

文章编号: 1003-0077(2014)01-0047-09

特定事件微博与新闻报道话题对比研究

周振宇, 李 芳

(上海交通大学 计算机科学与工程系 中德语言技术联合实验室, 上海 200240)

摘 要: 该文描述了基于特定事件的新闻报道和微博在话题层面的对比研究。首先利用 LDA 话题模型抽取两种媒体上关于特定事件的话题, 然后提出了话题关注度、差异度、演化度的定义和计算公式, 改进了不同媒体话题差异度的计算方法, 最后, 选取四个不同种类的事件, 进行实验对比与分析, 结果显示, 关于同一事件, 1) 微博上评论性话题较多, 话题关注度值比较接近; 新闻报道上事实性话题较多, 话题关注度值差异较大; 2) 微博与新闻报道对评论性话题词汇差异度大, 事实性话题词汇差异度小; 3) 微博上评论性话题持续时间较长, 内容变化较少; 新闻报道上事实性话题持续时间较长, 内容变化较少。

关键词: 话题模型; 微博; 新闻报道; 对比

中图分类号: TP391

文献标识码: A

Comparing Topics from Microblog and News Media about Specific Events

ZHOU Zhenyu, LI Fang

(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: This work conducts a contrastive study on the topics of specific events from microblog and news media. Firstly, we use LDA to extract topics from the two media, and then define three indexes: attention factor, diversity factor and evolution factor for an improved topic discrepancy calculation. Finally, we chose four events of different types for experiments and analysis. The results show: 1) There are more comment topics appearing on microblog with close attention factors in contrast to a high proportion of factual topics with varied attention factors in the news media. 2) In both microblog and news media, diversity factor of words used in the comment topics is bigger than that in factual topics. 3) In microblog, comment topics last longer with consistent contents, while the factual topics does so in the news media.

Key words: topic model, Microblog, news, contrast

1 引言

现代社会科技发达, 传播媒体在人们获取信息的过程中扮演着非常重要的角色。新闻、报刊这样的传统媒体具有信息量大、客观性内容多、不够及时等特点, 是过去人们获取信息的唯一途径。随着科技进步以及 Web 2.0 时代的到来, 博客、论坛、微博客等新型媒体已成为人们青睐的途径, 与传统媒体相比, 它单篇信息量小, 更多地表达了网友自己的观点, 互动性强, 即时性强。对比新型媒体与传统媒体的研究工作基本上处于新闻工作者的感性认识阶

段。因此, 利用话题模型对两种媒体的内容进行自动分析, 研究同一事件两种媒体的话题对比, 审视两种媒体的差异具有重要的现实意义。

本文主要研究以下三个主要问题:

- 1) 关于特定事件, 微博话题与新闻报道话题有什么不同? 关注点有哪些不同?
- 2) 相同话题在两种媒体上词汇有哪些差异?
- 3) 相同话题在微博与新闻媒体上随时间的变化有何不同?

为了解决以上三个问题, 本文提出了衡量话题的三个指标: 关注度、差异度和演化度。根据量化的数据研究两种媒体对同一事件的内容差异。

本文的主要工作包括三个方面。首先,利用话题模型 LDA 对特定事件的语料建模,挖掘出该事件在两种媒体下被讨论的热点话题;然后定义话题的关注度、差异度和演化度,并分别给出它们的计算方法;最后,选取四个特定事件,给出两种媒体对同一事件不同的关注度,话题内容上的差异度,以及单个话题的内容演化。

本文的组织结构如下:第 2 节介绍相关的工作,第 3 节是研究方法的描述,第 4 节是实验结果和分析,第 5 节为结论及展望。

2 相关工作

目前,基于微博与新闻报道话题的抽取主要采用 LDA 模型^[1]及其扩展^[2-3]。LDA 是无监督学习方法,不需要训练数据,已在新闻报道的话题抽取中有广泛应用。Hong^[4]采用了 LDA 模型对 Twitter 上的话题进行抽取^[4],证明 LDA 方法在微博话题抽取中也是可行的。Zhao^[5]使用了 Twitter-LDA 模型,考虑到每篇微博的字数较少,这一模型融入了作者信息,将同一作者的微博合并为一篇文档,同时模型也融入了背景词信息,并设置了变量控制一个词是来源于背景词还是话题词。也有研究者利用了微博中 tag、表情作为标签,使用半监督的 Labeled-LDA^[6-7],很好地利用了微博的特点。

