基于 LDA 的双通道在线主题演化模型

曹建平 [1] 王 晖 [1] 夏友清 [1] 乔凤才 [1] 张 鑫 [1]

摘 要 网络舆情分析中需要处理大量时效性较强的文本数据流. 针对在线时效性较强的文本数据流, 提出基于 LDA (Latent Dirichlet allocation) 的双通道在线主题演化模型 (Bi-path evolution online-LDA, BPE-OLDA), 在下一时间片生成文本时考虑文本的内容遗传和强度遗传, 很好地模拟了人在生成时效性较强的文本时的特征. 估算模型参数时对 Gibbs 采样算法进行了简化, 实验证明, 使用简化后的在线 Gibbs 重采样算法, BPE-OLDA 模型在提取时效性较强的文本数据流的主题方面具有明显的效果.

关键词 时效性, 强度遗传, Gibbs 采样, LDA 模型

引用格式 曹建平, 王晖, 夏友清, 乔凤才, 张鑫. 基于 LDA 的双通道在线主题演化模型. 自动化学报, 2014, **40**(12): 2877−2886

DOI 10.3724/SP.J.1004.2014.02877

Bi-path Evolution Model for Online Topic Model Based on LDA

CAO Jian-Ping[1] WANG Hui[1] XIA You-Qing[1] QIAO Feng-Cai[1] ZHANG Xin[1]

Abstract There are a large number of time-sensitive texts as data streams to be processed in open-source intelligence analysis. We design a new bi-path evolution model based online-LDA (BPE-OLDA) for the time-limited text streams. This model takes consideration of both content and intensity influences to model the composition process of human successfully. When estimating the parameters of this model, we simplify the Gibbs sampling. Experiments show that BPE-OLDA performs better than other approaches over time-limited text streams.

Key words Time-sensitive, intensity influence, Gibbs sampling, latent Dirichlet allocation (LDA)

Citation Cao Jian-Ping, Wang Hui, Xia You-Qing, Qiao Feng-Cai, Zhang Xin. Bi-path evolution model for online topic model based on LDA. *Acta Automatica Sinica*, 2014, **40**(12): 2877−2886

随着网络舆情系统广泛应用, 实时分析处理网络文档成为舆情分析处理的必然要求. 借助计算机迅速从大量庞杂的文档中计算出主题, 也是文本分析的重要研究内容[1−3]. 因此, 在线主题演化技术势必成为信息分析研究的重点. 在线主题演化模型的分析处理技术需具备以下几个功能: 1) 实时性, 从互联网海量信息中快速、及时地探测到热点主题; 2) 持续性, 追踪已知主题的后续关注焦点, 掌握其发展态势; 3) 定制性, 根据用户关注的焦点对特定主题自定义, 集中资源与精力处理用户的核心要求.

针对在线主题演化技术的功能需求, 涌现出许多主题演化模型. 基于 LDA (Latent Dirichlet allocation) 的在线主题演化模型 (Online-LDA, OLDA) 的研究是在线主题演化模型的分析处理技术的一个重要研究分支. LDA 模型是典型的主题模型 (Topic model), 能很好地模拟文本生成过程, 预测文本发展趋势. 如何借助 LDA 模型引入文本语料的时间信息, 研究主题随时间演化的规律与特征成为机器学习和文本挖掘领域研究的重要方向.

本文针对时效性较强的数据流, 提出基于 LDA 的双通道在线主题演化 (Bi-path evolution online-LDA, BPE-OLDA) 模型, 该模型既考虑了主题内容的遗传性, 又考虑了主题强度的遗传性, 较全面地模拟了人生成文本的过程. 同时, 为配合 BPE-OLDA, 在参数估算时对 Gibbs 采样算法进行了简化. 对比实验证明, 该模型在主题强度和困惑度方面有很好的优势. BPE-OLDA 模型主要应用于实时监控互联网中热点主题的形成、发展、爆发及消亡的舆情过程.

1 相关工作

随着 Web 2.0 的兴起, 近年来, 自然语言处理的研究也迎来了新的发展契机, 众多学者从不同方面研

收稿日期 2013-01-11 录用日期 2013-09-12
Manuscript received January 11, 2013; accepted September 12, 2013
国家自然科学基金 (61105124, 60902091) 资助
Supported by National Natural Science Foundation of China (61105124, 60902091)
本文责任编委 李乐飞
Recommended by Associate Editor LI Le-Fei
1. 国防科学技术大学信息系统与管理学院 长沙 410073
1. College of Information System and Management, National University of Defense Technology, Changsha 410073

究文本, 特别是短文本的主题[4]、情感[5−6]、语义[7]等. 针对文本主题的挖掘方法越来越深入细致, 针对不同类型文本特点设计的算法不断涌现.

1.1 基于传统聚类的主题挖掘算法

最早的主题挖据方法实际上可认为是传统的聚类算法在文本空间上的应用. 这类算法将文本中的非结构化数据看作是向量空间里的点, 可以通过向量空间模型 (Vector space model, VSM) 进行映射, 再利用传统的聚类算法实现文本聚类. 这些聚类方法包括基于划分的算法 (如 K-means 算法)、基于层次的算法 (自顶向下和自底向上算法)、基于密度的算法等[4, 8]. 虽然这类算法的聚类结果可以近似认为满足同一个主题, 但有两个明显缺陷: 1) 算法的核心是对距离的计算, 而海量的文本信息的距离很难定义; 2) 算法完全不涉及语义信息, 与人们的理解差距很大.

1.2 基于矩阵模型的主题挖掘算法

潜在语义分析 (Latent semantic analysis, LSA) 是 Deerwester 等[9] 提出的一种基于矩阵模型挖掘文本主题的新方法. LSA 通过奇异值分解 (Singular value decomposition, SVD) 的降维方法挖掘文档的潜在结构 (语义结构), 该方法可在低维的语义空间里进行查询和相关性分析, 挖掘文档中隐含的语义相关性.

根据 Landauer 等的研究[10], 当这个语义空间的维度和人类语义理解的维度相近时, LSA 的计算结果能很好地被人们理解, 即其表面信息转化为深层次的抽象具有一定的意义. 但由于一个单词在语义空间中只有一个坐标, 无法用多个坐标来表示多个意义, 所以 LSA 无法处理 “一词多义” 的问题; 且 SVD 涉及到矩阵运算, 计算复杂度比较高.

