一种话题演化建模与分析方法

胡艳丽 [1] 白 亮 [1] 张维明 [1]

摘 要 根据时序关系将文本流划分为连续时间片中的文本集, 在线抽取各时间片中隐含的子话题, 采用模型选择方法动态确定各时间片包含的子话题数, 以历史时间片的子话题信息作为当前子话题发现的先验知识, 基于 OLDA (Online latent Dirichlet allocation) 模型抽取各时间片包含的子话题, 通过 Gibbs 抽样对话题模型参数进行估计; 对子话题进行关联分析, 定义子话题产生、消亡、继承、分裂和合并五种演化类型, 提出基于相对熵的子话题关联分析方法, 根据子话题语义相似度和时序关系建立子话题间的关联, 由具有时序关系和内容关联的子话题组成话题, 通过子话题内容和强度的变化描述话题演化. 基于真实网络新闻的话题演化分析实验表明, 本文提出的话题演化分析方法能够有效检测网络新闻话题内容和强度的演化.

关键词 话题演化, OLDA 模型, 模型选择, Gibbs 抽样, 相对熵, 关联分析

引用格式 胡艳丽, 白亮, 张维明. 一种话题演化建模与分析方法. 自动化学报, 2012, **38**(10): 1690−1697

DOI 10.3724/SP.J.1004.2012.01690

Modeling and Analyzing Topic Evolution

HU Yan-Li[1] BAI Liang[1] ZHANG Wei-Ming[1]

Abstract Topic evolution of network public opinions is investigated. By treating topics as a set of correlated sub-topics, a topic evolution model is proposed, consisting of sub-topic detection and correlation analysis. Furthermore, a sub-topic detection algorithm based on OLDA is presented with Bayesian model selection for the appropriate topic numbers and parameters estimation via Gibbs sampling. The correlations are further defined for analysis of topic evolution, including emergence, extinction, development, merge and split of sub-topics. The method is experimentally verified to be efficient for detecting topic evolution of network public opinions.

Key words Topic evolution, online latent Dirichlet allocation (OLDA), model selection, Gibbs sampling, relative entropy, correlation analysis

Citation Hu Yan-Li, Bai Liang, Zhang Wei-Ming. Modeling and analyzing topic evolution. *Acta Automatica Sinica*, 2012, **38**(10): 1690−1697

互联网和 Web 2.0 技术在全球范围内的迅猛发展引发了一场影响深远的媒体革命. 网络以其表达的自由性、匿名性、交互性和跨时空性等特性为社会成员提供了空前的话语权, 逐渐成为人们发布信息和表达观点的主要载体. 2009 年国家自然科学基金委员会管理学部和信息学部联合设立重大研究计划项目 “非常规突发事件的应急管理”, 其中将网络信息处理作为核心科学问题之一.

话题 (Topic) 是事件相关报道的集合[1]. 网络信息来源多样, 其中可能蕴含着关系公共安全、社会稳定的热点、敏感话题, 受事件发展和社会、历史、文化等诸多因素的影响, 其内容和受关注程度处于动态变化之中. 话题演化包括话题内容演化和强度演化两方面[2]: 话题内容演化是指话题内容随时间推移发生变化, 话题强度演化表示话题受关注程度的变化.

与话题演化相关的研究包括话题发现和跟踪 (Topic detection and tracking, TDT) 技术. TDT 的研究始于 1996 年, 初衷是自动发现新闻报道流中的话题, 进而按话题组织各种事件及其相应的报道[3]. 但 TDT 早期的研究没有充分利用语料的时间信息研究话题随时间的演化.

近年来, 统计话题模型得到了深入研究. 普林斯顿大学的 Blei 等首先提出了 LDA (Latent Dirichlet allocation) 模型[4]: 该模型是一种具有文本话题建模能力的概率生成模型 (Generative model), 假设文本的隐含语义结构由一组相互关联的话题组成, 话题则由一组关键词组成, 通过隐含的话题建立文本和词间的关联, 将文本从 N 维的词汇空间降维到

收稿日期 2011-05-13 录用日期 2012-04-28
Manuscript received May 13, 2011; accepted April 28, 2012
国家自然科学基金 (60902094, 60903225, 41001260), 高等学校博士学科点专项科研基金 (20114307110008) 资助
Supported by National Natural Science Foundation of China (60902094, 60903225, 41001260) and Research Fund for the Doctoral Program of Higher Education of China (20114307110008)
本文责任编委 宗成庆
Recommended by Associate Editor ZONG Cheng-Qing
1. 国防科学技术大学信息系统工程重点实验室 长沙 410073
1. Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073

K 维的话题空间, 能够有效地解决由于文本高维特性带来的 “维灾”, 较传统的向量空间模型更加有效. 其他的概率生成模型通常假设一个文本只涉及一个话题, 难以对现实中一个文本通常涉及多个话题的情况有效建模. 基于上述优点, LDA 模型在文本表示和文本挖掘中得到广泛应用.