最近,有不少研究者提出了各种特征,对微博与新闻报道上的话题进行研究。Zhao^[5]比较了 Twitter 和 New York Times 上话题的类别与类型,将话题分为事件型、实体型、持续型三个类型,从分布、内容、覆盖程度、转发程度等方面比较话题在两种媒体上的区别。Ramage^[6]将话题分为物质类、状态类、风格类、社交类 4 种加以阐述,从整体上分析了 4 种类型话题的强度差异,还对比了两个用户(w3c 和 Oprah)的微博上 4 种类型话题的强度与内容差异。还有不少研究者针对特定事件分析微博话题^[8-9]。Qu^[8]分析了玉树地震后的微博内容,力图找出灾难后人们主要谈论什么话题,不同类型的话题的发送与转发行为是否不同,以及它们是如何传播的。研究发现灾难后人们关注的重点是发表观点、描述事件、捐赠默哀等话题。不同类型的话题的发送与转发行为也不同,事件刚发生时往往是描述事件居多,随着时间的推移,人们更多关注灾后重建,并在哀悼日发送表达感情的内容。研究还发现人们更乐意转发介绍救灾行动和事件情况的内容,从转发的平均

深度来看,也是行动类内容最多。

本文涉及了微博与新闻报道话题的对比,所以如何将不同的话题关联起来至关重要。Zhao^[5]使用了传统的 JS 距离的方法;Chu^[10]提出,直接利用 JS 距离有时候会忽略了话题侧重点的区别,故在传统的计算方法上融合了特征词的概念。将 LDA 话题结果中概率值较大的词作为特征词,通过计算特征词的余弦相似度,来修正分布的距离。

本文和文献^[5]的研究目的相同,区别是选取了特定事件的话题进行两种媒体的分析对比,提出了话题的关注度、差异度和演化度计算方法。本文与文献^[8]的不同之处使用了话题模型进行话题的抽取,以及分析了特定事件在微博和新闻报道上的话题,主要对两种媒体在话题层次上进行对比。

3 研究方法

本文从话题层面对微博与新闻媒体进行对比研究。首先针对特定事件,挖掘两种媒体上的语料;然后对两种语料应用 LDA 建模,挖掘潜在话题;接下来对两种媒体上的话题进行关注度计算,对比两种媒体不同的话题关注。然后研究两种媒体相同话题在词汇与语义上的差异度。最后通过演化度来观察两种媒体上的话题随时间的变化趋势。

在本文中,我们主要讨论两种不同话题。

1) 评论性话题:人们对某一现象或实体的评论,如呼吁停止冷漠,对道德现状的鞭答,用俚语调侃独裁者等。

2) 事实性话题:对客观事实的描述,如对目击者的采访,对病情的进展报道,对各界悼念的报道等。

3.1 话题建模

LDA 模型是一个生成概率模型,是三层的变参数层次贝叶斯模型。首先假设词由话题的概率分布混合产生,而每个话题是在词汇表上的一个多项式分布;其次假设文档是潜在话题的概率分布的混合;最后针对每个文档从 Dirichlet 分布中抽样产生该文档包含的话题比例,结合话题和词的概率分布生成该文档中的每一个词汇。本文对两个事件在两种媒体上的语料集按时间先离散建模,得到事件在两种媒体下,各时间段的多个话题结果。表 1 是文中使用到的符号。

表 1 文中使用到的符号

符号	符号的描述
α	LDA 模型的 <i>Dirichlet</i> 先验参数,表示文档-话题分布的先验
β	LDA 模型的 <i>Dirichlet</i> 先验参数,表示话题-词分布的先验
K	话题个数
V	词汇表
d	文档
z	话题
θ_d	文档 d 上的话题分布
p_z	话题 z 上的词汇分布
$AF(z)$	话题 z 的关注度
$DF(z)$	话题 z 的差异度
$EF(z)$	话题 z 的演化度

3.2 话题关注度计算

话题的关注度是衡量该话题被谈论的程度,即在新闻报道或微博中该话题所占的比例。LDA 建模后可以得到话题在各文档中的概率分布。我们可以计算话题在所有文档上的平均概率作为关注程度的衡量。但是一篇只有几个字或几十个字的微博,经过分词、去除停用词等处理之后,剩下的有效词数很少。实验中我们发现有的微博的有效词语可能只有一个,当这个词语被分给某话题后,该话题的概率为 1。故而对不同的文档字数,赋予不同的权值,从而使计算上更具科学性。我们定义话题 z 在某天的覆盖程度为式(1)。

$$s(z) = \sum_{d_i \in D} \theta_{d_i, z} \varphi_{d_i}$$

(1)

其中 D 为当日的文档(新闻报道、微博)全集, θ 是话题在文档上的分布, φ 根据文档字数多少而确定的权值。这个强度的指标衡量了一个话题在某日在所有文档中的关注度。在不同媒体间进行比较时,我们定义话题 z 的关注度 $AF(\text{attention factor})$ 的计算公式为强度归一化的值,如式(2)所示。

$$AF(z) = \frac{s(z)}{\sum_{z_i \in T} s(z_i)}$$

(2)

其中 T 是当日的话题全集。

3.3 话题差异度计算

话题的差异度是衡量新闻媒体与微博上相同话

题的差异度,用话题词汇分布的距离来计算。话题距离通常采用 JS 距离来计算,但对于本文的研究语料,如果直接使用 JS 距离,其效果较差。这是由于两种媒体本身用词的差异。如微博上人们可能会使用一些较为口语化的词汇,而新闻报道上可能更多地使用较为正式和官方的词汇。事实上, LDA 建模后,每个话题表示为具有相同语义的词汇集合。定义话题的词汇表示:

考察词在话题 z 上的分布 p_z ,若对于某词 w ,有 $p_z(w) > \xi$, ξ 为阈值,则认为 w 是话题 z 的词汇,记作 $w < z$ 。记话题 z 的词汇集为

$$D(z) = \{w \mid w < z, w \in V\}$$

其中 V 是词汇表。

假设话题 z 在两种媒体上分别表现为话题 z_1 和 z_2 。定义它们词汇的交集与并集:

$$\text{交集 } z_1 \cap z_2 = D(z_1) \cap D(z_2)$$

$$\text{并集 } z_1 \cup z_2 = D(z_1) \cup D(z_2)$$

从公式可以看出,交集表示两种媒体共有的词汇,并包含了它们所有的词汇。共有词汇数目越多,话题语义关系越接近,反之不同词汇数目越多,话题语义差异越大。定义词汇差异度 v_{z_1, z_2} 为 $\frac{|z_1 \cup z_2| - |z_1 \cap z_2|}{|z_1 \cup z_2|}$ 。JS 距离与词汇差异度值的取值均在 $(0, 1)$ 之间,用 λ 来控制词汇差异度值的权重。定义话题 z 在两个媒体的差异度 $DF(\text{diversity factor})$ 的计算公式为:

$$DF(z) = (1 - \lambda)JS\text{div}(p_{z_1} \parallel p_{z_2}) + \lambda v_{z_1, z_2}$$

(3)

其中 $JS\text{div}(p_{z_1} \parallel p_{z_2})$ 是话题 z_1 与 z_2 的 JS 距离。

3.4 话题演化度计算

话题演化度是衡量同一媒体相同话题随时间的变化。由于 LDA 的结果表征了话题在文档上的分布,以及词汇在话题上的分布。在讨论话题演化度的时候,通过计算话题间的语义相似度来表征,采用常用的 JS 距离(Jensen-Shannon divergence)来判断话题之间是否存在演化关系。之所以不像上一节中对 JS 距离进行修正,是因为对于同一种媒体形式来说,它的词汇使用的差异并不大。假设微博(新闻报道)上的某话题 z ,它在某时间段 t 表示为 z_t ,词汇表 V_t 的分布是 p_{z_t} ;在时间段 $t+1$ 上表示为 z_{t+1} ,词汇表 V_{t+1} 的分布是 $p_{z_{t+1}}$ 。由于词汇表 V_t 与 V_{t+1} 是取自两个不同的时间段,维度并不相同。故欲计算两个分布的距离之前须先统一维度,扩充词汇表。扩

充方法参照 Chu^[10]的方法：将两个词汇表合并，并置话题中未出现的词的分配次数为 0。则定义话题的演化度 EF(evolution factor)计算公式为两个分布的 JS 距离：

$$\begin{aligned} EF(z) &= JSdiv(p_{z_t} \parallel p_{z_{t+1}}) \\ &= \frac{1}{2}(KLdiv(p_{z_t} \parallel m) \\ &\quad + KLdiv(p_{z_{t+1}} \parallel m)) \end{aligned} \tag{4}$$

其中 $m = \frac{1}{2}(p_{z_t} + p_{z_{t+1}})$

4 实验结果与分析

本文主要针对微博与新闻报道上特定事件的话题进行多方面对比研究。有的事件由微博引发，有的事件则是由新闻报道引起。针对不同类型的事件进行对比，可以更好地研究出两种媒体形式的差异。我们选取了 2011 年 4 个比较有影响力的事件作为语料集。

1. 小悦悦事件：2 岁女童小悦悦被车碾压，18 个路人过路不闻，最终女童死亡。
2. 金正日事件：朝鲜最高领导人金正日在去视察的火车上因过度疲劳而去世。
3. 华尔街事件：示威者试图占领华尔街，声称反对美国的权钱交易及严重的金融危机。
4. 天宫一号事件：中国自主研发的首个空间目标飞行器天宫一号发射升空，并与神舟八号对接。

之所以选取这 4 个事件，是因为这 4 个事件分别为民生类事件、人物类事件、政治类事件、科技类事件，且 1、4 两个事件为国内事件，2、3 两个事件为国外事件。且这 4 个事件在其所在的时间段内均为热点事件，网络上讨论较多。故选取这 4 个事件进行本文的研究具有一定的代表性。

其中新闻报道部分均采用了新浪新闻下关于这两个事件的新闻报道全文的集合；微博部分采用新浪微博提供的 API 进行实时收集，直接使用“金正日”、“小悦悦”、“华尔街”、“天宫一号”作为关键词进行检索得到的微博、去除重复出现超过 20 次的微博、以及微博中所有的 hashtag。