1.3 基于概率模型的主题挖掘算法

主题模型是一种对文字中隐含主题的建模方法, 通过概率的产生式模型 (Generative model) 挖掘文本主题[11]. 主题模型假设文本的生成过程是不同的主题按一定的规则选择单词的过程; 反过来, 已知文本单词的分布情况, 就可通过概率方法反推出文本集的主题分布情况. 主题模型中最具代表性的模型是概率潜在语义分析 (Probabilistic latent semantic analysis, PLSA) 和 LDA. PLSA 是在 LSA 基础上提出的基于最大似然法 (Maximum likelihood) 和产生式模型的概率模型. Hofmann[12] 在提出 PLSA 时沿用了 LSA 的 “降维” 思想, 通过 “降维” 将文档从高维空间投影到了语义空间. PLSA 与 LSA 不同的是, LSA 是以共现表 (即共现矩阵) 的 SVD 的形式表现的, PLSA 是基于派生自最小公倍数 (Lowest common multiple, LCM) 的混合矩阵分解. PLSA 一般运用期望最大化 (Expectation-maximization, EM) 算法对模型进行求解. 实际运用中, 由于 EM 算法的计算复杂度小于 SVD 算法, 因此 PLSA 在性能和处理大规模数据方面通常也优于 LSA.

PLSA 在取得很大进步的同时, 由于过多的参数也会导致过拟合 (Overfiting) 现象[13]. LDA 的创始者 Blei 在 PLSA 的基础上加入了 Dirichlet 先验分布, 解决了过拟合问题, 是一个突破性的进展[14]. LDA 引入了超参数, 形成了一个 “文档 – 主题 – 单词” 3 层贝叶斯模型, 然后运用概率方法对模型进行推导, 寻找文本集的语义结构, 挖掘文本的主题.

主题模型的应用领域包括主题挖掘[15]、文本检索[16]、文本分类[17]、引文分析[18] 和社交网络分析[19] 等, 此外还应用于处理非文本信息, 包括计算机视觉、图像等领域.

近年来对主题模型的研究不断深化, 衍生出了各种各样的模型, 如 Dynamic topic model[20]、Syntactic topic model[21] 等. Link-PLSA-LDA[22] 和 HTM (Hypertext topic model)[23] 将文本间的关联作为影响因素, 更好地对超文本进行主题挖掘和文本分类, 这些模型都利用了特定文本集对象的特征.

2 双通道在线主题演化模型

人在创作一篇时效性较强的文本 (如新闻报道或论坛帖子) 时, 会不自觉地参考当前网络中的热点主题及其重要程度, 确定文本的具体内容. 新闻的及时性及网民浏览新闻和论坛的习惯决定了时效性较强的文本必然具备上述特征. 归结起来, 时效性很强的数据流具有如下特性: 当前时间片内的文本会受到先前时间片内的内容和强度的影响, 同时也影响着下一时间片内文本的生成.

传统的基于 OLDA 模型并没有从人在生成一篇文本时所受当前热点主题影响的角度去架构数学模型[20, 24−26], 或只考虑到文本内容的遗传而没有考虑强度的遗传. 本文针对具有较强时效性的文本流, 从人在选择文本主题时所受影响的角度, 结合基于 LDA 的在线主题演化模型, 提出基于 LDA 的双通道在线主题演化模型, 即在模拟生成文本时, 同时考虑主题内容遗传和主题强度遗传两个因素.

2.1 主题内容遗传度

主题内容遗传度 C[23−25] 体现了从历史文本集中获得的主题信息对未来的文本挖掘主题信息的

影响. 将当前时间片 t 及之前时间片的主题信息遗传到 $t+1$ 时间片的主题内容遗传度定义为: $C^t = H^t B^t$. 其中 H^t 是内容演化矩阵 B^t 中各时间片主题的混合比例, 表示 $t+1$ 之前的 L 个时间片的历史文本数据分别对 $t+1$ 时间片内的文本挖掘的影响方式. $t+1$ 时间片内的不同主题关于词的多项式分布的 Dirichlet 先验分布参数, 可通过 $\beta^{t+1} = C^t$ 给定. 该式保证了内容演化矩阵中的主题自动对齐, 即新时间片挖掘出的主题序号自动与先前时间片内具有相同语义含义的主题序号对齐. 利用主题的自动对齐, 可方便地探测新主题产生, 判断旧主题的消亡. 图 1 的下半部分即为文档的内容遗传部分的示意.

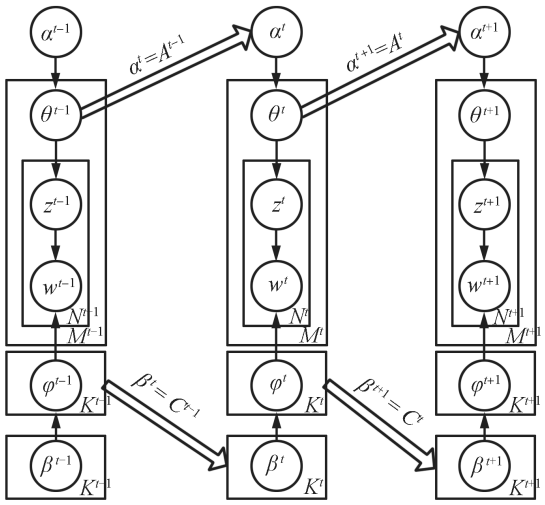

图 1　BPE-OLDA 模型框架图

Fig. 1　Pattern of BPE-OLDA model

本文用 $H^t = (\bar{N}^{t-(L-1)}, \bar{N}^{t-(L-2)}, \cdots, \bar{N}^t)$ 表示内容演化矩阵的混合比例, 其中 $\bar{N}$ 表示文本集 D 中所有词的数目. L 表示 $t+1$ 之前的 L 个时间片. 该内容遗传度混合比例考虑了到不同时间片内文本集所含的文本数和词数差异性, 文本集越大, 挖掘出的主题信息遗传影响力越大, 越符合人工判别的结果.

2.2　主题强度遗传度

主题强度遗传度是表示先前时间片的主题信息对新时间片内具体文本选择主题产生的影响程度的数学模型, 即在未知文本集内容的情况对文本下主题的分布先验假设. 该先验假设是基于文本数据流具有很强的时效性做出的.

假设当前时间片为 t, 将该时间片产生的主题强度遗传度记为 $A^t = E^t R^t$, 其中 R^t 为主题强度演化矩阵, E^t 为各时间片内的主题强度遗传混合比例向量. 主题强度演化矩阵 R^t 表示 $t+1$ 时刻之前 F 个时间片内的主题强度, 是二维 $F \times K^t$ 矩阵, $R^t = [R_1^t, R_2^t, \cdots, R_{K^t}^t]$, 其中 R_k^t 表示主题 k 在 F 个时间片内的主题强度值, 是 $F \times 1$ 的列向量.

