

基于动态主题模型的话题热度及演化趋势探索^①

曹丽娜,唐锡晋

(中国科学院 数学与系统科学研究院,100190)

摘 要:在当今中国,如何在互联网上追踪人们的中心话题及话题的热度变化趋势无论是在理论研究还是在实际的社会管理中都是非常重要的议题。而对社会管理而言,能够及时、准确地对社会危机事件的言论进行适当干预的前提条件是对公众意见的整体感知和准确把握。该文通过运用动态主题模型对天涯论坛下“天涯杂谈”版块的数据进行分析,探索了在该版块下网民每日首发帖的主要话题及这些话题的热度变化。

关键词:主题探测;动态主题模型;天涯论坛

Prevailing Trends Detection of Public Opinions Based on Dynamic Topic Model

Cao Lina, Tang Xijin

(Academy of Mathematics and Systems Science Chinese Academy Sciences, China, 100190)

Abstract: How to catch both the central topics and the trend of topics over the on-line discussions are not only of theoretical importance for scientific research, but also of practical importance for social management in current China. In social management perspective, making intervention toward crisis timely and precisely depends on the right image or perception of public opinions toward the crisis. In our research, topic modeling is applied to explore the changing topics of new posts collected from Tianya Zatan Board of Tianya Club. Those online data reflect the community opinions toward social problems.

Keywords: topics detection; dynamic topic models; Tianya Club

^① 资助项目:国家重点基础研究发展计划(2010CB731405),国家自然科学基金(71171187)。

1 引言

近年来,由于网民数量的增加以及 Web 2.0 时代的到来,社交网络成为人们沟通交流、分享信息、表达情感的重要场所。许多社会事件的发生和发展往往都是通过社交网络媒体直接曝光,网民在第一时间便可获得事件的信息从而参与讨论,进而在网络上形成强大的舆论动向。这些网络舆论表现出随时间不断演化发展的特点,比如“郭美美”事件,最早是 2011 年 6 月因新浪微博用户郭美美实名认证为“中国红十字会商业总经理”并在微博中炫富而引发的社会争议,后来随着网民的不断参与,矛头很快指向了中国红十字会,最终引起了一场慈善信任危机。同样类似的例子还有“小悦悦”事件等。由此可见,网络舆论会对现实中人们的生活产生深远的影响,同时也对政府的社会管理工作提出了更高的要求。如何准确地把握网络舆论,并在适当的时间采取监管措施,是政府进行社会管理时面临的实际问题。

在众多网络媒体中,网络论坛一直以来都是一个很重要的交流平台。它比博客具备更强的可交互性,比微博更利于网民分析事件的发展演变过程。即使许多事件是通过微博曝光的,但论坛依然为事件的发展提供了更多的讨论空间。在前期对天涯杂谈版块(天涯论坛的子版块)数据进行每日采集^[1]的基础上,本文进一步聚焦每日话题变迁的研究。这些帖子涵盖的话题不仅包括日常生活,更多的是对社会突发事件、社会不公、社会道德、反腐败等关乎民生的讨论,是研究社会舆情较好的数据源。

本文所针对的实际数据即首发帖,是一个文本时间序列集。相较于文本之间没有任何相互关联的文档集而言,文本时间序列中每篇文档所附带的时间信息对我们理解文本内容有非常大的帮助。从这些时间信息中,可以进行许多挖掘,例如对曝光的事件进行跟踪、分析某些话题关注度随时间的演化过程、话题的演化分支等。

本文尝试运用动态主题模型(dynamic topic models,DTM)研究各种话题的热度及其变动趋势,描述每个话题中词语的细小的变化趋势。文章结构如下:首先介绍话题挖掘的各种方法,包括定性与定量的分析方法,其次描述数据的清理过程,及对模型结果的分析处理过程,最后给出结论。

