网络舆情中一种基于OLDA的在线话题演化方法*

胡艳丽,白 亮,张维明
(国防科技大学 信息系统工程重点实验室,湖南 长沙 410073)

摘 要:研究网络舆情分析中话题演化方法。首先分析网络舆情信息的特点;在此基础上,建立网络舆情信息模型,基于话题模型抽象描述文本内容的隐含语义,进而建立文本流在时间序列上的关联模型;进一步,提出基于OLDA的话题演化方法,针对舆情信息的特点,建立不同时间片话题间的关联。实验结果表明,该方法能够有效检测话题演化,为网络舆情分析提供了有效途径。

关键词:网络舆情;话题模型;话题演化; Gibbs抽样
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2012)01-0150-05

OLDA-based method for online topic evolution in network public opinion analysis

HU Yanli, *BAI Liang*, *ZHANG Weiming*
(Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China)

Abstract: The topic evolution was investigated for network public opinion analysis. The properties of network public opinion information were analyzed firstly. Based on the properties, the latent semantics of textual data for network public opinion was described by using the topic model, and the text streams are modeled with a consideration of time for online analysis. Furthermore, a topic evolution method based on OLDA was proposed by incorporating the correlation of topics among time slices. The proposed method was experimentally verified to be efficient for detecting topic evolution of network public opinion.

Key words: network public opinion; topic model; topic evolution; Gibbs Sampling

互联网的普及和信息技术的发展使得网络成为人们获取信息和传播观点的主要渠道之一。网络舆情是社会舆论的一种表现形式,是通过互联网传播的公众对社会突发事件特别是热点、焦点问题所持的有较强影响力、带有倾向性的言论和观点[1]。

话题(Topic)是事件相关的报道的集合[2]。围绕某一话题的报道、言论和观点在网络上迅速扩散,能够在短时间、大范围形成具有强大影响力的网络舆情。话题演化表示话题随时间推移表现出的动态性、发展性和差异性。话题演化是网络舆情分析的重要内容之一,具有重要的理论价值和社会意义。

1 相关工作

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,研究人员对统计话题模型进行了深入研究。LDA(Latent Dirichlet Allocation)模型[4]是三层变参数层次贝叶斯模型,是一种具有文本话题表示能力的非监督学习模型,在文本表示和文本挖掘中得到广泛应用。进一步,研究人员通过引入时间信息对LDA模型进行扩展。根据话题模型是否具有在线处理能力,LDA模型面向时间的扩展模型可分为两类[5]。Topic Over Model(TOM)模型[6]将时间看作连续的可观测变量,通过考虑文本的时间属性计算话题在时间上的分布强度,话题由词和文档的时间属性共同决定。动态话题模型(Dynamic Topic Model, DTM)[7]根据时间分割文本集合,应用LDA模型发现每个时间窗口的话题,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

* **收稿日期**:2011-08-30
基金项目:国家自然科学基金资助项目(60902094)。
作者简介:胡艳丽(1979-),女,河南夏邑人,博士研究生,E-mail:smilelife1979@hotmail.com;
张维明(通信作者),男,教授,博士,博士生导师,E-mail:wmzhang@nudt.edu.cn

(Continuous Time Dynamic Topic Model, CTDTM)[8]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[9]研究多时间粒度的话题演化,允许用户根据需要增大或减小话题演化的时间粒度。但上述扩展模型需要对整个文本集进行建模,无法在线处理新到达的文本。

为应对非常规突发事件,舆情分析需要进行及时甚至近实时的话题演化分析。在线话题模型在这方面进行了有益的探索。动态混合(Dynamic Mixture Model, DMM)模型[10]严格按照时间先后顺序处理到达的文本,不依赖于文本可交换假设。增量 LDA(Incremental Latent Dirichlet Allocation, ILDA) 模型[11]根据文本到达时间对 LDA 模型进行增量建模,研究不同时间段话题数可变的话题内容演化问题。OLDA 模型[12]根据时间信息将文本集划分为一组时间窗口,应用 LDA 模型发现每个时间窗口文本集的话题,并且采用话题历史分布作为当前时间窗口话题发现的先验知识,研究话题内容和强度的演化。