构造主题强度演化矩阵 R^t 时需采取补零策略, 即若 F 个时间片内没有主题, 则将其主题强度设置为 0. 主题强度遗传混合比例向量 E^t 表示 $t+1$ 时刻之前的 F 个时间片内的主题强度的混合方式, 构造 E^t 时需考虑不同时间片文本集的规模, 规模越大的文本集产生的主题强度影响力越大, 故本文令 $E^t = [M^{t-(F-1)}, M^{t-(F-2)}, \cdots, M^t]$, 则该 E^t 的设定对不同时间片的文本集规模有平滑作用. 通过设置 E^t 的值决定当前时间片继承了之前的时间片的主题强度.

2.3　BPE-OLDA 模型

利用主题强度遗传度给定新时间片的文本关于主题分布的 Dirichlet 分布的先验参数 α 是 BPE-OLDA 模型区别基于 LDA 的在线主题演化模型的核心部分. 假设当前时间片为 t, 则根据 BPE-OLDA 模型有

$$\alpha^t = A^{t-1} = E^{t-1} R^{t-1} \tag{1}$$

式 (1) 体现的实际意义不仅是在单个时间片内预先给定了文本下主题分布的 Dirichlet 分布的先验参数, 然后在单个时间片内使用 LDA 模型挖掘主题信息, 而是与主题内容遗传结合, 使得在线主题演化模型能够充分利用各时间片遗传的主题信息, 从而细致地弥合了由于时间切片造成的主题信息间的裂痕, 更加准确地描述主题在数据流中的演化规律.

图 1 是考虑了主题内容和强度两方面遗传性的 BPE-OLDA 模型框架图. 从图 1 可以看出, 基于 LDA 的双通道在线主题演化模型是通过主题内容遗传和主题强度双通道的. 下一时间片内容的生成直接受之前主题内容和强度的影响, 可以更准确地描述时效性较强的文本的生成过程.

BPE-OLDA 模型通过对主题强度遗传性的引入, 很好地模拟了在较强时效性数据流中, 人生成文本时参考的因素. 该模型的详细生成过程如下:

将时间流划分成离散的时间片, 对其中每个时间片的文本集:

1) 对于时间片 t, $\beta^t = C^{t-1} = H^{t-1} B^{t-1}$, 当 $t = 1$ 时, β^t 弱化为常向量 $\boldsymbol{\beta}^0$, $\boldsymbol{\beta}^0$ 通过模型初始化给定;

2) 对于时间片 t, $\alpha^t = A^{t-1} = E^{t-1} R^{t-1}$, 当 t

$=1$ 时, α^t 弱化为常数 α^0, α^0 是模型通过初始化给定;

3) 对于 T 中每个主题 z^t, 根据 $\phi_z^t \sim Dir(\beta_z^t)$, 抽样得到 ϕ_z^t;

4) 对于 D^t 中每个文本 d^t, 根据 $\theta_d^t \sim Dir(\alpha^t)$, 抽样得到 θ_d^t;

5) 对于文本 d^t 的第 i 个词 $w_{d,i}^t$:

a) 根据多项式分布, $z_{d,i}^t \sim Multi(\theta_d^t)$, 抽样得到主题 $z_{d,i}^t$;

b) 根据多项式分布, $w_{d,i}^t \sim Multi(\phi_z^t)$, 抽样得到主题 $w_{d,i}^t$.

对上述生成过程, 本文利用简化的在线 Gibbs 重采样算法 (Rect-online Gibbs re-sampling, Rect-OGRS) 估算 BPE-OLDA 的参数 $(\boldsymbol{\theta}, \boldsymbol{\phi})$.

2.4 简化 Gibbs 采样

Gibbs 采样算法是马尔科夫链蒙特卡洛方法 (Markov chain Monte Carlo method, MCMCM) 的一种简单实现形式[27−28]. LDA 模型使用 Gibbs 采样算法可估算出文本下主题的分布和主题下词的分布, 即 $(\boldsymbol{\theta}, \boldsymbol{\phi})$, 但 Gibbs 采样算法不能进一步估算得到 LDA 模型的模型参数 (α, β), 如需得到 (α, β) 的值, 必须结合 EM 算法求解. 一般实际应用时只需得到 $(\boldsymbol{\theta}, \boldsymbol{\phi})$ 的估算值. 构造马尔科夫链进行 Gibbs 采样有多种方式[29], 可构造多条马尔科夫链并从每条链末抽样; 也可构造一条马尔科夫链, 当链长达到某个预先给定的长度时, 每隔一定数量的状态, 抽样一次, 利用采样得到的样品估算参数 $(\boldsymbol{\theta}, \boldsymbol{\phi})$.

使用简化的在线 Gibbs 采样算法 (Rect-online Gibbs sampling, Rect-OGS) 估算参数 $(\boldsymbol{\theta}, \boldsymbol{\phi})$ 包括 3 个步骤: 1) 将文本集 D 中所有的词, 按先文本集中文本的顺序, 后文本中词的顺序进行排序, 构造多元随机变量 $(z_1, z_2, \cdots, z_{\bar{N}})$; 2) 逐一构造马尔科夫链中的各个状态, 而马尔科夫链中的状态就是 $(z_1, z_2, \cdots, z_{\bar{N}})$ 依赖于前一个状态的一次抽样; 3) 在足够长的马尔科夫链末抽样 $(z_1, z_2, \cdots, z_{\bar{N}})$, 得到 $(\boldsymbol{\theta}, \boldsymbol{\phi})$ 的估算值 $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$.

$$P(z_i = j | z_{-i}, \bar{W}) = \frac{n_{-i,j}^{(w_i)} + \beta_{-i,j}^{(w_i)}}{n_{-i,j}^{(\cdot)} + \sum\limits_{w_i} \beta_{-i,j}^{(w_i)}} \times \frac{n_{-i,j}^{(d_i)} + \alpha_j}{n_{-i,\cdot}^{(d_i)} + \sum\limits_{j} \alpha_j} \tag{2}$$

式 (2) 是马尔科夫链根据前一个状态构造当前状态的更新公式. 式 (3) 和式 (4) 是从马尔科夫链中抽样所得的状态 $(z_1, z_2, \cdots, z_{\bar{N}})$ 估算 $(\boldsymbol{\theta}, \boldsymbol{\phi})$ 的公式:

$$\hat{\theta}_j^{(d_i)} = \frac{n_{-i,j}^{(d_i)} + \alpha_j}{n_{-i,\cdot}^{(d_i)} + \sum\limits_{j} \alpha_j} \tag{3}$$

$$\hat{\phi}_j^{(w_i)} = \frac{n_{-i,j}^{(w_i)} + \beta_{-i,j}^{(w_i)}}{n_{-i,j}^{(\cdot)} + \sum\limits_{w_i} \beta_{-i,j}^{(w_i)}} \tag{4}$$

