图书情报工作
第58卷 第2期 2014年1月

基于动态LDA主题模型的内容主题挖掘与演化*

胡吉明 陈果

摘 要 指出文本内容主题的挖掘和演化研究对于文本建模和分类及推荐效果提升具有重要作用。从分析基于LDA主题模型的文本内容主题挖掘原理入手,针对当前网络环境下的文本内容特点,构建适用于动态文内容本主题挖掘的LDA模型,并通过改进的Gibbs抽样估计提高主题挖掘的准确性,进而从主题相似度和强度两个方面研究内容主题随时间的演化问题。实验表明,所提方法可行且有效,对后续有关文本语义建模和分类研究等具有重要的实践意义。

[关键词] 主题挖掘 主题演化 动态LDA模型
[分类号] G202
DOI: 10.13266/j.issn.0252-3116.2014.02.023

文本内容挖掘与语义建模是信息推荐和数据挖掘领域的研究热点与核心内容,而文本内容的主题挖掘则是语义建模的重要基础。当前网络环境下,信息内容具有呈动态交互和随时间发展演化等特征,因此要求创新信息内容挖掘方法,提升内容主题挖掘的准确性,动态描述其演化趋势。基于此,本文对传统潜在狄利克雷分布(LDA)主题模型进行动态化改进,运用增量Gibbs抽样估计算法,实现文本内容主题的准确挖掘;在文本时间片划分的基础上,基于主题相似度和强度度量,描述内容主题的时间演化趋势。本文研究对语义层次的信息内容建模以及提高内容描述的准确性具有重要作用。

1 引 言

文本内容的主题提取即选择合适的文本内容主题和特征词汇,以此对文本内容进行特征描述和建模。主题模型作为一种文本内容的概率生成模型或产生式模型,如潜在语义分析(LSA)[1]、概率潜在语义分析(PLSA)[2]和LDA[3],通过对人类思维过程的模拟,找到产生文本的最佳主题和词汇,能够最大程度地表示文本中所蕴含的含义,信息丢失较少,较好地解决了词汇、主题和文本之间的语义关联问题[4],是目前最常用的文本主题提取方法[5]。更重要的是,LDA主题模型基于产生式的三层贝叶斯概率计算得到通过潜在主题有限混合表示的文本,并且通过词汇表中所有词汇的概率分布来表示每个主题,文本内容则根据主题和词汇的混合分布来区分[6]。LDA主题模型采用Dirichlet分布简化了模型的推导过程,避免了LSA和PLSA模型产生的过拟合的问题[7],因此具有很好的先验概率假设,参数数量不会随着文本数量的增长而线性增长,泛化能力强,在算法复杂度和展示效果方面表现优越,广泛应用于文本主题挖掘、文本分类聚类、文本检索、内容主题演化等领域[8]。

近年来,网络信息内容主题的挖掘受到国内外研究者和机构的广泛关注,旨在准确捕捉网络信息内容的动态演化特征,跟踪或准确发现其发展变化趋势。如M. Mohd等设计了交互事件跟踪(iEvent)系统,以此发现用户交互所产生的热点内容主题[9]。C. Aksoy等构建了基于语言模型的新奇新闻检测系统BilNov-2005,实现了新奇新闻主题的动态实时挖掘[10]。余传明等基于LDA模型研究了用户评论内容主题和热点关键词的挖掘方法,实验表明该模型具有较好的热点主题识别效果[11]。刘洪涛等针对内容主题不明确和热点问题难以跟踪的问题,通过计算文献作者的舆论评价得到每个评价社区的关键词概率描述,实现了社区中评论主题的发现,对文本语义挖掘和共享等具有重要意义[12]。黄颖通过基于LDA和主题词的相关性新事件监测模型,结合报道发生的时间确定合理的主