实验包括 3 个方面，一是在同一时间点上，研究两种媒体的话题关注度；二是通过式(3)计算话题差异度，研究相同话题在两种媒体上的词汇差异性；三是通过演化度的计算确定话题的演化路径，研究话题随时间的变化，以及这种变化在两种媒体上有什

么不同。

4.1 实验数据

实验数据分为四组：如表 2 所示。语料均为全文，并过滤停用词、hashtag。实验使用了开源的 Gibbs Sampling 工具，话题个数 K 设为 6，模型参数 α, β 分别设置为 50/K 和 0.01。关注度计算中的权值 φ 的取值为：文档字数少于 2 时为 0.2，文档字数在 3 到 5 之间为 0.45，文档字数在 6 到 9 之间为 0.6，文档字数大于 10 为 0.8。话题差异度计算中，话题词汇表示的阈值 ξ 设为 0.4，距离式中的词汇差异度权值 λ 设为 0.3。标准 JS 距离的阈值 η_{JS} 为 0.8，话题差异度的阈值 η_{DF} 为 0.64。

表 2 四个事件的话题条数、持续时段

事件	微博语料条数	新闻报道 语料篇数	话题持续时段 (2011 年)
小悦悦	208 546	162	10/17~10/28
金正日	429 539	87	12/19~12/31
华尔街	243 202	48	10/04~10/19
天宫一号	296 532	126	09/29~10/11

4.2 话题关注度分析

根据式(2)分别计算 4 个事件微博和新闻话题的关注度。表 3~6 分别列出 4 个事件前 5 天的关注度最高的三个话题的 top3 话题词与关注度值。

从结果可以看出，微博上的话题，人们谈论的较多的是评论性话题。如小悦悦事件中，人们纷纷呼吁要停止冷漠，以及强烈谴责路人的这种见死不救的行为；金正日事件中，人们纷纷表示悼念，以及表示他是中国人民的朋友；华尔街事件中，人们表示对占领行为的声援；天宫一号中，人们对成功发射而欢呼。而对事实性话题，如小悦悦事件中政府援助问题，金正日事件中外国的表态和事件造成的经济影响，华尔街事件中各方面抗议活动，天宫一号各种背景知识这样的话题的关注度则相对较小。而对于新闻报道上的话题，可以看到，基本上都是事实性话题，如小悦悦事件中对小悦悦病情的介绍，政府出台保障政策，金正日去世事件中外国的悼念与表态，发布讣文信息，华尔街事件中对事件进程的报道，天宫一号事件中对工程各方面的详细介绍等。诸如小悦悦事件中对道德的反思，金正日去世事件中对金正日独裁事实的批判，华尔街事件中对资本家的揶揄，

表 3 金正日事件两种媒体上每日关注度最高的三个话题(top3 话题词与关注度,前 5 日)

	话题	2011/12/19	2011/12/20	2011/12/21	2011/12/22	2011/12/23
微博	1	全世界,日头,播报,0.189 6	世界,民众,知道,0.188 7	逝世,悼念,金日成,0.180 2	金正恩,儿子,留给,0.195 6	默哀,联合国,大会,0.203 4
	2	卡扎菲,知道,萨达姆,0.179 2	中国人,同志,悲痛,0.170 6	领袖,知道,世界,0.173 8	去世,哀悼,平壤,0.171 2	金正恩,吊唁,政府,0.18
	3	逝世,领导人,最高,0.172 2	中国,萨达姆,卡扎菲,0.169 6	中国,生于,萨达姆,0.173 5	完全,逝世,情感,0.170 1	逝世,领导人,去世,0.177 4
新闻报道	1	韩国,去世,消息,0.240 9	韩国,表示,逝世,0.222 2	吊唁,同志,中国,0.221 2	表示,逝世,哀悼,0.185 2	进行,社论,同志,0.216
	2	逝世,最高,政府,0.193	劳动党,委员会,最高,0.201 6	韩国,政府,哀悼,0.197 8	日本,政府,关系,0.183 7	金正恩,表示,领导,0.202
	3	日本,消息,已经,0.182	去世,美国,韩国,0.157 2	半岛,平壤,表示,0.192	吊唁,平壤,领导人,0.167 5	韩国,经济,中国,0.170 8

表 4 小悦悦事件两种媒体上每日关注度最高的三个话题(top3 话题词与关注度,前 5 日)

	话题	2011/10/17	2011/10/18	2011/10/19	2011/10/20	2011/10/21
微博	1	路人,漠视,司机,0.179 4	帮助,身边,关心,0.176 8	冷漠,帮助,良心,0.178 3	冷漠,成都,良心,0.177 7	一路走好,孩子,天堂,0.187
	2	情况,婆婆,中国人,0.171 2	希望,世界,社会,0.171 8	社会,道德,中国,0.170 3	生命,路人,正义,0.172 4	天堂,世界,希望,0.178 7
	3	危重,无奈,父亲,0.165 1	良心,人性,冷漠,0.166 4	路人,冷漠,社会,0.167 6	社会,世界,人性,0.167 1	冷漠,一路走好,路人,0.166 9
新闻报道	1	司机,肇事,医院,0.238 8	陈贤妹,医院,媒体,0.208 3	社会,医院,女士,0.211	医院,媒体,抢救,0.201 4	医院,司机,情况,0.253
	2	金城,警方,男子,0.194	下午,阿姨,昨日,0.196 1	司机,下午,媒体,0.196 3	社会,道德,冷漠,0.189 7	功能,出现,抢救,0.248 9
	3	阿姨,小孩,时分,0.167 2	司机,肇事,妈妈,0.160 4	道德,见死不救,法律,0.173 1	见义勇为,人员,奖励,0.188 3	父母,社会,道德,0.187 7