其中, $\beta_j^{t,(w_i)} = (C_j^{t-1})^{(w_i)}$. Rect-OGS 算法将式 (3) 和式 (4) 简化为

$$\hat{\theta}_j^{(d_i)} = \frac{n_{-i,j}^{(d_i)}}{n_{-i,\cdot}^{(d_i)}}, \quad 0 \le j \le K^t,\ 1 \le i \le \bar{N}^t \tag{5}$$

$$\hat{\phi}_j^{(w_i)} = \frac{n_{-i,j}^{(w_i)}}{n_{-i,j}^{(\cdot)}}, \quad 0 \le j \le K^t,\ 1 \le i \le \bar{N}^t \tag{6}$$

式 (5) 和式 (6) 对式 (3) 和式 (4) 的简化不仅保持主题信息间的遗传性, 同时排除了式 (3) 和式 (4) 估算时由于主题内容遗传度 C^{t-1} 的影响造成当前时间片 t 内主题下词相互关系扭曲的可能性. 本文默认当前时间为 t, 除特别说明外, 均默认所有符号都有 t 上标.

1) Rect-OGS 算法和在线 Gibbs 采样算法保持主题内容信息遗传的简单解释.

由第 2.1 节可知, 主题信息的遗传通过公式 $\beta = C^{t-1}$ 实现. 假设有文本 d, 主题 k_1 和词 $w = m$, m 表示词表 S 中的序号. 假设词 w 在 Dirichlet 分布的先验参数 β 中有 $\beta_{k_1}^{(w)} > \beta_{k_{-1}}^{(w)}$, 其中 k_1 表示主题, k_{-1} 表示除主题 k_1 的其他任何主题, 则在构造马尔科夫链的状态时, 需要对文本集 D 中每个位置的词 (如第 i 个位置的词 $w_i = m$), 根据式 (2) 进行抽样得到产生该词的主题. 由于 $\beta_{k_1}^{(w)}$ 较大, 使得 $n_{-i,k_1}^{(w_i)} + \beta_{k_1}^{(w)}$ 较大. 如不考虑 $n_{-i,k_1}^{(w_i)} + \alpha_{k_1}$, 则该位置 i 抽样得到主题 k_1 的概率就较高. 对其他的词也是如此. 从而促使文本集中的各个词向内容遗传度 β 中所在 "概率值" 较高的主题下集中, 形成主题内容的遗传性. 值得注意的是, β 中主题的 "概率值" 被放大了若干倍, 具体放大倍数由内容演化矩阵的主题混合比例向量 H^{t-1} 决定; 最后通过式 (2) 影响文本下主题的分布. 将词向量内容遗传度 β 中置于 "概率值" 较高的主题下保持了时间片主题内容信息的遗传性, 确保了不同时间片内主题对齐. 由此可见, Rect-OGS 算法的简化方式没有影响在线 Gibbs 采样算法关于主题内容信息的遗传性 (式 (5) 对式 (3) 的简化理由类似).

2) 在线 Gibbs 采样算法关于主题下词分布的估算式 (4) 造成当前时间片 t 内主题下词相互关系扭

曲的可能性.

当前时间片 t 的主题下词分布的 Dirichlet 分布的先验参数 β 由 $C^{t-1} = H^{t-1}B^{t-1}$ 给定, 如 β 中主题下词的 "概率值" 不够大, 会造成主题内容信息的遗传性减弱, 但是如果 β 中主题下词的 "概率值" 过大, 就可能出现如下情况: 假设对于主题 j 下的两个词 $w_i = m_1$, $w_h = m_2$, 这两个词在文本集 D 中的位置分别为 i, h , 如果存在 $\beta_j^{(w_i)}$ 的值远大于 $n_{\cdot,j}^{(w_h)} + \beta_j^{(w_h)}$, 而 $n_{\cdot,j}^{(w_i)}$ 远小于 $n_{\cdot,j}^{(w_h)}$, 就会造成虽然在文本集 D 中词 w_i 由主题 j 产生的次数远远小于词 w_h 由主题 j 产生的次数, 但词 w_i 在主题 j 下的概率值却远大于词 w_h 在主题 j 下的概率值. 这种情况是矛盾的, 也与人工判别结果相差很远.

3) 在线 Gibbs 采样算法关于估算式 (4) 造成当前时间片 t 内主题下词相互关系扭曲的可能性存在的原因.

由式 (2) 构造的马尔科夫链 $Z_1, Z_2, \cdots$, 其中每个状态有 $Z = (z_1, z_2, \cdots, z_{\bar{N}})$, 根据 Gibbs 采样算法的收敛性可得到 $z_i|w_i$ 的概率分布函数 $P(z_i|w_i)$, 其中 $1 \le i \le \bar{N}$. 由 LDA 模型的生成过程可得到式 (7) 和式 (8):

$$P(z_i|w_i) = \theta_j^{(d_i)}\phi_j^{(w_i)} \tag{7}$$

$$P(z_i = j|z_{-i}, \bar{W}) = \hat{\theta}_j^{(d_i)}\hat{\phi}_j^{(w_i)} \tag{8}$$

根据式 (7) 和式 (8) 可得到在线 Gibbs 采样算法的估算式 (6). 式 (7) 的估算方式是通过独立获得文本集中每个位置上 $z_i|w_i$ 的概率分布函数 $P(z_i|w_i)$ 所得, 显然这种方式只考虑词与主题的关系, 没有考虑词与词之间的关系, 由 LDA 模型的词包假设可知, 文本中词与词之间没有语义结构, 但词与词之间并不是独立的, 因此式 (3) 最后的估算主题下词分布的方式与词包假设相违背.

4) Rect-OGS 算法简化式 (6) 的合理性.

假设 $\boldsymbol{Z}_p$ 是从足够长的马尔科夫链末得到状态. 考虑状态 $\boldsymbol{Z}_p$ 的第 i 个位置的情况, 若 $z_i = j$, 则 $z_i = j|w_i$ 不仅可理解为主题 j 是服从概率分布 $P(z_i|w_i)$ 的一个样品, 也可理解为第 i 个位置的词 w_i 由主题 j 产生的一个实例, 即词 w_i 是从概率分布函数 $P(w|z = j)$ 的文本中抽取的一个样品, 同样, 对文本集 D 中所有位置的词都如此, 于是可用词 w_i 由主题 j 产生的次数除以主题 j 产生所有词的总次数作为 $\phi_j^{(w_i)}$ 的估算值, 即式 (6). 该简化公式不仅保持了时间片之间的主题内容遗传性, 同时也考虑了主题下词的相互关系, 避免了由于 H^{t-1} 放大倍数设置过大造成挖掘出的主题信息失真, 但该简化公式会造成相邻时间片人工解读的同一个主题的相似度下降.