为了研究话题随时间的变化, 研究人员通过引入时间信息对 LDA 模型进行扩展. TOM (Topic over model) 模型[5] 通过考虑文本的时间属性计算话题在时间上的分布强度. 动态话题模型 (Dynamic topic model, DTM)[6] 采用状态空间记录话题内容和分布强度的变化. 连续时间的动态话题模型 (Continuous time dynamic topic model, CTDMT)[7] 采用布朗运动模型对连续时间上的话题演化进行建模. MTTM (Multi-scale topic tomography) 模型[8] 研究多时间粒度的话题演化. 但上述扩展模型采用 LDA 模型对整个文本集进行全局建模, 无法增量处理新到达的文本, 导致较大的时间和空间开销. 为解决上述问题, 研究人员提出了具有增量处理能力的话题模型动态混合模型 (Dynamic mixture model, DMM)[9] 对严格按时间顺序到达的文本进行话题发现. 增量 LDA (Incremental latent Dirichlet allocation, ILDA) 模型[10] 只考虑话题上词分布在不同时间片间的关联, 获取话题内容的演化. OLDA (Online latent Dirichlet allocation) 模型[11] 考虑了不同时间片间词分布的关联, 并且话题数是固定的.

近年来, 国内基于统计话题模型的话题分析研究也逐步展开. 石晶等基于 PLSA 模型和 LDA 模型进行文本分割, 提取片段主题词[12−14], 但未研究话题演化问题. 楚克明等基于 LDA 模型进行话题抽取, 定义话题相似度和散度, 但不考虑不同时间片间的联系[15−16]. 崔凯等提出基于 LDA 的在线主题演化模型[17], 只考虑了不同时间片间话题所含关键词的联系. 2011 年 Wang 等将协同过滤 (Collaborative filtering) 用于话题建模, 利用在线社会网络中兴趣相似用户群体的推荐进行信息过滤、筛选, 以提高话题建模的准确性[18].

本文研究网络话题演化问题, 建立在线话题演化分析模型, 抽取网络信息中隐含的子话题, 建立子话题随时间推移的关联, 分析话题内容和强度随时间的演化.

1　网络话题演化分析建模

网络舆情监测过程中获取的网络信息是具有时序关系的文本流. 为了更好地理解话题的发展变化, 本文采用一组具有时序关系和内容关联的子话题描述话题, 一个子话题对应话题的一个片断, 在特定时间对话题某个侧面的报道形成话题的一个子话题. 因此, 网络话题演化分析归结为单个时间片中的子话题发现和不同时间片间的子话题关联分析, 如图 1 所示.

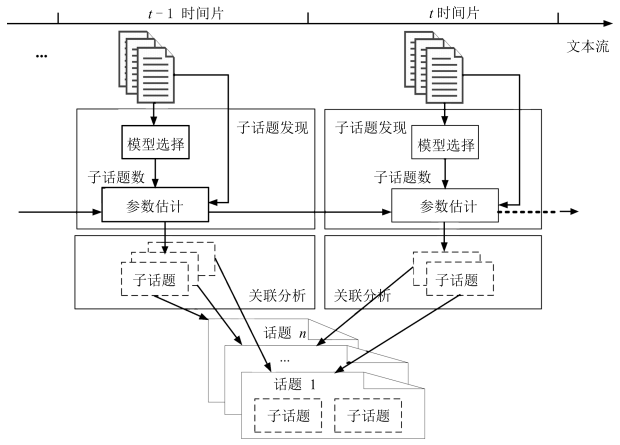


图 1　话题演化分析框架

Fig. 1　Analysis process of topic evolution

根据舆情分析的粒度设置时间片, 将文本流划分为连续时间片中具有时序关系的文本集. 由于网络信息的不确定性, 各时间片包含的子话题数往往是动态变化的, 因此框架中采用模型选择方法动态确定各时间片涉及的子话题数, 进而通过参数估计求解模型参数, 得到各时间片包含的一组子话题. 在此基础上, 通过关联分析将相关的子话题组成话题, 由各时间片中相互关联的子话题共同构成一个话题, 通过组成同一话题的子话题随时间的变化分析话题内容和强度随时间的演化.

网络话题演化分析中时间片的划分与实际应用密切相关, 并对话题发现以及演化分析的结果产生影响. 互联网中各种形式的舆情信息通常具有确定的、细粒度的时间信息, 如网络新闻发布时间、论坛发帖/回帖时间、聊天时间等, 可以支持不同时间粒度的话题演化分析. 网络舆情监测过程中通常根据舆情分析的需要决定演化分析的粒度, 如以 “天” 为单位进行分析.

2　基于 OLDA 模型的子话题发现

假设时间片 t 中文本集 D 包含 M 个文本, 涉及 K 个子话题. 文本 $d \in D$ 中词汇 w 的出现概率由子话题的出现概率和词汇 w 在子话题中的出现概率共同决定:

$$P(w) = \sum_{k \in [1,K]} P(w|\boldsymbol{z} = k)P(\boldsymbol{z} = k) \qquad (1)$$

其中, $\boldsymbol{z}$ 是 K 维子话题向量, $\boldsymbol{z}=k$ 表示向量 $\boldsymbol{z}$ 的第 k 维为 1, 即对应子话题 k. $P(\boldsymbol{z}=k)=\theta_k^{(d)}$ 表示文本 d 中子话题 k 的出现概率, 描述子话题 k 的强度; $P(w|\boldsymbol{z}=k)=\varphi_w^{(k)}$ 表示子话题 k 中词汇 w 的出现概率, 满足 $\sum_{k\in[1,K]}\theta_k^{(d)}=1$ 且 $\sum_{w\in V}\varphi_i^{(k)}=1$.