表 5 华尔街事件两种媒体上每日关注度最高的三个话题(top3 话题词与关注度,前 5 日)

	话题	2011/10/04	2011/10/05	2011/10/06	2011/10/07	2011/10/08
微博	1	占领,华尔街,运动,0.182	华尔街,美国人,占领,0.178	占领,运动,华尔街,0.185	华尔街,占领,民主,0.178	华尔街,占领,美国人,0.183
	2	美国人,世界,媒体,0.180	日报,苹果,华尔街,0.171	华尔街,苹果,仗义,0.168	华尔街,日报,乔布斯,0.174	华尔街,占领,减肥,0.183
	3	中国,日报,华尔街,0.164	华尔街,占领,纽约,0.169	华尔街,日报,公司,0.166	华尔街,占领,美国人,0.170	华尔街,占领,运动,0.167
新闻报道	1	示威者,纽约,示威,0.228	美国,示威者,3 日,0.226	占领,高校,运动,0.201	游行,工会,华尔街,0.242	美国,抗议,活动,0.201
	2	活动,抗议,经济,0.224	占领,活动,愤怒,0.212	游行,纽约,工会,0.186	抗议,占领,华盛顿,0.167	美国,华尔街,运动,0.180
	3	华尔街,占领,抗议,0.171	华尔街,巨额,理解,0.152	美国,抗议,活动,0.163	示威,示威者,加入,0.164	占领,金融,民众,0.176

表 6 天宫一号事件两种媒体上每日关注度最高的三个话题(top3 话题词与关注度,前 5 日)

	话题	2011/09/29	2011/09/30	2011/10/01	2011/10/02	2011/10/03
微博	1	成功,发射,天宫一号,0.185	成功,发射,天宫一号,0.195	发射,成功,天宫一号,0.186	升空,天宫一号,热烈,0.198	天宫一号,燃料,北航,0.196
	2	天宫一号,发射,成功,0.181	发射,升空,成功,0.175	天宫一号,关注,今晚,0.171	天宫一号,发射,成功,0.174	太空,国庆,原来,0.173
	3	天宫一号,发射,直播,0.171	天宫一号,成功,我国,0.168	今天,骄傲,祖国,0.164	天宫一号,天宫,航天,0.171	发射,成功,庆祝,0.165
新闻报道	1	发射,中心,北京,0.290	发射,火箭,卫星,0.188	轨道,天宫,环境,0.214	时间,观测,朱进,0.283	火箭,变化,控制系统,0.206
	2	天宫一号,进行,航天员,0.175	对接,载人,空间站,0.183	天宫一号,中国,发射,0.179	任务,测控,航天,0.221	测控,任务,载人,0.187
	3	航天,飞行,系统,0.164	天宫一号,任务,空间,0.162	飞行器,天宫一号,进行,0.174	太空,中国,天宫一号,0.140	航天员,太空,设备,0.178

天宫一号事件中骄傲的心情等话题没有出现在新闻报道中,小悦悦事件中的保护见义勇为和金正日去世事件中的与中国关系都排在靠后的位置。

两者各有一些独有话题,金正日事件中微博独有的话题是关于金正日的俚语,新闻报道是对平壤实况的介绍。这样的独有话题也反映了两种媒体各自的特点,即微博人们可以随意戏谑,新闻报道更乐于展现事件的实况。

从以上结果可以看出,微博主要谈论的内容是对现象或人的评论,而新闻报道更侧重于客观事实,基本上处于大体上相反的局面。这也正说明了微博作为一种新兴媒体,它可以使广大网民直抒胸臆,参与度远远高过新闻报道。网民从新闻报道中被动接受客观信息,而在微博中表达自己的主观倾向。

从两者的关注度值来看,微博上的话题关注度差异远不及新闻报道上的差异。即使将话题数定在 6,新闻报道上强度较大的话题的关注度都超过了 1/5,最少的仅有 1/8 左右。而且对于相类似的话题,在不同时间上的关注度波动很大。反观微博上,6 个话题的相对强度均在 1/6 上下。说明微博上的话题本身的区别没有新闻报道上那么明显。

4.3 话题差异度分析

首先通过对比实验验证本文提出式(3)的有效性。我们对 4 个事件每天抽出的话题进行人工比对,用式(3)分别计算两种媒体上每天的任两个话题之间的差异度,通过阈值确定相同话题,人工评判其准确性。以 JS 距离为 baseline,表 7 是 4 个事件的计算结果。

