。

MCMC 是一套从复杂的概率分布抽取样本值的近似迭代方法，Gibbs 抽样作为 MCMC 的一种简单实现形式，其目的是构造收敛于某目标概率分布的 Markov 链，并从链中抽取被认为接近该概率分布值的样本。于是目标概率分布函数的给出便成为使用 Gibbs 抽样的关键。

记后验概率 $p(z_i = j \mid z_{-i}, w_i)$，计算公式如下：

$$p(z_i = j \mid z_{-i}, w_i) = \frac{\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\bullet)} + V\beta} \bullet \frac{n_{-i,j}^{d_i} + \alpha}{n_{-i,\bullet}^{d_i} + T\alpha}}{\sum_{j=1}^{r} \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{\bullet} + V\beta} \bullet \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\bullet}^{(d_i)} + T\alpha}} \tag{4}$$

其中，$z_i = j$ 表示将词汇记号 w_i 分配给主题 j，这里 w_i 被称为词汇记号是因为其不仅代表词汇 w，而且

与该词所在的文本位置相关，z_{-i} 表示所有 $z_k(k \neq i)$ 的分配。$n_{-i,j}^{(w_i)}$ 是分配给主题 j 与 w_i 相同的词汇个数；$n_{-i,j}^{(\bullet)}$ 是分配给主题 j 的所有词汇个数；$n_{-i,j}^{d_i}$ 是文本 d_i 中分配给主题 j 的词汇个数；$n_{-i,\bullet}^{(d_i)}$ 是 d_i 中所有被分配了主题的词汇个数；所有的词汇个数均不包括这次 $z_i = j$ 的分配。

Gibbs 抽样详述如下：

1）z_i 被初始化为 1 到 T 之间的某个随机整数。i 从 1 循环到 N，N 是语料库中所有出现于文本中的词汇记号个数。此为 Markov 链的初始状态；

2）i 从 1 循环到 N，根据公式(4)将词汇分配给主题，获取 Markov 链的下一个状态；

3）迭代第(2)步足够次数以后，认为 Markov 链接近目标分布，遂取 z_i （i 从 1 循环到 N)当前值作为样本记录下来。为了保证自相关较小，每迭代一定次数，记录其他样本。舍弃词汇记号，以 w 表示唯一性词，对于每个单一样本，可以按下式估算 ϕ 和 θ；

$$\hat{\phi}_w^{z=j} = \frac{n_j^{(w)} + \beta}{n_j^{(\bullet)} + V\beta}, \hat{\theta}_{z=j}^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\bullet^{(d)} + T\alpha} \tag{5}$$

其中，$n_j^{(w)}$ 表示词汇 w 被分配给主题 j 的频数；$n_j^{(\bullet)}$ 表示分配给主题 j 的所有词数；$n_j^{(d)}$ 表示文本 d 中分配给主题 j 的词数；$n_\bullet^{(d)}$ 表示文本 d 所有被分配了主题的词数。

1.3 LDA

概率主题模型近年来得到了国内外学者广泛研究。普林斯顿大学 David M.Blei 首先提出了 LDA 模型[8]，用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文档主题混合比例，模拟文档产生过程，之后考虑主题间相关性提出了 CTM 模型[11]，考虑时间信息提出了动态主题模型 DTM[12]，引进监督信息建立了 SLDA 模型[13]。LDA 的扩展模型包括作者-主题模型[14]、作者-角色-主题模型[15]、OLDA 模型[16]等。文献[17]提出了嵌套中国餐馆式层次主题模型。文献[18]考虑文档之间存在网络链接关系，结合内容与网络链接关系，为两个文档之间增加了一个二元随机变量，根据其内容特征，来刻画这种隐含的链接关系。文献[19]针对原始的 LDA 考察两个词只是基于共现的角度，不能够精确地刻画一些句子结构信息，提出了每个句子的生成都是基于语法树，且整个概率生成过程完全附着在语法树上，并且每个句子内，不同的词都有可能去选择更适合自己的主题。文献[9, 20]对 LDA 模型中主题自动标注和主题的可解释性进行了研究。

本文针对传统 LDA 模型可解释性不足，提出了一种词位置相关的 LDA 模型（PLDA），在传统的 LDA 模型基础上，对词抽样产生每个词出现位置的概率分布，结合主题-词和词-位置的概率分布修正主题-词的概率分布，使得产生的结果具有更好的可解释性。