引入 LDA 模型[4] 假设, 令文本中子话题的出现服从参数为 $\boldsymbol{\theta}^{(d)}=(\theta_1^{(d)},\cdots,\theta_K^{(d)})^{\mathrm{T}}$ 的多项分布 (以下简称子话题分布), 记作:

$$\boldsymbol{z}\sim\mathrm{Multinomial}(\boldsymbol{\theta}^{(d)}) \tag{2}$$

类似地, 词汇集中的词汇在子话题 k 中的出现服从参数为 $\boldsymbol{\phi}^{(k)}=(\phi_1^{(k)},\cdots,\phi_N^{(k)})^{\mathrm{T}}$ 的多项分布 (以下简称词分布), 记作:

$$w|\boldsymbol{z}=k\sim\mathrm{Multinomial}(\boldsymbol{\phi}^{(k)}) \tag{3}$$

假设子话题分布和词分布的先验分别服从参数为 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 的 Dirichlet 分布, 分别记作:

$$\boldsymbol{\theta}^{(d)}\sim\mathrm{Dirichlet}(\boldsymbol{\alpha}) \tag{4}$$

$$\boldsymbol{\phi}^{(k)}\sim\mathrm{Dirichlet}(\boldsymbol{\beta}) \tag{5}$$

其中, $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 分别表示各子话题以及子话题中词汇在取样前的权重分布向量, $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_K)^{\mathrm{T}}$, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_N)^{\mathrm{T}}$, 且 $\alpha_k>0$ ($k\in[1,K]$), $\beta_i>0$ ($i\in[1,N]$).

先验知识的确定对后验求解的准确性具有重要影响, 也是应用话题模型进行子话题发现的难点之一. 由于已出现的话题可能在后续时间片中再次出现, 因此可以利用历史时间片中的词分布的后验为当前时间片的子话题发现提供先验知识. 对于当前时间片 t, 以时间片 $t-1$ 中词分布的加权作为时间片 t 中相应分布的先验, 即时间片 t 中词分布的 Dirichlet 先验满足[11]:

$$\boldsymbol{\beta}^t=\Psi^{t-1}*\boldsymbol{\omega} \tag{6}$$

其中, Ψ^{t-1} 是 $|V|\times K$ 维矩阵, 每一列对应时间片 $t-1$ 中的一个词分布, $\boldsymbol{\omega}$ 为 $K\times 1$ 维权重向量.

由于各时间片中的子话题数由对应的文本集决定, 通常是动态变化的. 因此子话题发现采用模型选择的方法根据各时间片对应的文本集动态确定子话题数[19]. 令时间片 t 中的子话题数为 K, 矩阵 Ψ^{t-1} 的设置如下: 若时间片 $t-1$ 中的子话题数大于 K, 则选择强度最大的前 K 个词分布构成 Ψ^{t-1}; 否则, 若时间片 $t-1$ 中的子话题数小于 K, 则除上述 K 个词分布外, 设置 Ψ^{t-1} 中的其余列为词汇集上的均匀分布.

因此, 基于 OLDA 模型, 本文提出增强 LDA 模型 (Enhanced LDA, ELDA), 进行子话题发现的基本步骤如下:

步骤 1. 对于当前时间片 t, 根据式 (6) 确定子话题分布和词分布的先验分布; 若 $t=1$, 则根据先验知识确定子话题分布和词分布的先验, 或者将子话题分布和词分布初始化为均匀分布.

步骤 2. 通过模型选择[20] 确定时间片 t 的子话题数 K.

步骤 3. 对文本集 D^t 进行抽样, 通过参数估计[21] 获得时间片 t 中子话题分布和词分布的参数 $\boldsymbol{\theta}^t$ 和 $\boldsymbol{\phi}^t$.

3 基于相对熵的子话题关联分析

3.1 子话题相似性度量

相对熵又称 KL 散度 (Kullback-Leibler divergence) 或 KL 距离[22], 表示概率分布间的差异性. 对于概率分布 $\boldsymbol{P}=(p_1,\cdots,p_n)$ 和 $\boldsymbol{Q}=(q_1,\cdots,q_n)$, $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 间的相对熵定义为

$$KL(\boldsymbol{P}\|\boldsymbol{Q})=\sum_{i\in[1,n]}p_i\log\frac{p_i}{q_i} \tag{7}$$

相对熵是不对称的, 即 $KL(\boldsymbol{P}\|\boldsymbol{Q})\neq KL(\boldsymbol{Q}\|\boldsymbol{P})$. 而子话题的语义相似性应该是对称的, 即对于任意子话题 T_i^t 和 T_l^m, T_i^t 和 T_l^m 的相似性与 T_l^m 和 T_i^t 的相似性相等. 因此基于相对熵的子话题相似度定义如下:

$$\begin{aligned}Sim(T_i^t,T_l^m)=&\\ -\frac{1}{2}KL(T_i^t\|T_l^m)-\frac{1}{2}KL(T_i^t\|T_l^m)=&\\ -\frac{1}{2}\sum_{w\in V}p_w\log\frac{p_w}{q_w}-\frac{1}{2}\sum_{w\in V}q_w\log\frac{q_w}{p_w}&\end{aligned} \tag{8}$$

其中, $p(w)$ 和 $q(w)$ 分别表示特征词汇 w 在子话题 T_i^t 和 T_l^m 中的出现概率.