* 本文系教育部人文社会科学青年基金项目"社会网络环境下信息内容主题挖掘与语义分类研究"(项目编号:13YJC870008)和国家自然科学青年基金项目"社会网络环境下基于用户-资源关联的信息推荐研究(项目编号:71303178)"研究成果之一。
[作者简介] 胡吉明,武汉大学信息资源研究中心讲师,E-mail: whuhujiming@qq.com;陈果,武汉大学信息资源研究中心博士研究生。
收稿日期:2013-11-13 修回日期:2014-01-04 本文起止页码:138-142 本文责任编辑:王传清

题数目以探知新事件[13]。

2 基于动态 LDA 的内容主题挖掘模型

网络环境下文本信息所具有的短文本结构特征加大了文本挖掘和表示的难度[14]，因此，本文在现有 LDA 主题挖掘基础上，结合微博、博客、社交网络等社会化网络服务中的交互式信息特点，构建动态 LDA 主题模型，按时间片划分文本信息，将增量 Gibbs 抽样算法引入其中，通过参数估计得到时间片文本集中连续的主题 - 词汇分布和文本 - 主题分布。

2.1 LDA 主题模型的动态化改进

首先采用滑动时间窗把文本划分到时间片内，时间片内的文本数根据其主题和词汇分布的不同而不同，且允许不同时间片内存在相同的文本(因文本存在主题交叉或相似现象)，组成文本时间片集；然后采用 LDA 主题模型对每个时间片文本集进行主题挖掘，提取出 T 个主题，运用增量 Gibbs 抽样算法[15]得出文本内容和主题之间的概率分布关系(文本 - 主题和主题 - 词汇)。进而对前一个时间片文本集中文本的主题 - 词汇概率分布关系加权处理(W)后，作为当前时间片文本集中主题 - 词汇分布的先验概率，求出随时间变化的主题 - 词汇和文本 - 主题概率分布，最终得到此文本内容主题的时间演化模式，如图 1 所示：

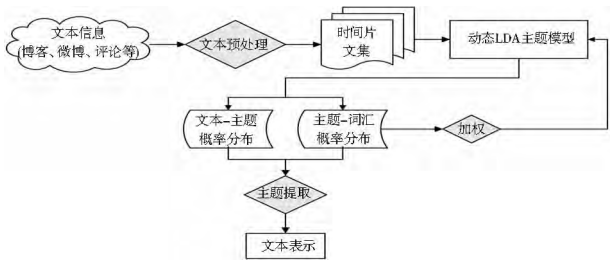

图 1 基于动态 LDA 主题模型的文本主题挖掘框架

在基于 LDA 主题模型进行文本主题提取的过程中，本文改进的重点是基于时间窗口将动态演化的文本按时间窗划分，按照文本内容主题的连续性和差异性，运用增量 Gibbs 抽样算法进行抽样计算。

首先，确立时间 t 内的文本集合 $D_t=\{d_1,d_2,\cdots,d_l\}$，时间窗大小根据用户需求、具体应用领域和文本分析的粗细粒度设定(M_t)。文本时间片一旦划分，则保证不同时间片内的文本不能交换，而同一时间片内的文本可以交换。其次，根据前一时间的主题 - 词汇分布的后验概率 φ_{t-1} 乘上权重 W($W=\frac{V_t W_U}{V_{t-1}}$，$V_t$ 为 t 时刻的词汇数，W_U 为用户自行设定的权重，本文认为当前时间片内的文本信息受到上一时间片文本信息的影响)作为当前时间片文本主题提取的先验概率 φ_t，从而建立动态 LDA 文本主题挖掘模型，如图 2 所示：