2 PLDA

2.1

本节首先给出一些相关定义。

定义 1. 单词（word）：单词即基本的离散数据单元，为固定词汇表中 $\{1,...,V\}$ 的一个元素，一个单词即一个 v 维单元向量 $(w^0,...,w^j,...,w^v)$ ，其中 $w^j = 1$，j 为单词在词汇表中的位置，$w^u = 0(u \neq i)$ 。

定义 2. 文档（document）：文档是由 N 个单词组成的序列，即 $\boldsymbol{w} = (w_1, w_2, ..., w_N)$ ，其中 w_i 为序列中的第 i 个词。

定义 3. 词位置（word position）：词位置即单词在文档中出现的位置，词位置集合 $wp = \{wp_1, wp_2, ..., wp_n\}$ ，其中 $wp_1, wp_2, ..., wp_n$ 分别表示一个文档中的不同位置项。

定义 4. 词贡献度（word degree）：词贡献度即单词对主题的辨别能力。很显然，常用背景单词对主题的辨别能力差，词贡献度低。一个文档词贡献度为一个 m 维向量 $(d_1, d_2, ..., d_m)$ ，其中 $d_1, d_2, ..., d_m$ 分别表示一个文档不同位置的贡献度。

定义 5. 文档集（corpus）：文档集即由多个文档组成的集合 $D=\{w_1,w_2,...,w_M\}$。

定义 6. 样本空间（sample space）：样本空间即最可能代表一个主题的的单词集合，$U=\{w_1,w_2,...,w_u\}$，样本空间大小为 $|U|$。

2.2 PLDA

本文考虑词位置信息，在传统的 LDA 基础上，通过抽样生成词在文本中位置的概率分布。具体产生式模型如图 4 所示，其中空心圆圈代表隐含变量，实心圆圈代表可观察变量，有向边代表条件概率依赖，方框代表的是重复次数。

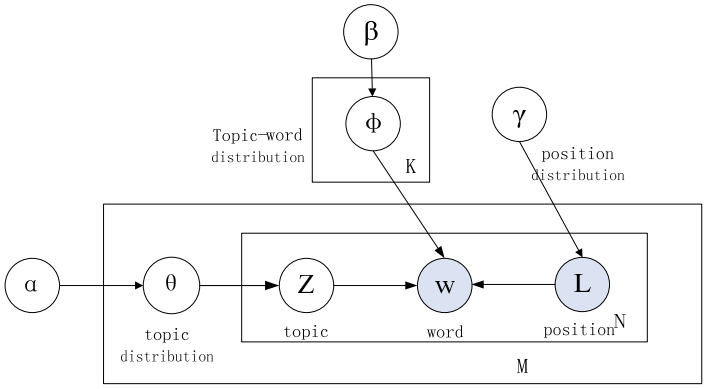

Fig. 4 The PLDA model

图 4 PLDA 模型

词位置相关 LDA 模型首先对主题反复抽样产生文本中的每个词，然后对词位置抽样产生每个词出现的位置。词位置相关 LDA 模型假设语料 D 中的每一篇文本生成过程算法 2 所示。