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[13-15],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[16-17]。崔凯等提出基于 LDA 的在线主题演化模型[18],只考虑了不同时间片间话题所含关键词的联系。

统计话题模型模拟文本的生成过程,对文本预测具有很好的应用效果。舆情分析中,基于话题模型的话题演化目前还鲜有研究。本文研究网络舆情信息的话题演化问题,提出基于 OLDA 的话题演化方法,为网络舆情分析提供基础。

2　网络舆情信息特性

网络舆情信息的表现形式主要为网络新闻、新闻评论、BBS 论坛帖子、博文及电子邮件等,经过预处理后主要以文本的形式存储,通常具有下述特点:

(1)舆情信息具有海量高维特性:通过网络表达和传播信息的便捷性使得舆情信息具有海量特性,特别是非常规突发事件通过网络可以在短时间内迅速形成大规模的舆情信息。另一方面,文本形式的舆情信息其特征具有高维特性,为舆情分析带来很大难度;

(2)舆情信息在时间上具有继承性和延续性:随着舆情经历酝酿、发展、形成等一系列过程,相应的舆情信息在对应的时间段内根据舆情变化具有继承性和延续性,内容随时间不断更新;

(3)舆情信息在内容上具有交互性:随时间更新的过程中舆情信息并非完全独立,而是彼此关联的,并且内容上往往存在交互关系。例如 BBS 论坛中发帖人围绕主帖的讨论,其回帖通常是对主帖内容的回复;

(4)舆情信息涉及的话题具有演化性:舆情信息通常会同时涉及经济、法律、道德等诸多方面,并且随时间发展,话题的内容和关注度都会发生改变。

上述特点决定了舆情分析技术面临巨大挑战:一方面,舆情分析要能够从海量文本中准确发现舆情信息涉及的话题,另一方面,针对舆情信息随时间不断更新的特点,舆情分析还需要高效的在线处理能力,能够及时甚至近实时发现话题随时间的变化,从而对舆情进行有效监测和引导。

3　网络舆情在线话题演化分析

3.1　网络舆情信息建模

网络舆情演化可分为形成、高涨、波动、淡化等阶段[19]。相应地,网络舆情信息在时间上具有继承性和延续性。舆情监测过程中采集的文本流按时间顺序到达,每个文本通常会涉及一组话题,一个话题可由一组关键词表示。网络舆情信息的继承性和延续性体现为相邻时间片的文本内容间存在关联,其对应的话题在时间上具有延续性,为舆情监测和预警提供了可能。

随时间发展变化的舆情信息可用时间序列上相互关联的文本集表示,每个文本视为一组话题的混合分布,话题是一组关键词的分布。根据舆情分析的需要将文本流按一定的时间粒度进行划分。时间片 t 内到达的文本集记作 $D^t = \{d_1, \cdots, d_n\}$,其中 $d_i (i = 1, \cdots, n)$ 为舆情信息对应的一个文本。令词汇集为 V,对于文本 d_i,令 $P(z)$ 表示话题出现的概率,$P(w|z)$ 表示话题包含词 $w \in V$ 的概率。根据贝叶斯规则,词对于话题的后验概率如(1)式所示。

$$P(z|w) = \frac{P(w|z) * P(z)}{P(w)} \tag{1}$$

其中,z 是由一组话题构成的话题向量,话题 k 出现对应于向量 z 的第 k 维。

历史时间片中文本涉及的话题(以下简称话题分布)以及话题所含关键词的分布(以下简称词分布)为当前时间片的话题演化分析提供了先

验知识。例如,假设相邻时间片间文本内容存在关联,前一时间片中话题分布以及词分布的后验为当前时间片的话题分布和词分布提供了先验知识,在时间序列上的关系如图 1 所示。上述模型是网络舆情信息话题演化分析的基础。