3.2 子话题关联分析

组成同一话题的相关子话题可能出现在相邻时间片中, 也可能出现在相隔若干中间时间片的时间片中. 令滑动窗口包含 n 个时间片, 对于时间片 t 中的子话题, 考虑其与时间片 t 相邻滑窗内子话题间的关联.

定义 1. 令滑动窗口包含 n 个时间片, 对于时间片 t 中的子话题 T_i^t, 滑窗内各时间片 i ($i\in[t-1,t-n]$) 中与 T_i^t 相似度最大的子话题称为 T_i^t 的前向关联子话题, 记作 $prior(T_i^t)$.

定义 2. 令滑动窗口包含 n 个时间片, 对于时间片 t 中的子话题 T_i^t, 时间片 i $(i \in [t+1, t+n])$ 中与 T_i^t 相似度最大的子话题称为 T_i^t 的后向关联子话题, 记作 $post(T_i^t)$.

借鉴文献 [23] 的思想, 根据子话题与其前向和后向关联子话题间的关系, 可以将子话题的演化分为产生、消亡、继承、分裂和合并等多种类型, 定义如下:

定义 3 (子话题产生). 对于子话题 T_i^t, 若不存在前向关联子话题 T_l^m 使得 T_i^t 和 T_l^m 的相似度大于阈值 ε, 即不满足 $Sim(T_i^t, T_l^m) \geq \varepsilon$, 则 T_i^t 是在时间片 t 中产生的新子话题.

定义 4 (子话题消亡). 对于子话题 T_i^t, 若不存在后向关联子话题 T_l^m 使得 T_i^t 和 T_l^m 的相似度大于阈值 ε, 即不满足 $Sim(T_i^t, T_l^m) \geq \varepsilon$, 则 T_i^t 在时间片 t 中消亡.

定义 5 (子话题继承). 对于子话题 T_i^t, 若存在前向关联子话题 $T_l^m = prior(T_i^t)$ 使得 T_i^t 和 T_l^m 的相似度大于阈值 ε, 即 $Sim(T_i^t, T_l^m) \geq \varepsilon$, 并且 T_i^t 也是 T_l^m 的后向关联子话题, 即 $T_i^t = post(T_l^m)$, 则 T_i^t 是 T_l^m 的后继.

定义 6 (子话题分裂). 对于子话题 T_i^t, 若存在前向关联子话题 $T_l^m = prior(T_i^t)$ 使得 T_i^t 和 T_l^m 的相似度大于阈值 ε, 即 $Sim(T_i^t, T_l^m) \geq \varepsilon$, 但 T_i^t 不是 T_l^m 的后向关联子话题, 即 $T_i^t \neq post(T_l^m)$, 且存在子话题 T_k^r, 使得 $Sim(T_k^r, T_l^m) \geq \varepsilon$, 则称 T_i^t 和 T_k^r 是 T_l^m 的分裂.

定义 7 (子话题合并). 对于子话题 T_i^t, 若存在后向关联子话题 $T_l^m = post(T_i^t)$ 使得 T_i^t 和 T_l^m 的相似度大于阈值 ε, 即 $Sim(T_i^t, T_l^m) \geq \varepsilon$, 但 T_i^t 不是 T_l^m 的前向关联子话题, 即 $T_i^t \neq prior(T_l^m)$, 且存在子话题 T_k^r, 使得 $Sim(T_k^r, T_l^m) \geq \varepsilon$, 则称 T_l^m 是 T_i^t 和 T_k^r 的合并.

对于任意时间片 t 中的子话题 T_i^t, 分析 T_i^t 与时间片 t 相邻滑窗内的子话题间的关系, 根据相似度建立子话题间的关联. 基于相对熵进行子话题关联分析的基本步骤如下:

步骤 1. 对于当前时间片 t 中的子话题 T_i^t, 计算子话题 T_i^t 在滑窗内各时间片 $[t-1, t-n]$ 相似度大于阈值 ε 的前向关联子话题 T_l^m, 以及 T_l^m 的后向关联子话题 T_k^r.

步骤 2. 根据子话题 T_i^t、T_l^m 与 T_k^r 间的关系建立子话题间的关联:

1) 如果 T_i^t 和 T_k^r 为同一子话题, 则 T_i^t 是 T_l^m 的后继, 相应的将 T_i^t 并入 T_l^m 的后向关联子话题集, 并将 T_l^m 并入 T_i^t 的前向关联子话题集;

2) 如果 T_i^t 和 T_k^r 不为同一子话题, T_i^t 和 T_k^r 是 T_l^m 的分裂, 将 T_i^t 和 T_k^r 并入 T_l^m 的后向关联子话题集, 并将 T_l^m 分别并入 T_i^t 和 T_k^r 的前向关联子话题集.