由于单纯使用 OGS 会挖掘出噪音主题, 因此需在 OGS 挖掘出主题的基础上进行 Gibbs 重采样 (Gibbs resampling) 去除噪音主题. BPE-OLDA 模型上 Rect-OGRS 算法如算法 1 所示. 其中, 步骤 7)、15)、18) ∼ 21)、24) 和 25) 是 Rect-OGRS 算法用于 BPE-OLD 模型主题信息挖掘所做改动之处.

算法 1. Rect-OGRS 算法

1) 对每个时间片 t ($t = 1 : +\infty$)
2) 对 t 内文本集 D_t, 随机初始化 z_i, 初始化值为 $1 \sim K$, 且 $i = 1 : \bar{N}$
3) For $l = 1 : 5\,000$
4) For $i = 1 : \bar{N}$
5) 根据式 (1) 对 z_i^l 采样
6) EndFor
7) 当 $l \ge 2\,000$ 且 $l \equiv 0 \pmod{100}$ 时, 根据式 (5) 和式 (6) 得到估算值 $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$
8) EndFor
9) 将上述循环中获得 $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ 所有相加并除以 31, 得到该时间片的 $(\boldsymbol{\theta}^t, \boldsymbol{\phi}^t)$
10) 根据 $\boldsymbol{\theta}^t$ 获得各个主题的强度, 根据主题强度阀值过滤噪音主题
11) For $\bar{l} = 1 : 1\,000$
12) For $\bar{i} = 1 : \bar{N}$
13) 根据式 (7) 对 $z_i^{\bar{l}}$ 采样, 采样值不能是标注的噪音主题
14) EndFor
15) 当 $l \ge 400$ 且 $l \equiv 0 \pmod{100}$ 时, 根据式 (5) 和式 (6) 得到估算值 $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$
16) EndFor
17) 将步骤 7) 及步骤 15) 获得的 $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ 求平均值, 得到该时间片的 $(\boldsymbol{\theta}^t, \boldsymbol{\phi}^t)$
18) 根据 A^{t-1} 与 $\boldsymbol{\theta}^t$, 构造强度演化矩阵 R^t
19) 根据 B^{t-1} 与 $\boldsymbol{\phi}^t$, 构造内容演化矩阵 B^t
20) 根据 B^{t-1} 与 $\boldsymbol{\phi}^t$ 探测新主题, 更新内容演化矩阵 B^t, 强度演化矩阵 R^t
21) 判定消亡的旧主题, 更新内容演化矩阵 B^t, 强度演化矩阵 R^t
22) 构造主题内容遗传度 $C^t = H^t B^t$
23) $\beta^{t+1} = C^t$ 为时间片 $t+1$ 内 LDA 模型主题下词分布的 Dirichlet 分布先验参数
24) 构造主题强度强度遗传度 $A^t = E^t R^t$
25) $\alpha^t = A^t$ 为时间片 $t+1$ 内 LDA 模型文本下主题分布的 Dirichlet 分布先验参数

3 案例

为验证 BPE-OLD 模型, 本文选择台湾时政新闻作为在线文本数据流, 通过爬行器采集了来源于 TVBS、东森新闻等 20 余家台湾主流网站上的新闻,

时间从 2012 年 3 月 19 日至 2012 年 5 月 13 日. 该数据集中每篇报道至少包含一个主题.

通过分词, 将文本切成有完整独立意义的词语; 保留名词与动词, 去掉形容词、副词、英文及数字等; 丢弃文中的停用词; 将一些常见的同义词替换成同一个词, 如 “阿扁”, “陈水扁”, “扁” 统一替换成 “陈水扁”; 统计新闻数据集内每个词语出现的次数, 去掉出现次数小于 5 的词语, 从而过滤掉一些不常见的人名、地名及错写的词语; 最后将词总数小于 5 的新闻文本丢弃掉. 完成上述预处理后得到台湾时政类新闻数据集, 分为 7 个时间片, 时间片长大部分为 7 天, 个别时间片长为 14 天; 每个时间片内大约有 140 篇文档, 共有 1 011 篇文本; 其中有 3 675 互异的词, 共有 151 975 个词.

为便于观察比较模型和估算方法的优劣, 本文选择 OLDA + 在线 Gibbs 采样 (简称 OGS), OLDA + 简化在线 Gibbs 重采样算法 (简称 Rect-OGRS) 和 BPE-OLDA + Rect-OGRS 共 3 个组合作为观察对象进行对比实验.

主题内容的演化体现了同一个主题随着时间的推移, 主题下词分布的改变, 从该变化可跟踪网络中舆情关注焦点的变化. 在线主题演化模型中, 主题是通过该主题下词的概率值标识的, 因此该主题的实际语义含义需人工进行判别. 经人工判别该文本集主要包含 “陈水扁特赦”、“美牛事件” 和 “马英九国情咨文” 3 个主要话题. 以 “陈水扁特赦” 为例分析比较 OLDA + OGS、OLDA + Rect-OGRS 和 BPE-OLDA + Rect-OGRS 对主题内容演化规律刻画的效果. 选择主题下概率值最高的前 15 个词为代表, 表 1 ∼ 3 中第 1 行均表示时间片的序号. 需要说明的是, 由于 3 个组合都探测到 “陈水扁特赦” 主题从第 5 个时间片后消亡, 且经人工判别符合实际文本内容, 故没有在表中列出.

观察表 1 可以发现, 每个时间片 “陈水扁特赦” 主题下出现的词一方面大致相同, 另一方面词与词之间的相互关系大致相同, 但是表 2 和表 3 中 “陈水扁特赦” 主题除去诸如 “陈水扁”、“特赦”、“马英九” 等几个特别高频词外, 每个时间片中该主题的内容仍有改变, 如表 2 和表 3 在时间片 2 都出现了新人物柯建民和陈唐山, 时间片 3 出现了许信良, 时间片 4 出现了郑文龙, 这些人物经人工判别均为大力呼吁特赦陈水扁的重要人物, 并且在文本中反复出现. 出现该现象的原因是 OGS 算法即在线 Gibbs 采样算法估算主题下词分布的方式没有考虑到主题下词与词之间的相互关系, 如果主题内容演化度各个主题下某些词的 “概率值” 值过大, 就会造成各个时间片内主题下词的相对位置固定的现象; 而 Rect-OGRS 算法即修正的在线 Gibbs 重采样算