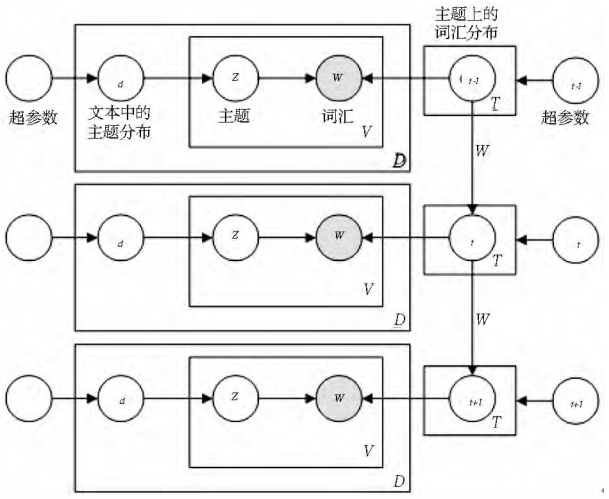

图 2 动态演化 LDA 文本生成模型

图 2 中，可直接观测变量(词汇)用实心圆表示，隐含的潜在变量(主题)用空心圆表示；图中矩形表示重复过程，大矩形表示从狄利克雷(Dirichlet)分布中为文本集中的每个文本 d 反复提取的主题分布 θ_d，小矩形则表示从主题分布中反复抽样产生的文本词汇 $\{w_1,w_2,\cdots,w_V\}$。

根据传统 LDA 模型的文本生成过程，动态 LDA 主题模型运算过程如图 3 所示：

动态演化 LDA 文本生成过程

首先将文本按照设定好的时间间隔划分为 t 个时间片文本集，每个时间片文本集内对应一个 φ_t 和 θ_d^t。

对每一个主题 $z_t \in \{z_1,z_2,\cdots,z_T\}$，选择第一个时间片文本集 $t=\{1,2,\cdots,t\}$

1 如果是第一个时间片 $t=1$，则 $\varphi_t \sim Dir(\beta_1)$

2 否则，计算 $\varphi_t=\varphi_{t-1}W$

3 对每一个文本 d，抽样计算 $\theta_d^t \sim Dir(\alpha)$

4 对文本中的每一个词汇 w_i

4.1 从多项式分布 θ_d^t 中抽样生成主题的概率分布 z_j：$P(z_j \mid \alpha^t)$

4.2 从多项式分布 φ_t 中抽样生成词汇的概率分布 w_i：$P(w_i \mid z_j,\beta_{z_j}^t)$

图 3 动态演化 LDA 文本生成过程

LDA 模型推理的依据就是文本生成过程的逆过程，根据文本的生成规则和已知参数，通过概率推导求得文本的主题结构；本文中所要推理的参数为时间片文本集内的主题 - 词汇概率分布 φ 和文本 - 主题分布 θ，Gibbs 抽样[16]是其常用且最有效的推导方法。基于此，本文增量改进原始 Gibbs 抽样算法[17]，并将其运用于 LDA 主题模型中实现其动态化运算。

2.2 基于增量吉布斯抽样估计的主题确定

本文在进行动态 LDA 模型构建时，首先引入先验加权，重新计算时间片 t 时刻的后验概率 P_t

$(z_i=j|z_{-i},w_i,d_i,\bullet)$，即目标函数的计算公式变为：

$$P_t(z_i=j|z_{-i},w_i,d_i,\bullet)=\frac{(n_{-i,j}^{(w_i)})_t+v(n_{-i,j}^{(w_i)})_{t-1}+\beta}{(n_{-i,j}^{(\bullet)})_t+v(n_{-i,j}^{(\bullet)})_{t-1}+V\beta}\cdot\frac{(n_{-i,j}^{(d_i)})_t+\alpha}{(n_{-i,\bullet}^{(d)})_t+T\alpha}\Big/\sum_{j=1}^{T}\frac{(n_{-i,j}^{(w_i)})_t+v(n_{-i,j}^{(w_i)})_{t-1}+\beta}{(n_{-i,j}^{(\bullet)})_t+v(n_{-i,j}^{(\bullet)})_{t-1}+V\beta}\cdot\frac{(n_{-i,j}^{(d_i)})_t+\alpha}{(n_{-i,\bullet}^{(d_i)})_t+T\alpha} \quad (1)$$