步骤 3. 对于时间片 $t-n$ 中的子话题 T_i^{t-n}, 计算子话题 T_i^{t-n} 在滑窗内各时间片 $[t-1, t-n]$ 的后向关联子话题 T_l^m, 以及 T_l^m 的前向关联子话题 T_k^r, 若存在子话题 T_l^m 与 T_i^t 的相似度大于阈值 ε 且 $T_i^{t-n} \neq T_k^r$, 则 T_l^m 是 T_i^{t-n} 和 T_k^r 的合并, 分别将 T_l^m 并入 T_i^{t-n} 和 T_k^r 的后向关联子话题集, 并将 T_i^{t-n} 和 T_k^r 并入 T_l^m 的前向关联子话题集.

关联分析过程中, 每个子话题 T_i^t 的前向关联和后向关联子话题集初始化为空, 若算法结束时, T_i^t 的前向关联子话题集仍为空, 则表明该子话题不存在前向关联子话题, 即为新产生的子话题; 若 T_i^t 的后向关联子话题集为空, 则表明该子话题不存在后向关联子话题, 即子话题 T_i^t 消亡.

4 实验与分析

实验采用 NIPS 标准数据集和从网易新闻采集的 2010 年网络新闻进行话话题演化分析. 实验在 IBM XSERIES226 服务器上部署实施, 该服务器 CPU 为 Intel(R) XEON(TM) 3.00 GHz, 内存为 2.00 GB. 实验设置权重向量 $\boldsymbol{\omega}$ 中每个元素的取值为 0.5, 相似度阈值 $\varepsilon = -2.0$.

4.1 基于 NIPS 数据集的话题发现

NIPS 数据集[24] 是目前用于文本分析测试的标准数据集之一, 该数据集包含 1988 年 ~ 2000 年神经信息处理系统 (Neural information processing systems, NIPS) 会议论文集共计 1 866 篇文献, 去除停用词、数词和出现次数少于 5 次的词, 包含 2 532 891 个词例 (Word tokens), 14 972 个唯一性词汇.

将 NIPS 语料划分为 13 个时间片, 各时间片包含的文本集由历届 NIPS 会议录用发表的论文构成. 采用 ELDA 方法进行话题发现, 得到话题 11 与强化学习 (Reinforcement learning) 相关, 选择各时间片中该话题出现概率最大的一组关键词如表 1 所示, 并与 OLDA 方法所得结果进行比较.

由表 1 可见, ELDA 方法得到的强化学习相关话题均包含关键词 Learning, 且其中有 10 个时间片中关键词 Reinforcement 和 Learning 共同出现. 相比而言, OLDA 方法仅有 6 个时间片中包含关键词 Reinforcement 和 Learning.

对支持向量机 (Support vector machine, SVM)

表 1 ELDA 与 OLDA 方法比较

Table 1 Comparison of ELDA and OLDA on NIPS corpus

(a) ELDA: Topic 11 reinforcement learning

88 :	information	training	state	learning	algorithm
89 :	state	learning	reinforcement	internal	algorithm
90 :	control	model	reinforcement	learning	arm
91 :	learning	task	state	tasks	rule
92 :	learning	state	reinforcement	task	policy
93 :	learning	state	reinforcement	policy	control
94 :	learning	state	policy	action	reinforcement
95 :	learning	state	reinforcement	policy	action
96 :	learning	state	control	policy	reinforcement
97 :	learning	state	policy	reinforcement	control
98 :	state	learning	policy	reinforcement	action
99 :	state	learning	policy	action	reinforcement
00 :	state	learning	policy	action	time

(b) OLDA: Topic 12 reinforcement learning

88 :	state	learning	system	states	time
89 :	node	system	state	rule	learning
90 :	state	learning	rule	system	node
91 :	learning	state	reinforcement	system	world
92 :	state	learning	action	task	exploration
93 :	learning	state	reinforcement	control	dynamic
94 :	learning	state	optimal	control	dynamic
95 :	learning	state	optimal	action	policy
96 :	learning	state	action	policy	reinforcement
97 :	learning	state	action	reinforcement	time
98 :	state	learning	policy	reinforcement	optimal
99 :	state	learning	policy	action	reinforcement
00 :	state	learning	policy	action	reward

及其相关话题, 我们同样发现 ELDA 方法抽取的关键词较 OLDA 方法能够更清晰地表达话题的语义. 根据模型选择方法, ELDA 方法根据文本内容动态确定话题数, 采用使抽样概率最大化的话题数进行话题发现, 即根据文本集抽样使得文本所涉及话题对应的词汇出现概率最大, 因此抽取的话题内容更加准确.

实验进一步根据困惑度 (Perplexity)[4] 对 ELDA 模型与 OLDA 模型进行了分析:

$$perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum\limits_{d\in D_{\text{test}}} \log P(w_d)}{\sum\limits_{d\in D_{\text{test}}} N_d}\right\} \tag{9}$$

对于测试语料 D_{test}, N_d 为文本 d 包含的词汇数, w_d 为文本 d 包含的词例.

给定语言模型, 困惑度可以理解为每个词后续可能接的词的平均数量. 因此, 困惑度越小模型的性能越好, 即模型对上下文的约束能力就越强. ELDA 模型与 OLDA 模型 (此时各时间片话题规模固定为 50) 的困惑度如图 2 所示.

可以发现, 当文本集变化时, 多数情况下 ELDA 方法较 OLDA 方法的困惑度更小, 进行话题发现的性能更好.