表 1 OLDA + OGS 挖掘出的 “陈水扁特赦” 内容演化表

Table 1 Content evolution of “Chen Shui-Bian amnesty” by OLDA + OGS

1		2		3		4	
陈水扁	0.05305	陈水扁	0.06123	陈水扁	0.06214	陈水扁	0.06416
总统	0.04825	特赦	0.04836	特赦	0.04915	总统	0.05224
特赦	0.03610	总统	0.04241	总统	0.04703	特赦	0.04868
民进党	0.02495	民进党	0.03005	民进党	0.03172	民进党	0.03179
马英九	0.01668	马英九	0.01595	马英九	0.01714	马英九	0.01922
陈菊	0.01061	连署	0.01533	连署	0.01494	连署	0.01325
问题	0.01008	立委	0.01316	立委	0.01246	立委	0.01288
立委	0.00949	问题	0.01197	问题	0.01166	问题	0.01240
保外就医	0.00927	保外就医	0.00913	保外就医	0.01012	保外就医	0.01035
高志鹏	0.00891	陈菊	0.00861	陈菊	0.00978	陈菊	0.01005
人权	0.00789	司法	0.00843	关心	0.00970	人权	0.00819
医疗	0.00746	医疗	0.0745	应该	0.00928	司法	0.00793
立场	0.00738	人权	0.00733	高志鹏	0.00745	医疗	0.00731
司法	0.00720	政治	0.00703	人权	0.00727	立场	0.00729
质询	0.00554	社团	0.00635	司法	0.00721	政治	0.00688

表 2 OLDA + Rect-OGRS 挖掘出的 “陈水扁特赦” 内容演化表

Table 2 Content evolution of “Chen Shui-Bian amnesty” by OLDA + Rect-OGRS

1		2		3		4	
陈水扁	0.06389	陈水扁	0.06680	特赦	0.06060	特赦	0.06711
总统	0.05800	特赦	0.06336	陈水扁	0.05704	陈水扁	0.06006
特赦	0.04349	民进党	0.03394	民进党	0.03530	连署	0.03876
民进党	0.02904	连署	0.03090	总统	0.03290	民进党	0.03343
马英九	0.01717	总统	0.02512	连署	0.03111	陈菊	0.02046
陈菊	0.01343	立委	0.01748	许信良	0.02054	问题	0.01956
应该	0.01248	社团	0.01478	立委	0.01258	立委	0.01424
立委	0.01173	柯建铭	0.01401	马英九	0.01249	郑文龙	0.01231
保外就医	0.01117	移监	0.01184	决议	0.01035	陈菊	0.01221
高志鹏	0.01069	马英九	0.01164	保外就医	0.00986	本土社团	0.01185
人权	0.00942	问题	0.01159	行动	0.00824	人权	0.001001
医疗	0.00894	本土社团	0.01064	政治	0.00822	司法	0.001023
司法	0.00872	司法	0.01052	本土社团	0.00799	医疗	0.00845
立场	0.00731	政治	0.01032	党主席	0.00695	党主席	0.00763
连署	0.00653	陈唐山	0.00971	应该	0.00693	政治	0.00606

表 3 BPE-OLDA + Rect-OGRS 挖掘出的 “陈水扁特赦” 内容演化表

Table 3 Content evolution of “Chen Shui-Bian amnesty” by BPE-OLDA + Rect-OGRS

1		2		3		4	
陈水扁	0.06361	陈水扁	0.06694	特赦	0.06316	特赦	0.06334
总统	0.05636	特赦	0.06351	陈水扁	0.05943	陈水扁	0.06098
特赦	0.04340	民进党	0.03193	民进党	0.03478	连署	0.05465
民进党	0.02641	连署	0.03099	连署	0.03243	民进党	0.03032
马英九	0.01817	总统	0.02425	许信良	0.02139	陈菊	0.02334
民进党	0.016	立委	0.01622	立委	0.01291	郑文龙	0.02010
陈菊	0.01379	社团	0.0147	移监	0.01026	立委	0.01705
问题	0.01235	柯建铭	0.01360	保外就医	0.01024	绿营	0.01506
立委	0.01144	马英九	0.01199	决议	0.00860	陈菊	0.01001
医疗	0.01101	移监	0.01187	政治	0.00848	本土社团	0.010
人权	0.0097	政治	0.01083	柯建铭	0.00748	人权	0.00945
保外就医	0.01096	陈唐山	0.01079	主张	0.00707	司法	0.00967
高志鹏	0.01048	本土社团	0.00975	本土社团	0.00703	医疗	0.00689
人权	0.00926	决议	0.00864	参与	0.00663	党主席	0.00606
连署	0.00651	保外就医	0.00776	绿营	0.00606	政治	0.00554

法通过基于主题下词相对关系, 修正了 OGS 算法的估算方式, 避免了该问题.

从 OLDA + OGS、OLDA + Rect-OGRS 和 BPE-OLDA + Rect-OGRS 对本小节的主题内容演化规律的刻画, 可知 OGS 算法存在 3 个问题: 1) 主题下词与词之间相对关系, 导致一些较重要的词

无法成为关键词; 2) 在描述主题消亡时存在滞后效应; 3) 对单个时间片内的主题信息挖掘不够细腻. 而 Rect-OGRS 算法则避免了上述问题.

3.1 主题强度演化的结果与分析

“陈水扁特赦”、“马英九国情咨文” 和 “美牛事件” 的主题强度演化图如图 2～4 所示.

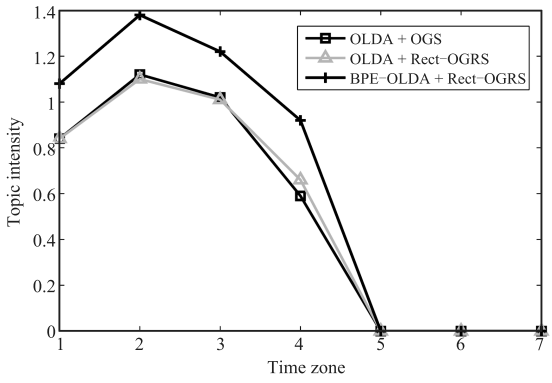

图 2 “陈水扁特赦” 主题强度演化图

Fig. 2 Topic intensity evolution of “Chen Shui-Bian amnesty”

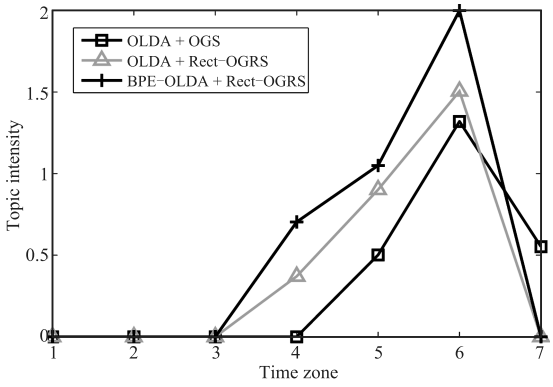

图 3 “马英九国情咨文” 主题强度演化

Fig. 3 Topic intensity evolution of “Ma Ying-Jiu state address”