其中，$z_i=j$ 表示把主题 j 赋给词汇 w_i 作为其主题，$\bullet$ 表示其他所有已知的或可见的信息（如其他所有词汇 w_{-i} 和文本 d_{-i}，以及超参数 α 和 β），z_{-i} 表示当前词汇外的所有其他词汇的主题 $z_k(k\neq i)$ 赋值（即分配给 $z_k(k\neq i)$ 的词汇数），$v(n_{-i,j}^{(w_i)})_{t-1}$ 是上一时间片内分配给主题 z_j 与词汇 w_i 相同的词汇个数，$v(n_{-i,j}^{(\bullet)})_{t-1}$ 是上一时间片内分配给主题 z_j 的所有词汇个数，$n_{-i,j}^{(w_i)}$ 是分配给主题 z_j 与词汇 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 z_j 的词汇个数，$n_{-i,\bullet}^{(d_i)}$ 是文本 d_i 中所有被分配了主题的词汇个数，但所有的词汇个数去掉这次 $z_i=j$ 的分配。

因此，基于增量 Gibbs 抽样估计的主题确定步骤为：①马尔可夫链初始状态的确定。$z_i(z_i=j)$ 被初始化为从 1 到 T 之间的每个随机整数，i 从 1 循环到 V。②马尔可夫下一状态的获取。在 i 从 1 循环到 V 的过程中，根据公式（1）将词汇 w_i 分配给主题 z_j，获取马尔可夫链的下一个状态。③马尔可夫链稳态确定。不断迭代第②步，直到其概率分布趋于稳定即马尔可夫链接近目标函数分布（迭代次数取决于特定的文本集，依据为相邻的马尔可夫链状态的关联大小，一般为 300－400 次左右后就会趋于稳定），记录下 z_i 的当前值作为样本；其他的样本值随每次迭代达到稳态时不断记录，以保证马尔可夫链的自相关较小。

基于此，对于每一个时间片文本集，便可重新估算 φ 和 θ 的值，如公式（2）和公式（3）所示：

$$(\hat{\varphi}_w^{(z=j)})=\frac{(n_j^{(w)})_t+v(n_j^{(w)})_{t-1}+\beta}{(n_j^{(\bullet)})_t+v(n_j^{(\bullet)})_{t-1}+V\beta} \quad (2)$$

$$(\hat{\theta}_{z=j}^{(d)})=\frac{(n_j^{(d)})_t+\alpha}{(n_\bullet^{(d)})_t+T\alpha} \quad (3)$$

其中，词汇 w 表示唯一词汇，$n_j^{(\bullet)}$ 表示时间片文本集中分配给主题 z_j 的所有词汇数，$n_j^{(w)}$ 则表示时间片文本集中词汇 w 被分配给主题 z_j 的总次数，$n_j^{(d)}$ 表示时间片文本集中某个文本 d 中分配给主题 z_j 的所有词汇数，而 $n_\bullet^{(d)}$ 则表示文本 d 中所有被分配了主题的词汇数。

3　基于主题相似度和强度度量的主题演化

随着时间的发展，信息内容的主题和强度也会发生变化，表现为从开始到高潮再到衰落的过程，甚至循环往复。有效地组织大规模文本信息，并按时间顺序描述其主题的演化过程，从而帮助用户追踪所需求偏好的主题，具有实际意义。

文本主题随时间的演化主要从不同时间片的主题相似度和强度变化来衡量[18]。在基于动态 LDA 主题模型的文本挖掘和演化研究中，本文采用 KL 距离（Kullback-Leibler divergence）[19] 计算主题－词汇概率分布之间的相似度，观测时间片文本集中内容主题的差异，描绘主题随时间变化的脉络和趋势；与此同时，主题强度的变化采用主题在时间片文本集内所占的比例来衡量（θ 的平均值），从而得出时间片内内容主题强度的变化趋势。