2 话题挖掘的方法

社会系统是一个开放复杂巨系统^[2]。现实中,社会管理、网络问政等对于管理者来说都是 wicked 问题,这类问题的求解过程是一个问题结构化的过程。

2.1 定性综合集成方法

以往我们的工作中,两种定性综合集成的支持技术——CorMap 和 iView,被视为处理非结构问题的方法^[3-4]。它们基于不同的原理,针对文本信息计算处理,运用图论、社会网络、复杂网络等定量方法分析群体意见,从意见中抽取出结构信息或者系统视角,从而引导展开深入探索进而形成关于问题的全景视角,给出定性分析的结果。如果按时间对群体意见的进行分析,则可以探测中心思想的演化变动。

2.2 主题建模

一些更深层次对大规模文档集合进行分析与挖掘的技术也已被提出。这些技术可以自动分析文档集合中所讨论的热点话题、形成对整个文档集合的概括性描述。主题模型(topic model)就是其中一种文本挖掘方法。

主题模型还可以用来从时间序列文本中辨识和追踪动态的主题。根据时间是连续还是离散,基于主题模型的方法主要有三种:(1)最基本的 LDA 模型(latent dirichlet allocation)^[5],该模型采取“词袋法”,通过训练整个文档集,得到若干主题,然后根据文档的发生时间来划分文档集;(2)在生成过程之前用离散的时间标记文档,如 DTM 模型^[6]和 Online LDA (OLDL)模型^[7];(3)随时间变化的主题模型(topic over time, TOT)采用 Beta 分布对主题强度在给定时间范围内的变化进行建模,将文本、词项和时间三者联合起来作为观测数据,以及连续时间动态主题模型(continuous time dynamic topic model, cDTM)^[8]用布朗运动建模主题随时间的演化。

这些方法已经被成功用来探测各种数据的潜在主题结构,如科学文献^[7,9-10]、电子邮件^[8]、新闻^[9]、电影评论^[10]等。在本研究中,我们尝试用 DTM 来分析中国的论坛数据。

DTM 模型先将时间离散化切片,然后假定相邻时间片上整个文档的主题分布以及主题内容都是随时间演化的。在 t 时刻,文档集的主题分布 α_t 以及主题下词语分布为 $\beta_{t,k}$ 均依赖于上一时刻的 α_{t-1} 和 $\beta_{t-1,k}$ 。在 t 时刻,文档的生成过程如下^[6]:

(1)抽取词语的分布 $\beta_t | \beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I)$ 。

(2)抽取主题分布 $\alpha_t | \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$ 。

(3)为每篇文档:

①抽取主题分布 $\eta \sim N(\alpha_t, a^2 I)$ 。

②为每个词:

A. 抽取主题 $Z \sim \text{Mult}[\pi(\eta)]$;

B. 抽取词语 $W_{t,d,n} \sim \text{Mult}[\pi(\beta_{t,z})]$ 。

其中,函数 π 是多项式到正态分布的映射。即

$$f = \pi(\beta_{t,z})_w = \frac{\exp(\beta_{t,z,w})}{\sum_w \exp(\beta_{t,z,w})}, \theta = \pi(\eta) = \frac{\exp(\eta)}{\sum_z \exp(\eta_z)}$$

在 DTM 中,如何根据观测值 $W_{t,d,n}$ 学习参数构成了推断问题。这里,参数 θ 和 θ 是待估参数。文献^[6]提出了采用变分法。

3 数据源及数据清理

天涯论坛(天涯社区)是中国的一个网络社区,截至 2013 年 7 月,注册用户达 8 200 万,在线用户常在 80 万至 100 万左右。

3.1 数据源

“天涯杂谈”版块对社会性话题的关注在天涯所有版块里首屈一指。我们从 2010 年 10 月开始从包括“天涯杂谈”版块在内的若干个版块上获取每日首发帖和更新帖^[1]。在“天涯杂谈”版块下,每日首发帖量大约在 2 000 条左右,更新帖(即被回复的帖子)则达到了 4 000 条左右,截至 2013 年 7 月,总的首发帖数据量已经抓取达到了 300 多万条。