表 7 应用 JS 距离与式(3)计算结果

事件	方法	Precision	Recall	F1
小悦悦	差异度(3)	0.687 5	0.702 1	0.694 7
	JS 距离	0.647 3	0.685 5	0.665 9
金正日	差异度(3)	0.545 4	0.631 6	0.585 4
	JS 距离	0.428 6	0.552 6	0.482 8
华尔街	差异度(3)	0.575 6	0.701 2	0.632 2
	JS 距离	0.535 5	0.678 3	0.598 5
天宫一号	差异度(3)	0.520 5	0.626 3	0.568 5
	JS 距离	0.529 8	0.634 3	0.577 4

可以看出,本文提出的式(3)相比于直接使用 JS 距离,除天宫一号事件外,精度与召回率均有所提升。其中金正日事件的提升较为明显。

通过图 1 实验,式(3)定义差异度的阈值设为 0.64 时 F 值较高。将低于这一阈值的话题定为相同话题,以便于后续对相同话题进行差异度分析。

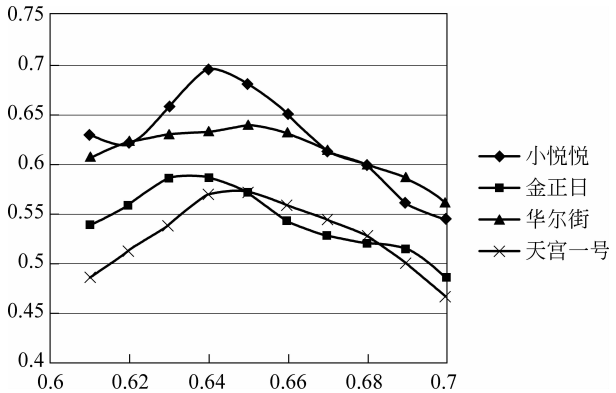


图 1 差异度计算式(3)的阈值-F 值曲线

差异度分析将主要侧重于对相同话题在不同媒体上的词汇差异进行分析。我们选取金正日事件第 2 日、小悦悦事件第 4 日和天宫一号第 1 日的各 3 组相同话题，观察它们的差异度，如表 8 所示。

由式(3)知，差异度越接近于阈值，则语义的差异越明显。从表 3~6 中可以看出事实性话题，如外国悼念情况、经济影响情况、政府保障老人这样的话

题，两种媒体的差异度相对较小，故词汇上的相似度较大。

比如两种媒体在外国悼念这一话题在语义上高度相似，基本都表达了韩国、美国、日本 3 个国家的表态，略有不同的是微博上谈到的表态以韩国方面的哀悼为主，而新闻报道上更强调美日基本保持半岛稳定。

表 8 两种媒体上 6 组相同话题的话题词对比

事件	话题	媒体	话题词	式(3) 差异度	Baseline 差异度
金正日	相同话题 1	微博	去世,韩国,逝世,哀悼,表示,美国,政府,时代,日本,关注	0.428 6	0.607 4
		新闻报道	韩国,表示,逝世,半岛,美国,总统,日本,稳定,政府,消息		
	相同话题 2	微博	逝世,领导人,消息,经济,最高,去世,遗产,影响,平壤,华盛顿	0.499 1	0.713 1
		新闻报道	去世,美国,韩国,权力,经济,进行,问题,领导人,发生,军方		
	相同话题 3	微博	金正恩,全世界,日头,领袖,生活,国度,独裁者,金日成,媒体,历史	0.611 4	0.732 5
		新闻报道	金正恩,接班人,父亲,军事,委员长,媒体,国际,工作,成为,电影		
小悦悦	相同话题 1	微博	道德,法律,问题,帮助,社会,中国人,老人,援助,提供,最后	0.540 0	0.771 4
		新闻报道	社会,道德,冷漠,陈贤妹,建议,主任,广东省,行为,好人,政府		
	相同话题 2	微博	见义勇为,保护,立法,见死不救,见义勇为者,法律,行为,诬陷,责任,惩罚	0.565 2	0.776 4
		新闻报道	网友,香港,责任,出现,代表,见死不救,司机,保护,生命,问题		
	相同话题 3	微博	社会,世界,人性,爱心,希望,良知,媒体,悲剧,温暖,善良	0.624 1	0.818 8
		新闻报道	法律,路人,父母,救助,最高,美国,好事,律师,希望,事情		
天宫一号	相同话题 1	微博	成功,发射,天宫一号,祝贺,激动,庆祝,预祝,正常,热烈,good	0.611	0.792 8
		新闻报道	中国,成功,重要,天宫一号,航天,空间站,世界,国际,计划,世界		
	相同话题 2	微博	中国,发射,太空,对接,航天,升空,目标,空间站,飞行器,酒泉	0.420 4	0.600 5
		新闻报道	对接,载人,交会,空间,发射,飞船,技术,任务,目标,工程		
	相同话题 3	微博	骄傲,天宫一号,中国,顺利,祖国,中国人,祝福,航天,希望,世界	0.617	0.835 3
		新闻报道	中国,成功,重要,天宫一号,航天,空间站,世界,国际,计划,世界		