3.1　基于主题相似性计算的演化

KL 距离是衡量两个主题概率密度分布差异最常用的度量标准，公式为：

$$D(P(w|s_1)\|P(w|s_1))=\sum_{w\in V}P(w|s_1)\log\frac{P(w|s_1)}{P(w|s_2)} \quad (4)$$

标准 KL 距离为非对称值且为非负，两者之间位置互换计算得出的 KL 距离值是不一样的。当 KL 为 0 时，表明两个主题的概率密度分布完全相同。在计算过程中，文本是按照时间片划分的，随着时间片的推移，文本不断加入文本集中，其数量将不断增加，而新的词汇和主题也将被引入。后续分析时基于假设“新词汇在之前时间片文本集中出现的次数为 0，只在当前时间片文本集出现”，并且词汇集在同时处理时间片文本集的过程中需要不断更新，从而可以在统一的概率分布空间中处理不同时间片内的主题－词汇概率分布，进而使得基于 KL 距离的主题相似度计算和比较更加方便。

3.2　基于主题强度计算的演化

主题强度表示文本主题受关注的程度，其演化过程可通过观测文本主题随时间变化的趋势来衡量。依据上述 Gibbs 抽样计算中的公式 3 获得的 $(\hat{\theta}_{z=j}^{(d)})$，求出当前的时间片文本集 t 中主题 z_j 的平均强度 $\bar{\theta}_t$：

$$\bar{\theta}_t=\frac{\sum_{i=1}^{M}\hat{\theta}_{z=j}^{(d)}}{M} \quad (5)$$

因此，可以根据 $\bar{\theta}_l$ 计算得出一系列时间片文本集中主题平均强度的不同值，绘出主题强度随时间的变化趋势图。主题强度变化趋势结合主题词（文本中概率值最大的词汇）和文本列表联合概率分布展示了该主题的具体含义，有利于主题价值的识别。

4 实验分析

本文采用中文文本分类语料库 TanCorpV1.0[19] 实验验证模型和方法的有效性，进行文本内容主题的挖掘和演化实现；利用汉语词法分析系统 ICTCLAS[20] 对文本集进行预处理后，以之作为本文实验的基础数据。在数据集中，选取科技类文本的前 60% 作为训练集，后 40% 作为预测集；在参数调整确定后，得到最终的实验结果。

为了有效利用增量 Gibbs 抽样算法，首先需确定 LDA 主题模型中 3 个变量（Dirichlet 分布中的超参数 α、β 以及主题数目 T）的最佳值。根据大量的文献调研[21]，本文令 $\beta=0.01$（根据其他研究实验效果及本次反复实验得出，此取值对实验效果较好），$\alpha=\frac{50}{T}$。这种取值在众多试验中具有较好的表现[22]。主题数目的取值对 LDA 模型的文本提取和拟合性能影响较大，其最佳值的确定主要通过两种方式：一种是词汇被选中的概率 $P(w|T)$，一种是困惑度（$perplexity(D)$）[23]。LDA 主题模型的困惑度是从模型泛化能力衡量 LDA 主题模型对于新文本的预测能力，困惑度越小表示模型的泛化能力越强。因其能够较为全面地评测模型的效果，本文将其作为评测指标，计算公式为：

$$perplexity(D_{test}|M)=exp\left\{-\frac{\sum_{l=1}^{M}log(P(d_l))}{\sum_{l=1}^{M}N_l}\right\} \quad (6)$$

其中，M 为测试文本数，N_l 为文本 d_l 的长度，$P(d_l)$ 为 LDA 主题模型产生文本 d_l 的概率。

对 T 的不同取值，分别运行 Gibbs 抽样算法，迭代次数为 300，观测困惑度取值的变化情况。从图 4 可以看出，困惑度值随着主题数的增加而变小，在 $T=50$ 时为最小值，而此时模型的性能达到最优，可见对于此文本集而言最佳主题数为 50。