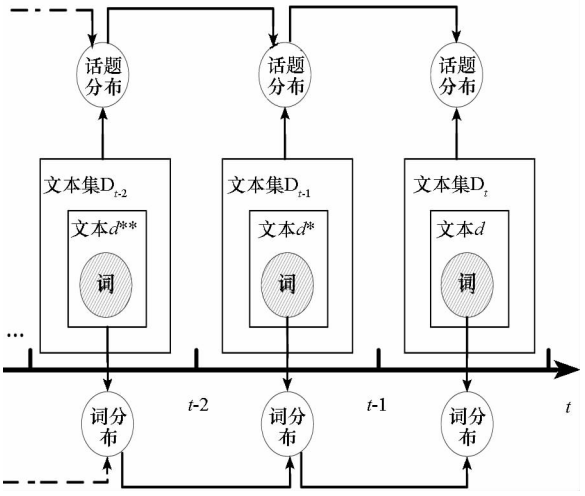

图 1 网络舆情信息关联示意图

Fig. 1 Correlated text stream of network public opinion

3.2 基于 OLDA 的话题演化方法

根据舆情分析的粒度,按时间信息将文本集离散到 N 个时间片中,设每个时间片中文本涉及的话题数为 K。

对于时间片 t,令文本 d 上的话题混合服从参数为 $\theta^{(d)}$ 的多项分布,记作 $z|\theta^{(d)}:Multi(\theta^{(d)})$,其中 $\theta^{(d)}=(\theta_1^{(d)},\cdots,\theta_K^{(d)})$;话题在词汇集上的混合服从参数 φ_k 的多项分布,记作 $w|z=k,\varphi_k:Multi(\varphi_k)$。为便于参数推理,采用 LDA 模型假设,令话题分布和词分布的先验服从 *Dirichlet* 分布,分别记作 $\theta:Dirichlet(\alpha)$ 和 $\varphi:Dirichlet(\beta)$,其中 $\theta=\{\theta^{(d)}|d\in D\}$,$\varphi=\{\varphi_k|k\in[1,K]\}$。

以时间片 $t-1$ 中话题分布和词分布的后验加权作为时间片 t 中话题分布和词分布的先验,即时间片 t 中话题分布和词分布的 *Dirichlet* 先验满足:

$$\alpha^t=\Phi*\omega \tag{2}$$

$$\beta^t=\Psi*\omega' \tag{3}$$

其中,Φ 是 $K\times|D^t|$ 维矩阵,每一列对应时间片 $t-1$ 上的一个话题分布 $\theta^{(d)}$,Ψ 是 $|V|\times K$ 维矩阵,每一列对应时间片 $t-1$ 上的一个词分布 φ_k。

参数 $\theta^{(d)}$ 和 φ_k 采用 Gibbs 抽样方法,通过对话题的词分配抽样进行估计。时间片 t 上,参数 $\hat{\theta}^{(d)}$ 对应话题 k 以及参数 φ_k^t 对应词 w 的估计公式如下:

$$(\varphi_k^{(w)})^t=\frac{(n_k^{(w)})^t+(n_k^{(w)})^{t-1}+\beta_k^t}{(n_k^{(\cdot)})^t+(n_k^{(\cdot)})^{t-1}+M\beta_k^t} \tag{4}$$

$$(\theta_k^{(d)})^t=\frac{(n_k^{(d)})^t+\alpha_k^t}{(n_\cdot^{(d)})^t+K\alpha_k^t} \tag{5}$$

其中,M 是词汇集 V 包含的词汇数,K 是话题数,$(n_k^{(w)})^{t-1}$ 和 $(n_k^{(w)})^t$ 分别表示时间片 $t-1$ 和时间片 t 上词 w 被分配给话题 k 的频数;类似,$(n_k^{(\cdot)})^{t-1}$ 和 $(n_k^{(\cdot)})^t$ 分别表示时间片 $t-1$ 和时间片 t 上分配给话题 k 的所有词数,$(n_k^{(d)})^t$ 表示时间片 t 上文本 d 中分配给话题 k 的词数,$(n_\cdot^{(d)})^t$ 表示时间片 t 上文本 d 所有被分配了话题的词数。

在 OLDA 模型考虑词分布相互关联的基础上,本文进一步考虑了话题分布在时间片间的联系,实现在线话题演化分析,满足舆情信息处理的需要。

4 实 验

天涯社区[20]是国内最主要的 BBS 论坛之一,本文采用天涯社区经济论坛部分数据进行话题演化分析。实验数据集包含 2011 年 1 月至 4 月的 2381 个帖子,大小为 122M。根据时间信息将数据集划分为 4 个时间片,各时间片所含帖子的规模如表 1 所示。实验设置话题数 $K=20$,权重矩阵 ω 和 ω' 中的元素取值为 0.5。