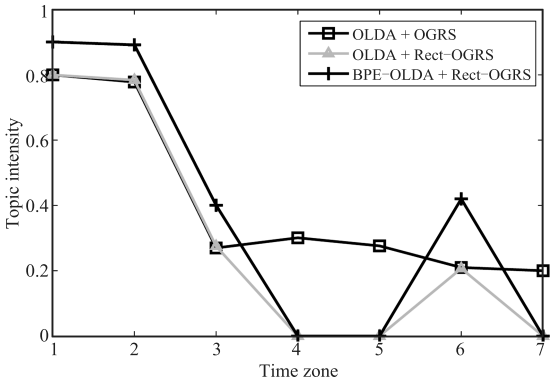

图 4 “美牛事件” 主题强度演化图

Fig. 4 Topic intensity evolution of “American beef event”

从图 2～4 可以看出, “陈水扁特赦” 主题从第 1 个时间片到第 4 时间片在台湾时政类新闻数据集中都是热点主题, 从第 5 个时间片后就消亡了. 而 “马英九国情咨文” 报告从第 4 个时间片后成为热点主题, 并在第 6 个时间片达到顶峰, 人工判断在第 6 个时间片内, 新闻文本都是关于马英九是否去立法院做国情咨文的蓝绿两营的口水战, 但在第 7 个时间片却不存在了, 经分析原因是台湾时政类新闻数据集选择的问题, 该文本集是利用爬行器从台湾相关网上获得的, 覆盖性存在一定问题. 同时, “美牛事件” 在这 7 个时间片中的热度不是很大, 原因是台湾时政类新闻数据集中还存在 “台湾猪肉含瘦肉精” 和 “台湾立法修改瘦肉精含量” 等主题, 这两个主题与本文的 “美牛事件” 主题有一定的重叠, 造成了 “美牛事件” 的强度值不是很大, 这也说明 BPE-OLDA + Rect-OGRS 能够聚焦某个大话题的侧面, 如本文的 “美牛事件” 主要是 “美牛事件” 造成的台湾与美国之间的贸易摩擦.

综合图 2～4 可以看出, Rect-OGRS 算法虽然修正了主题下词的分布, 但并没有对文本下主题分布产生明显的作用; 而 BPE-OLDA 算法则提高了各个主题的强度值, 进而提高了文本中词被正确的主题生成的概率.

BPE-OLDA + Rect-OGRS 不同主题的强度演化情况对比示意图如图 5 所示.

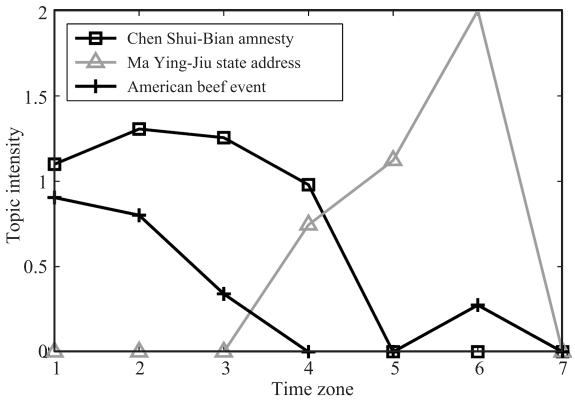

图 5 BPE-OLDA + Rect-OGRS 主题强度演化图

Fig. 5 Topic intensity evolution of BPE-OLDA + Rect-OGRS

3.2 困惑度的结果与分析

困惑度主要用来衡量 LDA 模型的泛化能力, 度量 LDA 模型对未知文本集的适用能力. 困惑度越小, 模型的泛化能力越好, 适用范围越广. 困惑度定义如下:

$$Perplexity(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log P(\boldsymbol{w}_d)}{\sum_{d=1}^{M}N_d}\right\} \quad (9)$$

图 6 是 OLDA + OGS、OLDA + Rect-OGRS 和 BPE-OLDA + Rect-OGRS 在台湾时政类新闻数据集中挖掘主题信息的困惑度走势图. 从图 6 可知, Rect-OGRS 算法可大幅度降低 OLDA 模型挖掘主题信息的困惑度. OGS 算法由于估算方式没有考虑到主题下词的相互关系, 造成所挖掘的主题与实际文本内容的贴合较差, 而 Rect-OGRS 很好地解决了这个问题. 另外, BPE-OLDA 模型通过引入主题强度演化, 提高了所挖掘主题贴合实际文本内容的程度. 同时 OLDA + OGS 的困惑度曲线走势比较缓, 是其本身困惑度值比较大造成的.

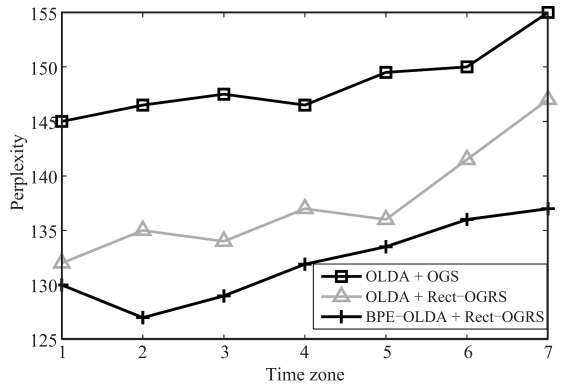

图 6　困惑度演化图

Fig. 6　Topic confusion evolution

4　结论与展望

本文提出了基于 LDA 的双通道在线主题演化模型 BEP-OLDA, 该模型针对时效性较强的数据流, 同时考虑了主题内容和强度遗传, 并采用简化 Gibbs 采样方法进行求解. 实验证明, 该模型对时效性较强的文本 (如时政新闻和论坛帖子) 有很好的效果. 经本文作者验证, 该方法在 NIPS 文本数据集等非时效性较强的文本上效果不明显. 因此, 该模型还有进一步改进的地方. 该模型对网络舆情的分析与处理, 尤其是对大量的时效性信息处理有很好的应用.