本文在进行主题演化的试验中，将权重 W 设为 0.3，只标识每个主题的前 10 个词汇，以时间片为顺序描述主题的概率分布变化情况。

从图 5 可以看出 3 个主题随时间的演化情况：主题 z_1 在此文本集中具有较高的强度，且在大部分时间片中较为稳定；主题 z_2 和主题 z_3 在前部分时间片波动较大，后期时间片文本集内的内容主题变化逐渐趋于平稳。因此，本文对 LDA 主题模型的动态化改进在基于时间片分布的文本内容挖掘中具有可行性和有效性，且能够较好地描述主题随时间的演化情况，对网络环境下动态文本内容主题的准确挖掘、随时间的演化趋势描述甚至文本建模具有一定的实际意义。

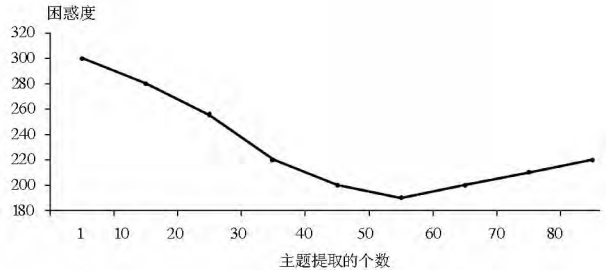

图 4 LDA 主题提取中的困惑度

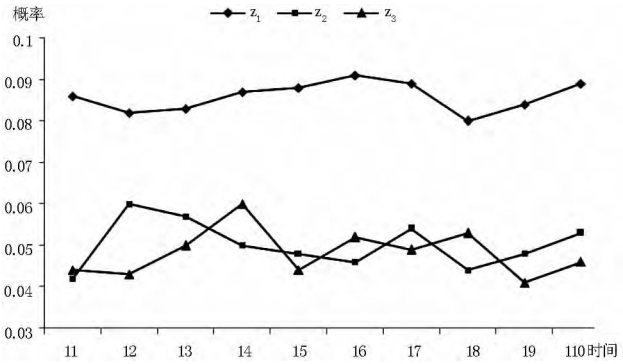

图 5 时间片文本集中文本内容主题的演化趋势

5 结 语

针对原始 LDA 模型中忽略文本时间信息而无法描述文本主题的演化问题，本文进行了动态 LDA 主题模型的构建。首先将文本集按照时间片划分，构建时间片文本集，对每一个时间片文本集进行主题提取，并将主题按照时间进行内容和强度两个方面的演化分析。为了适应文本主题的动态提取，本文对原始的 Gibbs 抽样算法进行增量改进，在不同时间片之间设置权重，对每一个时间片文本集进行参数 φ 和 θ 的估算。最后采用 KL 距离相似度计算主题－词汇概率分布之间的相似度，分析主题变化的差异；采用 θ 的平均值来代表时间片内主题的强度或主题在时间片内所占的比重，从而描绘主题在内容和强度上随时间的变化趋势。主题提取及演化的实验表明，本文所提方法可行且有效。因此，下一步将以此为基础进行文本语义建模等研究，提高文本描述的准确性。

参考文献：

[1] Deerwester S，Dumais S T，Furnas G W，et al. Indexing by latent

semantic analysis [J]. Journal of the American Society for Information Science, 1990,114(2):211 - 244.

[2] Hofmann T. Probabilistic latent semantic analysis [C]//Proceedings of the Twenty-Second Annual International SIGIR, Conference on Research and Development in Information Retrieval. New York: ACM,1999:50 - 57.

[3] Blei D M, Ng A Y, Jordan M L, et al. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003,3(2):993 - 1022.

[4] Blei D M. Probabilistic topic models [J]. Communications of the ACM,2012,55(4):77 - 84.

[5] Barbieri N, Manco G, Ritacco E, et al. Probabilistic topic models for sequence data [J]. Machine Learning,2013,93(1):5 - 29.