4.2 网络新闻话题演化分析

4.2.1 子话题数发现

实验进一步采用从网易新闻采集的 2010 年网络新闻作为数据集进行话题演化分析. 该数据集包

含 25 506 篇新闻报道, 共计 5 066 913 个词例, 包含 92 648 个唯一性词汇, 划分为 12 个时间片, 各时间片对应文本集的规模如表 2 所示.

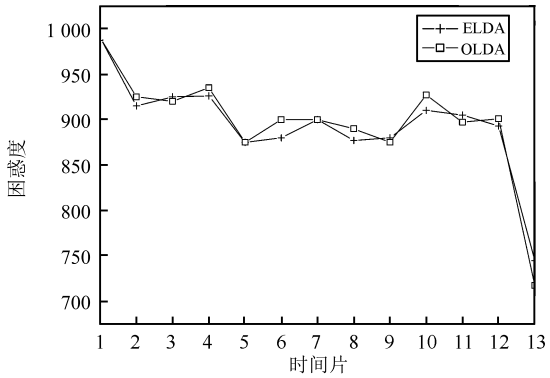


图 2　ELDA 模型与 OLDA 模型困惑度比较

Fig. 2　Comparison of perplexities of ELDA and OLDA trained on NIPS

表 2　数据集各时间切片所含新闻报道

Table 2　Text configurations of the experiments

时间片	规模	时间片	规模	时间片	规模
♯1	2 673	♯5	1 760	♯9	2 585
♯2	2 699	♯6	1 178	♯10	2 442
♯3	2 700	♯7	2 135	♯11	2 060
♯4	1 115	♯8	1 195	♯12	1 998

实验采用 ELDA 方法对各时间片中文本集进行子话题发现, 通过人工判读子话题发现的结果, 检测 ELDA 方法进行子话题发现的精度[25]:

$$Precision = \frac{\text{系统给出的正确答案数}}{\text{系统给出的答案数}} \tag{10}$$

对于本实验, “系统给出的正确答案数” 是指以子话题中出现概率最大的前五个词为判断依据, 若 ELDA 方法抽取的子话题包含的特征词汇与人工标注的子话题特征词汇存在三个 (含) 以上语义重合即认为 ELDA 方法抽取的子话题正确, 例如对于人工标注得到的子话题 {上海, 世博会, 波兰, 音乐, 肖邦} 和 ELDA 方法抽取的子话题 {中国, 上海, 世博会, 波兰, 肖邦}, 认为 ELDA 方法抽取的子话题正确.

根据上述方法, 实验重复 10 次, 以所有时间片中的总结果进行计算, 得到 ELDA 方法的最佳精度以及平均精度 (精确到个位), 并与 OLDA 方法进行了比较, 如表 3 所示. 实验分别测试了子话题数为 300 和 400 两种情况下 OLDA 方法进行子话题发现的结果, 分别对应 OLDA (300) 和 OLDA (400) 方法.

表 3　ELDA 与 OLDA 方法精度对比 (%)

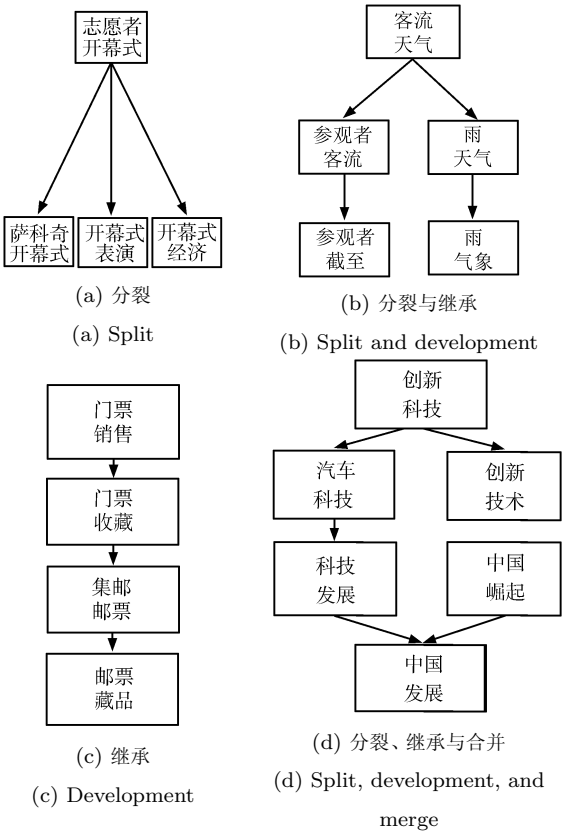
Table 3　Comparison of ELDA and OLDA (%)

	ELDA	OLDA (300)	OLDA (400)
最佳	87	78	85
平均	83	74	73

可以发现, ELDA 方法的平均精度优于 OLDA 模型. 对于最佳精度, OLDA 方法在不同的子话题数设置下具有较大波动, 且低于 ELDA 方法.

4.2.2　话题内容演化

以各时间片中包含特征词汇 “世博”、“世博会” 和 “世博园” 的话题 (以下简称世博会相关话题) 为例说明话题演化的过程, 如图 3 所示. 在本文语境清晰的前提下, 为了充分表达子话题的内容, 图 3 中未专门列出下列特征词汇 “中国”、“世博”、“世博会”、“世博园” 和 “上海”.