3.1 数据清理

本文抽取了2011年12月—2012年3月这四个月的数据进行分析。这四个月的数据分别作为数据源进行如下清理:(1)去除掉只有链接但已被删除的帖子;(2)用ICTCLAS^①对首发帖进行切词处理,同时只保留名词、动名词以及百度热搜词^[11],生成语料库;(3)去掉词频小于50次并且在所有文档中出现次数小于10次的词语,生成词典。表1列出了清理后帖子的个数、语料库及词典中的词数。

表1 数据集的统计

时间跨度 数据统计	2011年12月	2012年1月	2012年2月	2012年3月
帖子数量	12 155	12 032	20 124	37 549
语料库大小	48 318	44 912	58 911	82 915
词典大小	4 541	3 973	6 091	9 516

4 天涯论坛上话题的探测分析

在DTM模型训练中,时间粒度和时间跨度的选择对模型的结果至关重要。DTM一个重要的假设就是每个时间点下的主题分布都依赖于上一时刻的主题分布,具备一定的连续性,且是平滑演化的,而论坛上讨论的话题日新月异,更新较快。因此我们尝试了一系列的模型实验:在时间跨度上分别选择1个月、3个月、4个月去训练模型,同时用1天、3天和7天为文档标记时间点。最后我们选择按时间跨度为月分别训练四个模型,每个模型的时间间隔为1天。每个模型的主题个数设为60个。

4.1 词语变化趋势分析

在所有帖子中,人们在不同时刻会谈论不同的话题,即便是针对同一个话题,所用词语也会有相应的变化。根据后验推断,在文档集这个层级,每个主题都由一系列词语的分布构成。所以,词语分布的变化反映了每个话题中心的变迁。我们从12月份的模型中的60个主题中选择了“环境”这个主题做详细的分析。图1从两个方面来展示在这个主题下词语随时间的变化:图中(a)显示的是在整个12月份,概率值较高的前几个词语(如“环境”、“污染”等词)的分布变化趋势,y轴表示概率分布值;图(b)抽取了这个月其中的其中几个时间点,展示了在每个时间点下概率值排名前5的词语;图(c)显示了10条包含这个主题的帖子的标题。

另外,我们统计了图1(a)中所示5个词语在每个时间点下的每个词词频与总词频比率,如图2所示。

通过图1(a)与图2的比较,可以看出两种方法展现出明显不同的趋势,词频变化表现出来的趋势波动比较大,而DTM表现出更平滑的趋势,可以更好地表现出该主题下词语的变化趋势。进一步的,我们从文档集层级分析了每个时刻下主题的变化。