在独裁这一话题上，两种媒体都谈到了金正恩，但是从词语上看，差异较大，微博上主要讲的是独裁问题，是对金家三代领导人世袭的一种讽刺性的评价，而新闻报道上则主要谈到了金正恩接班的问题。

从以上的结果可以看出，从内容上看，越是事实性话题，两种媒体的差异度就越小，而越是评论性话题，两种媒体的差异度越大。

4.4 话题演化度分析

差异度着重研究同一时间点上两种媒体间的用词差异，而演化度则着重于研究话题在整个时间段的趋势变化在两种媒体上有什么不同。

我们通过计算相邻时间各话题间的演化度，得到话题的演化路径。在小悦悦事件中，我们选取“道德”这一评论性话题，观察这一话题在两种媒体上随时间的变化，如图 2 所示。

从上图可以看出，在整个时间段内，在微博上人们讨论的道德话题基本都围绕着对中国现今社会的道德问题各抒己见，着重谴责现在的中国社会缺少道德与良知。主要的谈论核心点在整个时间段上的变化不大。反观新闻报道上的道德话题，从高频词不断变化就可以看出，话题的着重点随着时间呈现一定的变化。如 19 日政府开会讨论见死不救的问题，18 日和 22 日均是对网友热议道德问题的报道，

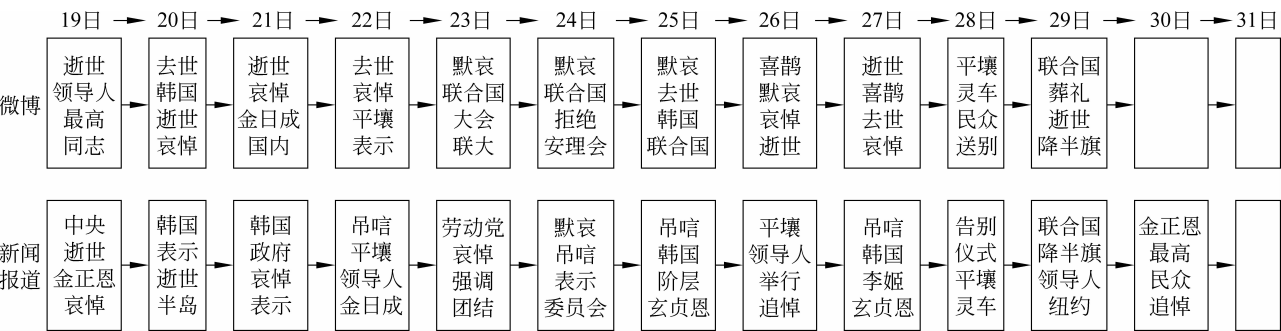


图2 小悦悦事件中“道德”话题在两种媒体上随时间的变化(前4个话题词)

而23日则提到了政府提倡的道德文化建设,27日又提到了对老百姓行为的讨论。整个时间段上关于道德的内容变化较大。

对于金正日去世事件,我们考察“悼念”这一事实性话题,如图3所示。由表8可知,在事件的初期,“悼念”话题在微博和新闻报道上的差异度是较小的,但是随着时间的变化,这一话题在两种媒体上的着重点也在发生着变化。

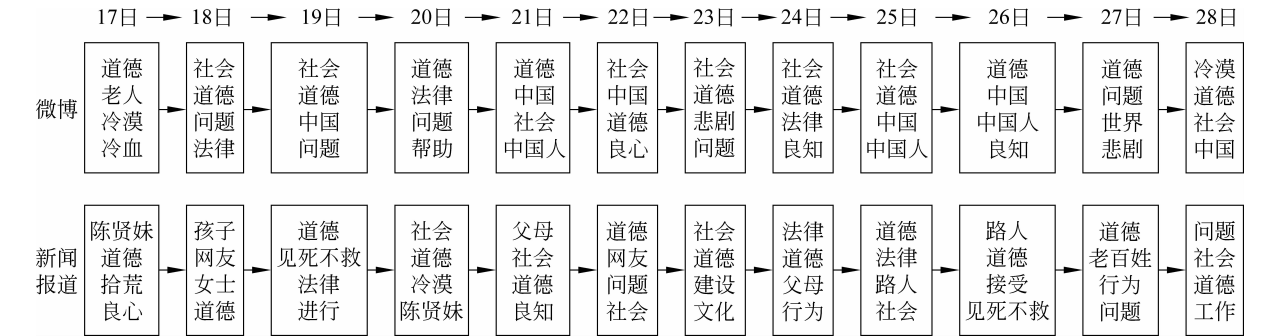


图3 金正日去世事件中“悼念”话题在两种媒体上随时间的变化(前4个话题词)