表 1 数据集信息

Tab. 1 Setting of data set

时间片	#1	#2	#3	#4
规模	165	199	775	118

通过预处理去除帖子内容中的停用词和主要的帖子标签,如:“天涯社区”、“经济论坛”、“作者”、“回复时间”及具体日期等。在此基础上,抽取得到的话题涉及股市、房地产、投资、银行、人民币升值、收入分配、增值税、国际经济形势等。

以日本的相关话题为例,选取每个时间片该话题中出现概率最大的一组词,如表 2 所示。可以看出时间片 1 和时间片 2 该话题主要与日本制造业、汽车行业相关,时间片 3 和时间片 4 对应的话题中地震、核电站、辐射等关键词占主导地位,反映了 3 月 11 日日本发生强烈地震,引发核泄漏危机这一重大突发事件。

表 2　日本相关话题所含关键词

Tab. 2　Top words of Japan

#1	#2	#3	#4
日本	日本	日本	辐射
企业	民族	地震	通货膨胀
中小企业	地震	日元	日本
公司	企业	核电站	房产
制造业	汽车	发生	影响
技术	丰田	影响	资金
通胀	公司	民族	手段
房价	制造业	历史	方式

实验进一步对本文模型和 OLDA 模型进行了对比分析。以中国经济相关话题为例,基于本文模型和 OLDA 模型进行话题演化分析得到的部分话题内容如表 3 所示。

表 3　热点话题所含关键词

Tab. 3　Top words of hot topics

OLDA	#1	资产	美元	市场	金融	中国	人民币
	#2	美元	美国	资产	市场	中国	金融
	#3	美元	升值	市场	美国	中国	黄金
	#4	美国	升值	美元	人民币	中国	黄金
本文模型	#1	中国	金融	市场	资产	人民币	银行
	#2	中国	市场	金融	人民币	资产	美国
	#3	中国	市场	金融	人民币	银行	美国
	#4	中国	人民币	市场	银行	经济	升值

可以发现,OLDA 模型仅考虑话题所含关键词间的关联,但不同时间片上的话题语义关联不强。本文提出的模型进一步考虑了不同时间片间的话题关联,有效增强话题随时间演化的语义联系。根据 Kullback-Leibler(KL)散度得到的话题距离验证了本文模型的有效性(见图 2)。

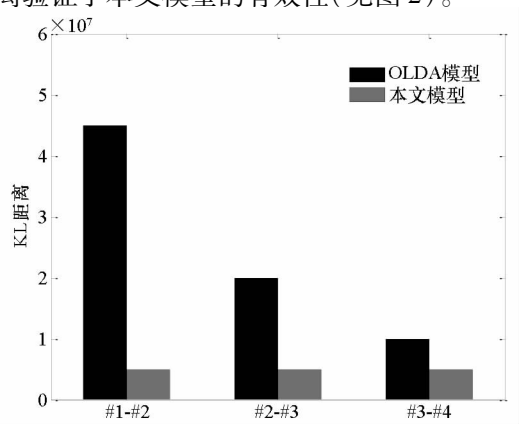

图 2　不同时间片话题 KL 距离对比

Fig. 2　KL divergence of topics between time slices

5　结束语

本文分析了网络舆情信息的特点,建立了网络舆情信息话题演化模型,在此基础上,提出了基于 OLDA 的话题演化方法,针对互联网数据集进行了话题演化分析。实验结果表明,该方法能有效检测话题内容随时间的变化。下一步将研究具有短文本形式的舆情信息的话题发现和演化问题。

参考文献(References)

[1] 徐晓日. 网络舆情事件的应急处理研究 [J]. 华北电力大学学报:社会科学版, 2007(1): 89 -93.
XU Xiaori. Study on the way to solve the parpxysmal public feelings on Internet [J]. Journal of North China Electric Power University (Social Sciences), 2007 (1): 89 - 93. (in Chinese)

[2] Allan J, Carbonell J, Doddington G, et a1. Topic detection and tracking pilot study: final report [C]//DARPA Broadcast News Transcription and Understanding Workshop, San Francisco, CA, Morgan Kaufmann Publishers Inc, 1998:194 -218.