算法 2. PLDA 文本生成过程

①选择 $N\sim Possion(\xi)$，这里 N 代表文档长度

②选择 $\theta\sim Dir(\alpha)$，这里 θ 是列向量，代表的是主题发生的概率，α 是 Dirichlet 分布的参数

③对 N 个单词中的每一个词 w_n：

(1)选择主题 $z_n\sim Multinomial(\theta)$

(2) 根据概率分布 $p(w_n|z_n;\beta)$ 选择一个词 w_n，其中 p 为主题 z_n 条件下的多项式分布

(3)对词 w_n 所在文档中的位置 pw_{ni}：

根据概率分布 $p(pw_{ni}|w_n,\gamma)$ 选择一个词 w_n 在文档中的位置 pw_{ni}，其中 p 为单词 w_n 条件下的多项式分布

其中 θ 是一个 $k\times1$ 的随机列向量，$Dir(\alpha)$ 是 θ 的分布，具体函数形式是一个 Dirichlet 分布，这一分布保证 θ 的 k 个分量 $\theta\_1,\theta\_2,...,\theta\_k$ 都取连续的非负值，且 $\theta\_1+\theta\_2+...+\theta\_k=1$。$z_n$ 是离散随机变量，在主题 T 中取 k 个离散值，$p(z|\theta)$ 是给定 θ 时 z 的条件分布。w_n 是离散随机变量，在词汇表 V 中取|V|个离散值，$p(w_n|z_n,\beta)$ 是给定 z_n 时 w_n 的条件分布。pw_{ni} 是离散随机变量，$p(pw_{ni}|w_n,\gamma)$ 是在给定 w_n 时 pw_{ni} 的条件分布，可以看作 $|V|\times m$ 的矩阵。词位置相关的 LDA 模型在产生一篇文档前，先根据 Dirichlet 分布生成一个 $k\times1$ 的列向量 θ，生成的这个 θ 非负且归一化，可以看作某个随机变量的分布；然后根据 $p(z|\theta)$ 随即选择一个主题 z_n；再根据 $p(w_n|z_n,\beta)$ 产生文本中所有单词；最后根据 $p(pw_{ni}|w_n,\gamma)$ 确定每个词在文本中的位置。

根据词位置相关 LDA 文本生成过程的算法，在给定 α、β 和 γ 后，主题 z 和文本 w 以及文本所在的词位置项 wp 的联合概率为：

$$p(\theta,z,\boldsymbol{w},wp\mid\alpha,\beta,\gamma)=p(\theta\mid\alpha)\times \prod_{n=1}^{N}p(z_n\mid\theta)p(w_n\mid z_n,\beta)p(pw_{ni}\mid w_n,\gamma) \tag{6}$$

其中 i 表示生成文本中不同的位置项。

则文档在不同位置项边缘分布函数为：

$$p(\boldsymbol{w},wp\mid\alpha,\beta,\gamma)=\int p(\theta\mid\alpha)(\prod_{n=1}^{N}\sum_{z_n}p(z_n\mid\theta)p(w_n\mid z_n,\beta)p(pw_{ni}\mid w_n,\gamma)) \tag{7}$$

最后，根据文档在不同位置项边缘分布函数，可以得文档集在不同位置项概率如下：

$$p(D,wp\mid\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d\mid\alpha)(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}\mid\theta_d)p(w_{dn}\mid z_{dn},\beta)p(pw_{dni}\mid w_{dn},\gamma)) \tag{8}$$

2.3

词位置相关 LDA 模型能够生成每个词在文本中出现的位置，不同位置项单词对主题的辨别能力不同。比如一篇文章中，标题项和摘要项单词比正文中单词具有更高的词贡献度；主语项和宾语项单词比谓语项、补语项和状语项单词具有更高的词贡献度。本文定义一个文档词贡献度为一个 m 维向量 $(d_1,d_2,...,d_m)$，其中 $1,2,...,m$ 分别为一个文档中不同位置项。则可以通过不同位置项贡献度 d_p 修正主题-词概率分布 ϕ 为 $d_p p(w_n\mid z_n,\beta)$。

2.4 PLDA

本文结合 Gibbs 抽样与词位置抽样，根据不同位置词贡献度修正 Gibbs 抽样中 ϕ 的结果，词位置抽样算法详述如算法 3 所示。根据该算法，抽样每个词位置概率分布为：

$$\hat{\gamma}_p^i=\frac{tokenposition[i][p]}{\sum_{j=0}^{m}tokenposition[i][j]} \tag{9}$$

算法 3. 词位置抽样算法

①对词汇表 $\{1,...,V\}$ 的每个词 w_i，初始化 $|V|\times m$ 数组 tokenposition[V][m]，其中 m 表示词位置项个数

②对文档集中的每篇文档 $d\in\{\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_M\}$

③对每篇文档 d 中的每个词 $w\in\{w_1,w_2,...,w_N\}$

(1) 计算每个词的位置项 $\{p_1,p_2,...,p_N\}$，其中 p_1 表示词 w_1 的位置项，p_i 表示词 w_i 的位置项，其中 $1\le i\le N$

(2) 索引每个词在固定词汇表 $\{1,...,V\}$ 中的具体位置 $l_1,l_2,...,l_N$，其中 l_1 表示词 w_1 在词汇表的位置，l_i 表示

词 w_i 在词汇表的位置，其中 $1 \le i \le N$

(3) $\text{tokenposition[k][p]}++$，其中 k 为词 w 在词汇表的位置，p 为词 w 的位置项

④循环②和③，对文档集 D 中每篇文档以及每个词都做抽样

⑤估算词汇表 $\{1,...,V\}$ 中每个词 w_i 位置分布：

$$\hat{\gamma}_p^i = \frac{tokenposition[i][p]}{\sum_{j=0}^{m} tokenpositionn[i][j]}$$