[6] Isaly L, Trias E, Peterson G. Improving the latent Dirichlet allocation document model with WordNet [C]//Proceedings of the 5th International Conference on Information Warfare and Security. London: Academic Conferences Ltd,2010:163 - 170.

[7] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning,2001,42(1):177 - 196.

[8] Du Lan, Buntine W, Jin Huidong, et al. Sequential latent Dirichlet allocation [J]. Knowledge and Information Systems, 2012,31(3):475 - 503.

[9] Mohd M, Crestani F, Ruthven I. Evaluation of an interactive topic detection and tracking interface [J]. Journal of Information Science,2012,38(4):383 - 398.

[10] Aksoy C, Can F, Kocberber S. Novelty detection for topic tracking [J]. Journal of The American Society for Information Science and Technology,2012,63(4):777 - 795.

[11] 余传明,张小青,陈雷,等. 基于 LDA 模型的评论热点挖掘: 原理与实现 [J]. 情报理论与实践,2010,33(5):103 - 106.

[12] 刘洪涛,肖开洲,吴渝,等. 带舆论评价的引文网络构建与主题发现 [J]. 情报学报,2011,30(4):441 - 448.

[13] 黄颖. LDA 及主题词相关性的新事件检测 [J]. 计算机与现代化,2012(1): 6 - 9,13.

[14] Kang J H, Lerman K, Plangprasopchok A. Analyzing microblogs with affinity propagation [C]//Proceedings of KDD Workshop on Social Media Analytics. New York: ACM,2010:67 - 70.

[15] Gohr A, Hinneburg A, Schult R, et al. Topic evolution in a stream of documents [C]//Proceeding of the Society for Industrial and Applied Mathematics. Washington: National Academy of Science, 2009:859 - 870.

[16] Griffiths T L, Steyvers M. Finding scientific topics [C]//Proceedings of the National Academy of Science. Washington: National Academy of Sciences, 2004:5228 - 5235.

[17] Walsh B. Markov chain monte carlo and Gibbs sampling [EB/OL]. [2014 - 01 - 05]. http://web.mit.edu/~wingated/www/introductions/mcmc-gibbs-intro.pdf.

[18] 楚克明. 基于 LDA 的新闻话题演化研究 [D]. 上海: 上海交通大学,2010.

[19] 谭松波,王月粉. 中文文本分类语料库 - TanCorpV1.0 [EB/OL]. [2011 - 11 - 10]. http://www.searchforum.org.cn/tansongbo/corpus.htm.

[20] 中国科学院计算技术研究所. ICTCLAS2011 [EB/OL]. [2010 - 12 - 21]. http://ictclas.org/ictclas_download.aspx.

[21] Guo Xin, Xiang Yang, Chen Qian, et al. LDA-based online topic detection using tensor factorization [J]. Journal of Information Science,2013,39(4): 459 - 469.

[22] 单斌,李芳. 基于 LDA 话题演化研究方法综述 [J]. 中文信息学报,2010,24(6):43 - 49,68.

[23] Cao Juan, Xia Tian, Li Jintao, et al. A density-based method for adaptive LDA model selection [J]. Neurocomputing, 2009,72(7 - 9): 1775 - 1781.

Mining and Evolution of Content Topics Based on Dynamic LDA

Hu Jiming　Chen Guo

Center for Studies of Information Resources, Wuhan University, Wuhan 430072

[Abstract] The study of mining and evolution of text topics is of important significance for text modeling and classification, as well as the recommendation service. Starting from the analysis of theory of text topic modeling based on LDA, aiming at dynamic characters of text contents under social networking environment, this article constructed a dynamic LDA model for mining of text topics. Subsequently, the accuracy degree of topic mining was improved by incremental Gibbs sampling and estimation. Furthermore, the evolution of dynamic topics of text contents was achieved from the aspects of topic similarity and intensity. The experiment demonstrated that methods proposed in this article were feasible and effective, which will be the foundation of further study about semantic modeling and classification text.

[Keywords] topics mining　topics evolution　dynamic LDA model