① 中文切词技术,网址 <http://www.ictclas.org/>。

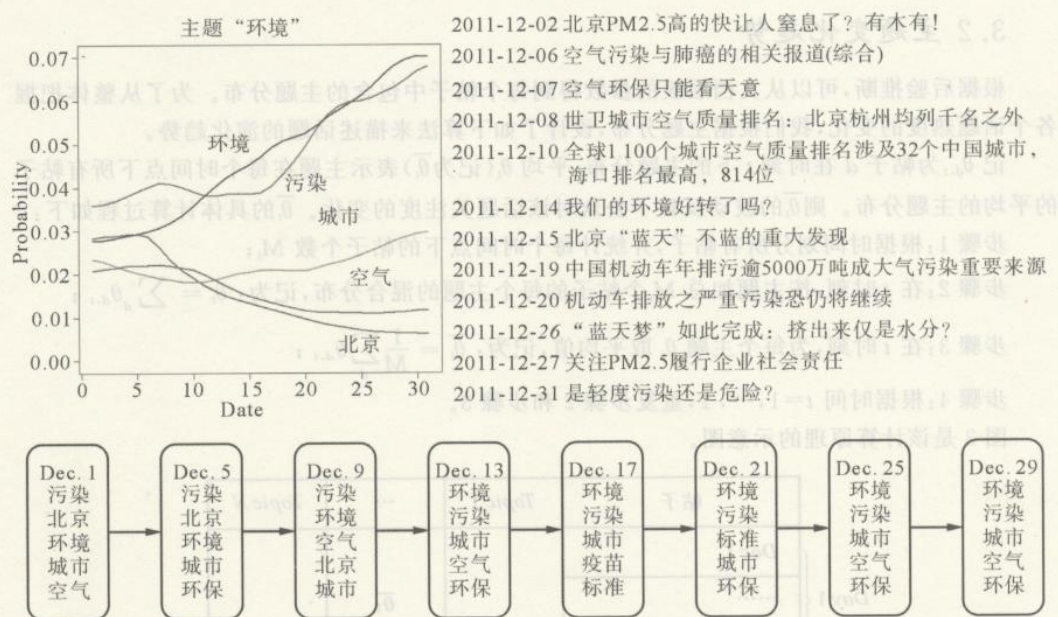


图1 主题“环境”下的词语随时间变化的趋势

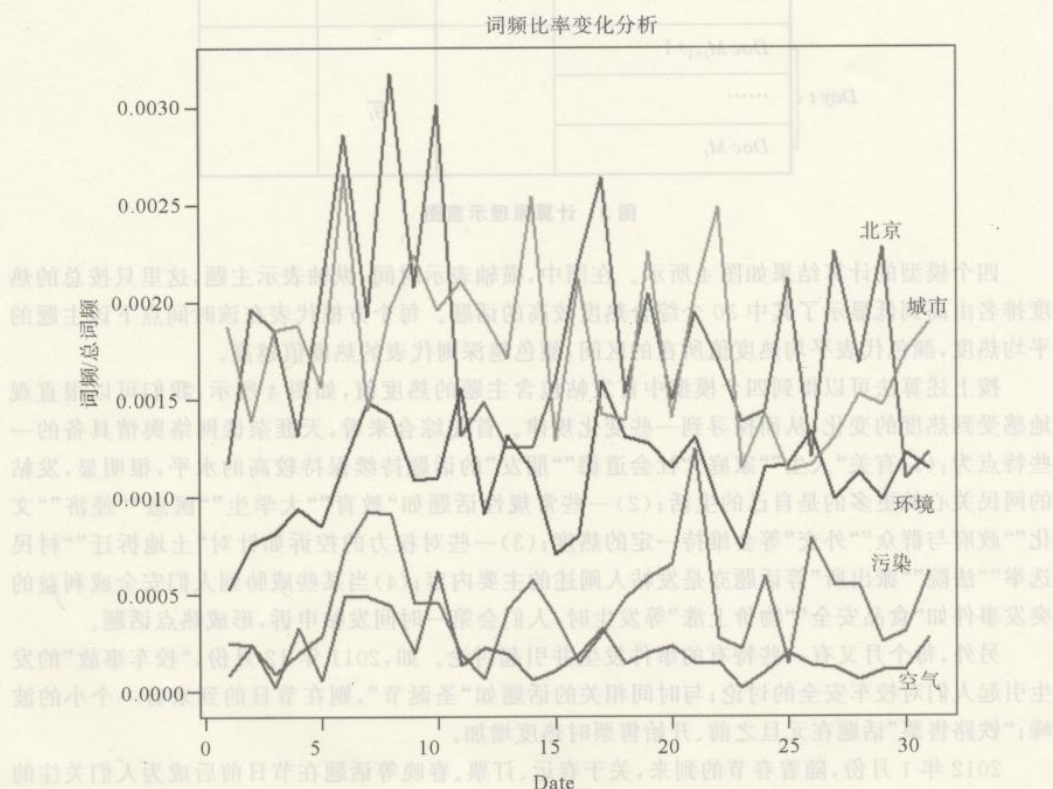


图2 词频/总词频的比率变化

3.2 主题变化趋势

根据后验推断,可以从文档层级的参数得到每个帖子中包含的主题分布。为了从整体把握各个话题热度的变化,我们根据主题分布,设计了如下算法来描述话题的演化趋势。

记 $\theta_{d,t}$ 为帖子 d 在时刻 t 下的主题分布,平均 θ_t (记为 $\bar{\theta}_t$) 表示主题在每个时间点下所有帖子的平均的主题分布。则 $\bar{\theta}_t$ 的波动反映了公众对该话题关注度的变化。 $\bar{\theta}_t$ 的具体计算过程如下:

步骤 1: 根据时间划分所有帖子,并统计每个时间点下的帖子个数 M_t ;

步骤 2: 在 t 时刻,按主题加总 M_t 个帖子的每个主题的混合分布,记为: $\theta_t = \sum_d \theta_{d,t}$;

步骤 3: 在 t 时刻,为每个主题 θ_t 取平均值,记为: $\bar{\theta}_t = \frac{1}{M} \sum_d \theta_{d,t}$;

步骤 4: 根据时间 $t=1, \dots, T$, 重复步骤 2 和步骤 3。

图 3 是该计算原理的示意图。

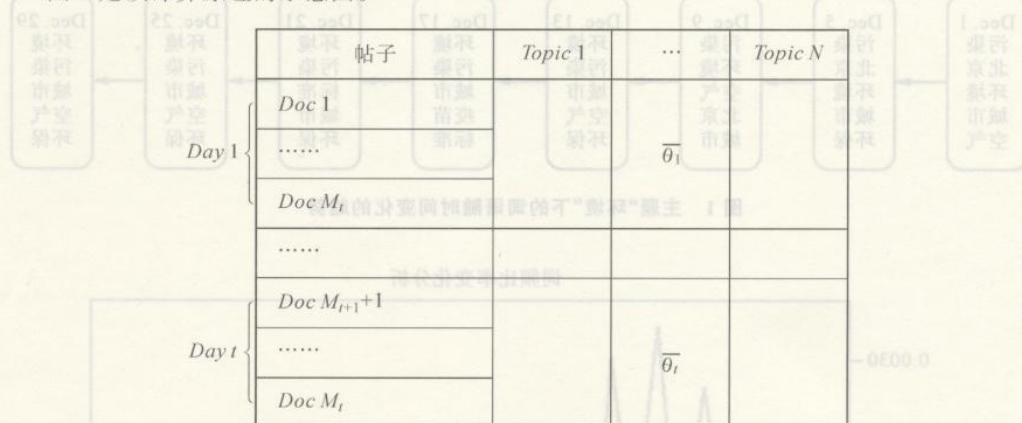


图 3 计算原理示意图

四个模型的计算结果如图 4 所示。在图中,横轴表示时间,纵轴表示主题,这里只按总的热度排名由高到低显示了其中 30 个综合热度较高的话题。每个方框代表在该时间点下该主题的平均热度,颜色代表平均热度值所在的区间,颜色越深则代表的热度值越高。

按上述算法可以得到四个模型中首发帖包含主题的热度值,如图 4 所示,我们可以很直观地感受到热度的变化,从而探寻到一些变化规律。首先综合来看,天涯杂谈网络舆情具备的一些特点为:(1)有关“人生”“家庭”“社会道德”“朋友”的话题持续保持较高的水平,很明显,发帖的网民关心的更多的是自己的生活;(2)一些常规性话题如“教育”“大学生”“医患”“经济”“文化”“政府与群众”“外交”等会维持一定的热度;(3)一些对权力的控诉如针对“土地拆迁”“村民选举”“法院”“派出所”等话题亦是发帖人阐述的主要内容;(4)当某些威胁到人们安全或利益的突发事件如“食品安全”“物价上涨”等发生时,人们会第一时间发帖申诉,形成热点话题。

另外,每个月又有一些特有的事件发生并引起讨论。如,2011 年 12 月份,“校车事故”的发生引起人们对校车安全的讨论;与时间相关的话题如“圣诞节”,则在节日的到来有一个小的波峰;“铁路售票”话题在元旦之前、开始售票时热度增加。

2012 年 1 月份,随着春节的到来,关于春运、订票、春晚等话题在节日前后成为人们关注的中心,一些社会事件如“张绍刚与刘莉莉”事件、“孔庆东骂香港”事件引起了不小的争论。“韩寒造假”事件在月末开始进入人们的视野。



图4 主题热度变化

2012年2月,“方舟子质疑韩寒”事件和关于韩寒本人的争论成为最热门的话题;“药家鑫”案的后续风波引起了人们的关注;“陶汝坤案”(一名高中男生因被拒而烧伤他人)的发生引起人们一片哗然。