其中，$\hat{\gamma}_p^i$ 表示词汇表中第 i 个词汇在 p 位置项的概率分布

词位置抽样结果可以估计每个词在各个位置项的概率分布，结合不同词位置贡献度 $(d_1, d_2, ..., d_m)$ 修正 Gibbs 抽样中 ϕ 的结果，修正后的主题-词概率分布 $\phi_w^{z=j}$ 如下：

$$\phi_w^{z=j} = \sum_{j=0}^{m} (\hat{\phi}_w^{z=j} \times \hat{\gamma}_j \times d_j) \tag{10}$$

其中，$\hat{\phi}_w^{z=j}$ 为 Gibbs 抽样估算的主题-词概率分布 ϕ，$\hat{\gamma}_j$ 为词在 j 位置项的概率分布，d_j 为 j 位置项的词贡献度。

4

本文实验使用数据集 NIPS[21]进行分析。其中 NIPS 数据集共包括 1988-2001 共 14 年会议 1958 篇文章的全文数据，大小为 35.1M，文件格式为文本格式，以年为单位划分成 14 个文件夹。其中全文数据包含标题，摘要、正文以及参考文献。本实验定义词位置项集合 $wp = \{title, abstract, text\}$，其中 $title, abstract, text$ 分别表示一个文档中的标题项，摘要项以及正文项，也就是词汇表 $\{1,...,V\}$ 中每个词出现的位置划分为标题、摘要以及正文。同时定义词位置项集合 $wp = \{title, abstract, text\}$ 的词贡献度分别为：$(0.49,0.49,0.02)$，$(0.05,0.05,0.9)$。

其中 $(0.49,0.49,0.02)$ 表示一个文档标题项与摘要项具有较高的词贡献度，$(0.05,0.05,0.9)$ 表示一个文档的正文项具有较高的贡献度。通过定义一个位置项的不同贡献度，验证不同位置的贡献度对主题-词可解释程度的影响。

实验设置主题数目为 50 个，迭代次数为 500 次，同时样本空间 $|U|$ 大小分别设为 4、8、12 以及观察函数 $h(|U|) = \frac{|U|}{|U|+1}$（即不减函数 $\omega(|U|) = 1$）。分别定义词贡献度为 $(0.49,0.49,0.02)$，$(0.05,0.05,0.9)$ 的模型为 $PLDA1, PLDA2$。实验从得到的主题中随机抽样 10 个主题做分析。

针对不同的样本空间 $|U|$ 的大小，实验结果如图 5 所示。

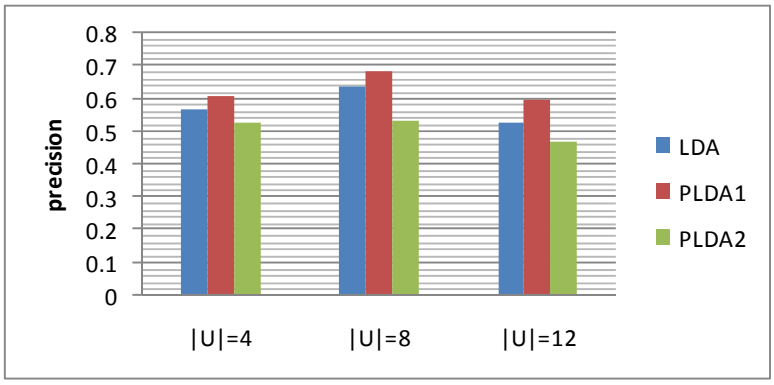

Fig. 5 The average precision of interpretability in different sample space

图 5 不同样本空间可解释平均精度

图 5 所示，在不同样本空间大小下，各模型可解释平均精度对比情况。

由图 5 所示可知，针对不同样本空间大小，*PLDA*1 模型可解释平均精度高于传统 *LDA* 模型可解释平均精度，精度提高分别约为 7.04%，7.66%以及 14.3%；同时，*PLDA*2 模型可解释平均精度却低于传统 *LDA* 模型可解释平均精度。实验结果说明当设置标题项与摘要项中的单词具有较高贡献度时，主题-词可解释平均程度优于传统 *LDA* 模型可解释平均程度；当设置正文项的单词具有较高贡献时，主题-词可解释平均程度相对较差。实验结果验证了词位置相关 *LDA* 模型能够在一定程度上影响主题-词的可解释能力，当设置不同位置项合适的词贡献度时，词位置相关 *LDA* 模型可解释平均精度可以提高约 9.67%。