[3] The 2002 Topic Detection and Tracking (TDT2002) Task definition and evaluation plan [EB/OL]. [2011 -03 -18]. ftp://jaguar. ncsl. nist. gov/tdt/tdt2002/evalplans/TDT02. Eval. plan. v1. 1. pdf.

[4] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003 (3):993 -1022.

[5] 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.
SHAN Bin, LI Fang. A survey of topic evolution based on LDA[J]. Journal of Chinese Information Processing, 2010, 24 (6): 43 -49. (in Chinese)

[6] Wang X R, McCallum A. Topic over time: a non-Markov continuous -time model of topical trends [C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 2006: 424 -433.

[7] Blei D M, Lafferty J D. Dynamic topic model [C]// Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, 2006: 113 -120.

[8] Wang C, Blei D M, Heckerman D. Continuous time dynamic topic models [C]// Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence, Corvallis, Oregon, 2008:579 -586.

[9] Nallapati R M, Cohen W, Ditmore S, et. al. Multi-scale topic tomography [C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, 2007: 520 -529.

[10] Wei X, Sun J, Wang X. Dynamic mixture models for multiple time series [C]// Proceedings of the 20th International Joint

Conference on Artifical Intelligence, Hyderabad, India, 2007: 2909 – 2914.

[11] Song X, Lin C Y, Tseng B L, et. al. Modeling and predicting personal information dissemination behavior [C]// Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 2005: 479 – 488.

[12] Alsumait L, Barbara D, Domeniconi C. On-line LDA: adaptive topic models of mining text streams with applications to topic detection and tracking [C]// Proceedings of the 8th IEEE International Conference on Data Mining, Washington, DC, USA: IEEE Computer Society, 2008: 3 – 12.

[13] 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.
SHI Jing, DAI Guozhong. Text segmentation based on PLSA model[J]. Journal of Computer Research and Development, 2007, 44(2): 242 – 248. (in Chinese)

[14] 石晶, 胡明, 石鑫,等. 基于 LDA 模型的文本分割 [J]. 计算机学报, 2008, 31(10): 1865 – 1873.
SHI Jing, HU Ming, SHI Xin, et al. Text segmentation based on model LDA [J]. Chinese Journal of Computers, 2008, 31(10): 1865 – 1873. (in Chinese)

[15] 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.
SHI Jing, FAN Meng, LI Wanlong. Topic analysis based on LDA model[J]. Acta Automatica Sinica, 2009, 35(12): 1586 – 1592. (in Chinese)

[16] 楚克明, 李芳. 基于 LDA 话题关联的话题演化 [J]. 上海交通大学学报, 2010, 44(11): 1496 – 1500.
CHU Keming, LI Fang. Topic evolution based on LDA and topic association[J]. Journal of Shanghai Jiaotong University, 2010, 44(11): 1496 – 1500. (in Chinese)

[17] 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.
CHU Keming. The research on topic evolution for news based on LDA model [D]. Shanghai:Shanghai Jiaotong University, 2010. (in Chinese)

[18] 崔凯, 周斌, 贾焰,等. 一种基于 LDA 的在线主题演化挖掘模型 [J]. 计算机科学, 2010, 37(11): 156 – 193.
CUI Kai, ZHOU Bin, JIA Yan, et al. LDA-based model for online topic evolution mining[J]. Computer Science, 2010, 37(11): 156 – 193. (in Chinese)

[19] 宋海龙, 巨乃岐, 张备,等. 突发事件网络舆情的形成、演化与控制[J]. 河南工程学院学报:社会科学版, 2010, 25(4):12 – 16.
SONG Hailong, JU Naiqi, ZHANG Bei, et al. Formation, evolution and control of network public opinion for emergencies [J]. Journal of Henan Institute of Engineering (Social Science Edition), 2010, 25(4): 12 – 16. (in Chinese)

[20] 天涯论坛 [EB]. [2011 – 04 – 20]. http://www.tianya.cn/bbs/index.shtml.
Tianya BBS [EB]. [2011 – 04 – 20]. http://www.tianya.cn/bbs/index.shtml. (in Chinese)