2012年3月份,“方舟子质疑韩寒”事件继续发酵并在月末渐渐平息;关于“电话诈骗”的申诉开始增加;两会会议的召开引起人们对政治的关注。

5 结论

网民以网络为平台,在参与政治生活、发挥民主监督等方面发挥了积极的作用,然而由此引发的网络群体事件对社会稳定及人民生活亦造成了不良影响。网络舆情是社会舆情在当今社会的重要组成部分,有效把握网络舆情动态是实时诊断社会和谐总体水平与演化趋势的重要入口。

我们的研究从中国最大的论坛之一——天涯论坛的数据切入,试图将现实中这种 wicked 问题结构化。定性综合集成技术是解决该问题的有效方法之一。同时,主题模型在对文档集合进行挖掘方面亦取得了较大的成功,为我们把握大数据集的结构、中心思想和演化趋势提供了

新的解决办法。本文即采用了该方法对中国 BBS 的发帖内容进行了分析,试图挖掘时间对话题热度的潜在影响,并感知网络舆论的规律特点。当然该模型也存在一些不足,比如由于模型的复杂性使得模型的学习过程效率不够高。

不同于对在线行为分析研究^[12]的是,本文的分析更多关注于对内容的分析,从而挖掘网民的一些言论特点。未来我们的工作会考虑:分析更新帖数据,这种分析可能会更好地反映人们对热门事件的关注;将内容分析与行为分析结合,去挖掘点击量或回复量与话题之间的关系。

参考文献

- [1]张泽代,唐锡晋. 面向天涯论坛的 Web 挖掘的初步研究[A]. 第十一届全国青年系统科学与管理科学学术会议暨第七届物流系统工程学术研讨会论文集[C]. 武汉理工大学出版社, 2011:199-204.
- [2]唐锡晋. 两个定性综合集成支持技术 [J]. 系统工程理论与实践, 2010, 30(9): 1593-1606.
- [3]Tang, X. J., Luo, Bin. : Understanding college students' thought toward social events by qualitative meta-synthesis technologies [J]. International Journal of Organizational and Collective Intelligence, 2011, Vol. 2, No. 4: 15-30.
- [4]Tang, X. J.. Qualitative meta-synthesis techniques for analysis of public opinions for in-depth study [A]. J. Zhou (ed.). Complex 2009, Part II, LNICST 5 [C]. Springer, 2009: 2338-2353.
- [5]Blei, D. M., Ng, A. Y., Jordan, M. I.. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
- [6]Blei, D. M., Lafferty, J. D.. Dynamic Topic Models [A]. Proceedings of the 23rd International Conference on Machine Learning, 2006.
- [7]Alsumait, L., Barbara, D., Domeniconi, C.. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking [A]. ICDM'08, Eighth IEEE International Conference on [C]. IEEE, 2008: 3-12.
- [8]Wang, C., Blei, D., Heckerman, D.. Continuous time dynamic topic models [R]. Uncertainty in Artificial Intelligence[UAI], 2008.
- [9]Wang, X., McCallum, A.. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends [R]. KDD'06, USA, 2006.
- [10]Meng, C., Zhang, M., Guo, W.. Evolution of Movie Topics Over Time [EB/OL]. <http://cs229.stanford.edu/projects2012.html>.
- [11]吴丹,唐锡晋. 百度新闻热词初析[A]. 第十一届全国青年系统科学与管理科学学术会议暨第七届物流系统工程学术研讨会论文集[C]. 武汉理工大学出版社, 2011: 478-483.
- [12]Cui, L. J., He, H., Liu, W.. Research on hot issues and evolutionary Trends in network forums [J]. International Journal of u- and e- Service, Science and Technology, 2013, Vol. 6, No. 2: 89-97.
- [13]Chen, X., Li, J., Li, S., Wang, Y.. Hierarchical Activeness State Evaluation Model for BBS Network Community [A]. Communications and Networking in China (CHINA-COM), 7th International ICST Conference on [C]. IEEE, 2012: 206-211.