从演化图上可以看出,“悼念”话题在30日左右便趋于消亡。在22日前和28日后,微博和新闻报道上讨论“悼念”的内容差别并不大。这主要是基于“悼念”是一个事实性话题。但是在23~27日这一时间段内,微博上讨论了安理会拒绝为金正日默哀,和平壤的喜鹊也为金正日“哀悼”,同期的新闻报道上未见相关内容。从讨论内容可以看出,微博上人们更乐于讨论一些较随意的内容,特别是新闻报道为了宣传需要而不方便报道的内容。这也体现了微博话题的随意性。比较而言,新闻报道则更侧重于客观事实。同时为了宣传需要,也会刻意隐去一些相关报道。

5 结论与展望

本文首先使用LDA话题建模,发现两种媒体中隐含的话题。接着,使用三个指标——关注度、差异度和演化度去研究评论性话题与事实性话题在两种媒体上的受关注程度、用词的差异和演化趋势。根据两个特定事件的实验结果可以得到以下结论:

- 1) 关于特定事件,两种媒体上的话题不完全相同。微博上评论性话题较多,且关注度较高,新闻报道则是事实性话题较多,关注度较高。带有调侃性的话题(如金正日去世事件中的俚语调侃)是微博上的特有话题,而纯粹描述事件进程的话题(如采访事发、病情恶化)是新闻报道上的特有话题。同时,微博上的话题之间的关注度差异不大,但新闻报道上的话题的关注度差异很大,即使是类似的话题,在不同时间的关注度波动也很大。
- 2) 评论性话题在两种媒体中的用词差异较大,这也反映了网友在评论或发表看法时的用词与新闻报道正规措词存在很大差异。而事实性话题在两种媒体中的词汇差异较小。如“独裁”这样的评论性话题,微博上的重点词汇有“独裁者”、“国度”、“金日成”等,而新闻报道上使用的词汇是“接班人”“委员长”等。而“哀悼”这样的事实性话题,两种媒体中的主要词汇都集中在“逝世”、“哀悼”、“韩国”、“美国”、“表示”等,差异较小。
- 3) 微博上评论性话题持续时间较长,话题内容随时间变化较小,事实性话题反之;新闻报道事实性

话题的持续时间较长,内容随时间变化较小,评论性话题反之。如“道德”这样的评论性话题在微博上一直持续,且内容基本都是对道德沦丧的斥责,新闻报道上该话题内容随时间一直有所变化;而“悼念”这样的事实性话题,在新闻报道上一直持续且内容变化不大,但在微博上,内容随时间不断变化。

今后的工作将考虑如何进一步更严谨地探索话题间的关联,从更多的角度去分析两种媒体间话题的差异性。特别是针对更多种不同类型的话题,如自然灾害类话题、社会民生类话题、政治事件类等。这些话题有的起源于微博,有的起源于新闻报道,这些特点也可能在话题的差异分析中体现。

参考文献

[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] Blei D M, Lafferty J D. A Correlated Topic Model of Science[J]. The Annals of Applied Statistics 2007,1 (1):17-35.

[3] Blei D M, Lafferty J D. Dynamic Topic Model[C]// Proceedings of International conference on Machine Learning, 2006: 113-120.

[4] Liangjie Hong, Davison B D. Empirical study of topic

modeling in Twitter[C]//Proceedings of the SIGKDD Workshop on SMA,2008.

[5] Xin Zhao, Jing Jiang, JianshuWeng, et al. Comparing Twitter and traditional media using topic models[C]// Proceedings of the European Conference on Information Retrieval, 2011.


[6] Ramage D, Dumais S, Liebling D. Characterizing Microblogs with Topic Models [C]//Proceedings of AAAI on Weblogs and Social Media, 2010.

[7] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2009.


[8] Yan Qu, Chen Huang, Pengyi Zhang, et al. Microblogging after a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake[C]//Proceedings of the ACM 2011 conference on Computer supported cooperative work, 2011: 25-34.

[9] Vieweg S, Hughes A L, Starbird K, et al. MicrobloggingDuring Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness[C]//Proceedings of the 28th International Conference on Human factors in computing systems, 2010: 1079-1088.

[10] 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): 1501-1506.




周振宇(1988—),男,硕士,工程师,主要研究领域为信息检索与抽取。
E-mail: lesbuy@qq.com




李芳(1963—),女,博士,副教授,主要研究领域为自然语言处理、信息检索与抽取。
E-mail: fli@sjtu.edu.cn


(上接第 46 页)



魏明川(1987—),男,硕士,主要研究领域为文本挖掘、舆情监测。
E-mail: weimingchuan@software.ict.ac.cn



朱俊杰(1986—),男,硕士,主要研究领域为文本挖掘、推荐系统。
E-mail: zhujunjie@software.ict.ac.cn



张瑾(1978—),男,博士,高级工程师,主要研究领域为舆情分析、自然语言处理、话题分析、分布式处理等。
E-mail: jinzhang@ict.ac.cn