References

1 Wang Fei-Yue, Wang Jue. Intelligence and security informatics: the state of the art and outlook. *China Basic Science*, 2005, **7**(2): 24−29
(王飞跃, 王珏. 情报与安全信息学研究的现状与展望. 中国基础科学, 2005, **7**(2): 24−29)

2 Chen H C, Wang F Y, Zeng D. Intelligence and security informatics for homeland security: information, communication, and transportation. *IEEE Transactions on Intelligent Transportation Systems*, 2004, **5**(4): 329−341

3 Wang Fei-Yue. Decision service and academic analytics for development of science and technology based on open source intelligence and big data. *Bulletin of Chinese Academy of Sciences*, 2012, **27**(5): 527−537
(王飞跃. 知识产生方式和科技决策支撑的重大变革 — 面向大数据和开源信息的科技态势解析与决策服务. 中国科学院院刊, 2012, **27**(5): 527−537)

4 Zhang Chen-Yi, Sun Jian-Ling, Ding Yi-Qun. Topic mining for microblog based on MB-LDA model. *Journal of Computer Research and Development*, 2011, **48**(10): 1795−1802
(张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘. 计算机研究与发展, 2011, **48**(10): 1795−1802)

5 Yang Zhen, Lai Ying-Xu, Duan Li-Juan, Li Yu-Jian. Short text sentiment classification based on context reconstruction. *Acta Automatica Sinica*, 2012, **38**(1): 55−67
(杨震, 赖英旭, 段立娟, 李玉鑑. 基于上下文重构的短文本情感极性判别研究. 自动化学报, 2012, **38**(1): 55−67)

6 Yin Chun-Xia, Peng Qin-Ke. Identifying word sentiment orientation for free comments via complex network. *Acta Automatica Sinica*, 2012, **38**(3): 389−398
(殷春霞, 彭勤科. 利用复杂网络为自由评论鉴定词汇情感倾向性. 自动化学报, 2012, **38**(3): 389−398)

7 Li Wen-Qing, Sun Xin, Zhang Chang-You, Feng Ye. A semantic similarity measure between ontological concepts. *Acta Automatica Sinica*, 2012, **38**(2): 229−235
(李文清, 孙新, 张常有, 冯烨. 一种本体概念的语义相似度计算方法. 自动化学报, 2012, **38**(2): 229−235)

8 Xu R, Wunsch D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 2005, **16**(3): 645−678

9 Deerwester S, Dumais S T, Furnas G W, Landauer T K, Harshman R. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 1990, **41**(6): 391−407

10 Landauer T K, Foltz P W, Laham D. Indexing by latent semantic analysis. *Introduction to Latent Semantic Analysis*, 1998, **25**(2): 259−284

11 Griffiths T, Steyvers M. Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning*. Hillsdale, NJ: Laurence Erlbaum, 2006.

12 Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 1999. 50−57

13 Salton G, McGill M. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1986.

14 Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 2003, **3**: 993−1022

15 Zhou Jian-Ying, Wang Fei-Yue, Zeng Da-Jun. Hierarchical Dirichlet processes and their applications: a survey. *Acta Automatica Sinica*, 2011, **37**(4): 389−407
(周建英, 王飞跃, 曾大军. 分层 Dirichlet 过程及其应用综述. 自动化学报, 2011, **37**(4): 389−407)

16 Wei X, Croft W B. LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2006. 178−185

17 Blei D M, Lafferty J. *Text Mining: Classification, Clustering, and Applications*. New York: Chapman & Hall/CRC, 2009.

18 Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences. In: Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM, 2007. 233−240

19 Mei Q Z, Cai D, Zhang D, Zhai C X. Topic modeling with network regularization. In: Proceedings of the 17th International Conference on World Wide Web. New York, USA: ACM, 2008. 10l−ll0

20 Blei D M, Lafferty J D. Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. New York, USA: ACM, 2006. 113−120

21 Boyd-Graber J, Blei D M. Syntactic topic models. In: Proceedings of the 20th Neural Information Processing Systems. Cambridge, USA: MIT, 2008.

22 Nallapati R, Cohen W. Link-pLSA-LDA: A new unsupervised model for topics and influence of blogs. In: Proceedings of the 2008 International Conference on Weblogs and Social Media (ICWSM). Menlo Park, CA: AAAI, 2008.

23 Sun C K, Gao B, Cao Z F, Li H. HTM: A topic model for hypertexts. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. New York, USA: ACM, 2008. 514−522

24 Alusumait L, Barber D, Domeniconi C. On-Line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 2008 English IEEE International Conference on Data Mining. Pisa, Italy: IEEE, 2008. 3−12

25 Manning C D, Raghavan P, Schittze H. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2007. 117−119

26 Fgueiredo M, Jain A K. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**(3): 381−396

27 Ching J, Chen Y J. Transitional Markov chain monte carlo method for Bayesian model updating, model class selection, and model averaging. *Journal of Engineering Mechanics*, 2007, **133**(7): 816−832

28 Xu Xin, Shen Dong, Gao Yan-Qing, Wang Kai. Learning control of dynamical systems based on Markov decision processes: research frontiers and outlooks. *Acta Automatica Sinica*, 2012, **38**(5): 673−687
(徐昕, 沈栋, 高岩青, 王凯. 基于马氏决策过程模型的动态系统学习控制: 研究前沿与展望. 自动化学报, 2012, **38**(5): 673−687

29 Griffths T. Gibbs Sampling In the Generative Model of LDA [Online], available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.8022, November 17, 2014

曹建平 国防科学技术大学信息系统与管理学院博士研究生. 主要研究方向为文本分析, 平行系统理论. 本文通信作者. E-mail: caojianping@nudt.edu.cn
(**CAO Jian-Ping** Ph. D. candidate at the College of Information System and Management, National University of Defense Technology. His research interest covers text analysis and parallel management theory. Corresponding author of this paper.)

王 晖 国防科学技术大学计算实验与平行系统技术研究中心教授. 主要研究方向为多媒体情报分析与数据挖掘. E-mail: huiwang@nudt.edu.cn
(**WANG Hui** Professor at the Research Center of Computational Experiments and Parallel System Technology, National University of Defense Technology. His research interest covers multi-media intelligence analysis and data mining.)

夏友清 国防科学技术大学信息系统与管理学院硕士研究生. 主要研究方向为文本分析与检索. E-mail: xiayouqing12@163.com
(**XIA You-Qing** Master student at the College of Information System and Management, National University of Defense Technology. His research interest covers text analysis and retrieval.)

乔凤才 国防科学技术大学信息系统与管理学院博士研究生. 主要研究方向为图像分类与检索. E-mail: qiaofengcai125@gmail.com
(**QIAO Feng-Cai** Ph. D. candidate at the College of Information System and Management, National University of Defense Technology. His research interest covers image categorization and image retrieval.)

张 鑫 国防科学技术大学计算实验与平行系统技术研究中心副教授. 主要研究方向为计算机视觉, 机器学习和图像分析. E-mail: zhangxin@nudt.edu.cn
(**ZHANG Xin** Associate professor at the Research Center of Computational Experiments and Parallel System Technology, College of Information System and Management, National University of Defense Technology. His research interest covers computer vision, machine learning, and image analysis.)