(a) 分裂

(a) Split

(b) 分裂与继承

(b) Split and development

(c) 继承

(c) Development

(d) 分裂、继承与合并

(d) Split, development, and merge

图 3　子话题关联分析示例

Fig. 3　Correlation analysis of sub-topicsm

图 3 中选取每个子话题中出现概率最大的一组词描述子话题的内容. 可以发现, 子话题间存在继承、分裂、合并等关系以及由上述基本关系组成的

复杂演化关系. 例如, 时间片 4 中子话题 “志愿者, 开幕式” 分裂为时间片 5 中子话题 “萨科奇, 开幕式”、“开幕式, 表演” 和 “开幕式, 经济” (见图 3 (a); 时间片 6 中子话题 “客流, 天气” 分裂为时间片 7 中的子话题 “参观者, 客流” 和 “雨, 天气”, 上述子话题在时间片 8 中分别发展为 “参观者, 截至” 和 “雨, 气象” (见图 3 (b)); 时间片 8 中子话题 “门票, 销售” 逐渐发展为时间片 9 至时间片 11 中的 “门票, 收藏”、“集邮, 邮票” 和 “邮票, 收藏” (见图 3 (c)); 时间片 9 中子话题 “创新, 科技” 首先分裂为时间片 10 中子话题 “汽车, 科技” 和 “创新, 技术”, 时间片 11 中子话题 “创新, 技术” 消亡, 而子话题 “汽车, 科技” 发展为 “科技, 发展”, 与新产生的子话题 “中国, 崛起” 在时间片 12 中合并为 “中国, 发展” (见图 3 (d)).

4.2.3 话题强度变化

本节根据话题混合的概率分析不同时间片话题强度的变化. 以上海世博会相关话题为例, 其强度变化如图 4 所示.

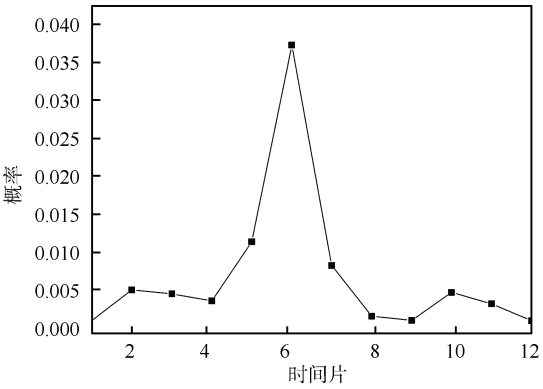


图 4 网络新闻世博会相关话题强度演化曲线

Fig. 4 Probability of Topic Expo 2010 over 12 slices

由图 4 可以看出, 上海世博会话题连续出现, 强度变化具有明显的起伏特点: 话题强度先逐渐上升, 经过一段时间到达峰值, 在较高强度保持一段时间后, 话题强度下降. 这是因为重大事件具有持续较高的关注度, 网络新闻不受时间、空间的限制, 能够针对重大事件进行全时全面的报道, 且报道频率随着事件临近逐渐增大, 特别是当事件发生时, 报道集中聚焦, 表现为话题强度急剧上升到达峰值, 事件发展后期随关注度下降报道频率下降, 使得话题强度大幅回落.

5 结论

本文提出了网络话题演化分析模型, 基于 OLDA 方法利用先验知识进行子话题发现, 通过模型选择方法动态确定子话题数, 抽取网络信息中隐含的话题片断; 提出了基于相对熵的子话题关联分析方法, 采用具有时序关系和内容关联的子话题描述话题, 定义了子话题产生、消亡、继承、分裂和合并五种演化类型, 根据子话题语义相似度和时序关系建立子话题间的关联. 实验基于标准数据集和真实网络新闻分析了话题内容和强度的演化. 结果表明, 本文提出的话题演化方法能有效检测话题内容和强度随时间的变化. 在此基础上, 下一步将研究如何采用更丰富的指标评价话题模型的性能以及基于话题进行网络新闻自动分类.

致谢

本文实验在开源程序 knowceans-ilda (http://arbylon.net/resources.html) 基础上实现, 特此表示感谢.

References

1 Allan J, Carbonell J G, Doddington G, Yamron J, Yang Y M, Umass J A, Cmu B A, Cmu D B, Cmu A B, Cmu R B, Dragon I C, Darpa G D, Cmu A H, Cmu J L, Umass V L, Cmu X L, Dragon S L, Van Mulbregt Dragon P, Umass R P, Cmu T P, Umass J P, Umass M S. Topic detection and tracking pilot study: Final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco, USA: Morgan Kaufmann, 1998. 194−218

2 Shan Bin, Li Fang. A survey of topic evolution based on LDA. *Journal of Chinese Information Processing*, 2010, **2**4(6): 43−49, 68
(单斌, 李芳. 基于 LDA 话题演化研究方法综述. 中文信息学报, 2010, **24**(6): 43−49, 68)

3 NIST. Topic Detection and Tracking Evaluation (TDT 2002) [Online], available: http://www.itl.nist.gov/iad/mig//tests/tdt/, April 28, 2012

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

5 Wang X R, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD). Philadelphia, USA: ACM, 2006. 424−433

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

7 Wang C, Blei D, Heckerman D. Continuous time dynamic topic models. In: Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence. Helsinki, Finland: AUAI, 2008. 579−586