5

概率主题模型近年来出现了许多理论研究成果，但如何更加全面有效地解释概率主题模型以及有效地应用概率主题模型还存在很多问题，下一步工作将结合本文的研究工作，将概率主题模型应用到话题检测、web 挖掘中，更加有效地检测 web 中的话题，且更加全面有效地解释话题。

[1] Shi Jin, Dai Guozhong. Text segmentation based on PLSA model [J]. Journal of Computer Research and Development, 2007, 44(2): 242-248(in Chinese)(石晶, 戴国忠. 基于 PLSA 模型的文本分割[J]. 计算机研究与发展, 2007, 44(2): 242-248)

[2] Shi Jin, Hu Minig, Shi Xin, Dai Guozhong. Text Segmentation Based on Model LDA [J]. Journal of Computers, 2008, 31(10): 1865-1873(in Chinese)(石晶, 胡明, 石鑫, 戴国忠. 基于 LDA 模型的文本分割[J]. 计算机学报, 2008, 31(10): 1865-1873)

[3] Bíró I, Siklósi D, Szabó J, A. Benczúr A. Linked Latent Dirichlet Allocation in Web Spam Filtering [C] // In: Proc. of the 5th international workshop on Adversarial information retrieval on the web(AIRWeb). Madrid, Spain, 2009, 37-40

[4] Li Wenbo, Sun Le, Zhang Dakun. Text Classification Based on Labeled-LDA Model [J]. Journal of Computers, 2008, 31(4): 620-627(in Chinese)(李文波, 孙乐, 张大鲲. 基于 Labeled-LDA 模型的文本分类新算法[J]. 计算机学报, 2008, 31(4): 620-627)

[5] Shi Jin, Fan Meng, Li Wanlong. Topic Analysis Based on LDA Model [J]. Acta Automatica Sinica, 2009, 35(12): 1586-1592(in Chinese)(石晶, 范猛, 李万龙. 基于 LDA 模型的主题分析[J]. 自动化学报, 2009, 35(12): 1586-1592)

[6] T.Dumais S, W.Furnas G, K.Landauer T. Using latent semantic analysis to improve access to textual information [C] //In: Proc. of the Conference on Human Factors in Computing Systems(CHI'88). Washington, D.C., United States, 1988, 281-285

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

[8] M.Blei D, Y.Ng A, I.Jordan M. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[9] Chang J, B.Graber J, Gerrish S, Wang C, M.Blei D. Reading Tea Leaves: How Humans Interpret Topic Models

[C]//Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS2009). Vancouver, B.C., Canada, 2009, 288-296

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

[11] M.Blei D, D.Lafferty J. A Correlated Topic Model of Science [J]. Annals Of Applied Statistics, 2007, 1(1): 17-35

[12] Wang C, Blei D, Heckerman D. Continuous Time Dynamic Topic Models [C]//In: Proc. of the 23rd Conference on Uncertainty in Artificial Intelligence. Helsinki, Finland, 2008, 579-586

[13] M.Blei D, D.McAuliffe J. Supervised topic models [J]. Advances in Neural Information Processing Systems, 2008, 20: 121-128

[14] Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The Author-Topic Model For Authors And Documents [C]//In: Proc. of the 20th conference on Uncertainty in artificial intelligence. Banff, Canada, 2004, 487-494

[15] McCallum A, Wang X, Corrada-Emmanuel A. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email [J]. Journal of Artificial Intelligence Research, 2007, 30: 249-272

[16] AlSumait L, Barbar´a D, Domeniconi C. On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking [C]//In: Proc. of the 2008 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008, 3-12

[17] M.Blei D, L.Griffiths T, I.Jordan M, B.Tenenbaum J. Hierarchical Topic Models and the Nested Chinese Restaurant Process [J]. Advances in neural information processing systems, 2004, 16: 17-24

[18] Chang J, M.Blei D. Relational Topic Models for Document Networks [C]//In: Proc. of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS2009). Florida, USA, 2009, 81-88

[19] Boyd-Graber J, Blei D. Syntactic Topic Models [J]. Advances in neural information processing systems, 2008, 21: 185-192

[20] Mei QZ, Shen XH, Zhai CX. Automatic Labeling of Multinomial Topic Models [C]//In: Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. California, USA, 2007, 490-499

[21] NIPS dataset: http://nips.djvuzone.org/txt.html[EB/OL]