8 Nallapati R M, Cohen W, Ditmore S, Lafferty J, Ung K. Multiscale topic tomography. In: Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD). San Jose, USA: ACM, 2007. 520−529

9 Wei X, Sun J M, Wang X R. Dynamic mixture models for multiple time series. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India: ACM, 2007. 2909−2914

10 Song X D, Lin C Y, Tseng B L, Sun M T. Modeling and predicting personal information dissemination behavior. In: Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD). Chicago, USA: ACM, 2005. 479−488

11 AlSumait L, Barbar á D, Domeniconi C. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 8th IEEE International Conference on Data Mining. Washington, USA: IEEE, 2008. 3−12

12 Shi Jing, Dai Guo-Zhong. Text segmentation based on PLSA model. *Journal of Computer Research and Development*, 2007, **44**(2): 242−248
(石晶, 戴国忠. 基于 PLSA 模型的文本分割. 计算机研究与发展, 2007, **44**(2): 242−248)

13 Shi Jing, Hu Ming, Shi Xin, Dai Guo-Zhong. Text segmentation based on model LDA. *Chinese Journal of Computers*, 2008, **31**(10): 1865−1873
(石晶, 胡明, 石鑫, 戴国忠. 基于 LDA 模型的文本分割. 计算机学报, 2008, **31**(10): 1865−1873)

14 Shi Jing, Fan Meng, Li Wan-Long. Topic analysis based on LDA model. *Acta Automatica Sinica*, 2009, **35**(12): 1586−1592
(石晶, 范猛, 李万龙. 基于 LDA 模型的主题分析. 自动化学报, 2009, **35**(12): 1586−1592)

15 Chu Ke-Ming, Li Fang. Topic evolution based on LDA and topic association. *Journal of Shanghai Jiaotong University*, 2010, **44**(11): 1496−1500
(楚克明, 李芳. 基于 LDA 话题关联的话题演化. 上海交通大学学报, 2010, **44**(11): 1496−1500)

16 Chu Ke-Ming. The Research on Topic Evolution for News Based on LDA Model [Master dissertation], Shanghai Jiao Tong University, China, 2010
(楚克明. 基于 LDA 的新闻话题演化研究 [硕士学位论文]. 上海交通大学, 中国, 2010)

17 Cui Kai, Zhou Bin, Jia Yan, Liang Zheng. LDA-based model for online topic evolution mining. *Computer Science*, 2010, **37**(11): 156−193
(崔凯, 周斌, 贾焰, 梁政. 一种基于 LDA 的在线主题演化挖掘模型. 计算机科学, 2010, **37**(11): 156−193)

18 Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD). New York, USA: ACM, 2011. 448−456

19 Hu Yan-Li, Bai Liang, Zhang Wei-Ming. OLDA-based method for online topic evolution in network public opinion analysis. *Journal of National University of Defense Technology*, 2012, **34**(1): 150−154
(胡艳丽, 白亮, 张维明. 网络舆情中一种基于 OLDA 的在线话题演化方法. 国防科技大学学报, 2012, **34**(1): 150−154)

20 Griffiths T L, Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(Suppl 1): 5228−5235

21 Heinrich G. Parameter estimation for text analysis [Online], available: http://www.arbylon.net/publications/text-est.pdf, September 26, 2012

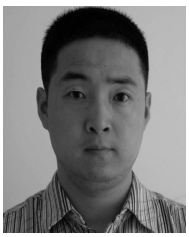
22 Manning C D, Schuetze H. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT, 1999

23 Lv Nan. Research on Topic Tracking and Evolution Analysis Technique [Master dissertation], PLA Information Engineering University, China, 2009
(吕楠. 话题追踪与演化分析技术研究 [硕士学位论文]. 信息工程大学, 中国, 2009)

24 NIPS Online Repository [Online], available: http://nips.djvuzone.org, September 26, 2012

25 Jurafsky D, Martin J H. *Speech and Language Processing*. New Jersey: Pearson, 2005

胡艳丽 国防科学技术大学信息系统与管理学院讲师. 2012 年获国防科学技术大学信息系统与管理学院博士学位. 主要研究方向为信息资源管理和网络舆情分析. 本文通信作者.
E-mail: smilelife1979@hotmail.com
(**HU Yan-Li** Lecturer at the School of Information System and Management, National University of Defense Technology. She received her Ph. D. degree from National University of Defense Technology in 2012. Her research interest covers information resource management and analysis of network public opinions. Corresponding author of this paper.)



白 亮 国防科学技术大学信息系统与管理学院讲师. 2008 年获国防科学技术大学信息系统与管理学院博士学位. 主要研究方向为智能信息处理和社会媒体分析. E-mail: liangbai@nudt.edu.cn
(**BAI Liang** Lecturer at the School of Information System and Management, National University of Defense Technology. He received his Ph. D. degree from National University of Defense Technology in 2008. His research interest covers intelligent information processing and social media analysis.)

张维明 国防科学技术大学信息系统与管理学院教授. 主要研究方向为信息系统和体系工程.
E-mail: wmzhang@nudt.edu.cn
(**ZHANG Wei-Ming** Professor at the School of Information System and Management, National University of Defense Technology. His research interest covers information system and system of engineering.)