理论探索

突发公共卫生事件的微博主题演化模式和时序趋势

——以 Twitter 和 Weibo 的埃博拉微博为例 *

安 璐[1] 杜廷尧[1] 余传明[2] 周利琴[1] 李 纲[1]
([1]武汉大学信息管理学院 湖北 430072;
[2]中南财经政法大学信息与安全工程学院 湖北 430073)

摘要 文章利用潜在狄利克雷分配(LDA)模型和自组织映射(SOM)方法比较分析了 Twitter 与 Weibo 平台上关于西非埃博拉(Ebola)病毒爆发的微博热点主题类别,揭示其演化模式和时序趋势的异同点,最后根据这些特点为突发公共卫生事件管理部门的应急决策提供了实际的建议。

关键词 时序分析 主题演化模式 埃博拉爆发 微博 突发公共卫生事件

Microblogging Topic Evolution Pattern and Timing Trends of Public Health Emergencies: Taking Ebola Microblogging on Twitter and Weibo for Example

An Lu[1] Du Tingyao[1] Yu Chuanming[2] Zhou Liqin[1] Li Gang[1]
([1]School of Information Management, Wuhan University, Hubei, 430072;
[2]School of Information and Safety Engineering, Zhongnan University of Economics and Law, Hubei, 430073)

Abstract The authors utilize the Latent Dirichlet Allocation (LDA) model and Self-Organizing Map (SOM) technique to compare and analyze similarities and differences between salient topic categories and characterize their evolution patterns and temporal trends in microblogs regarding Ebola outbreak in West African on Twitter and Weibo. Salient topic categories were identified and compared on two platforms. Finally, practical suggestions were provided for public health emergency response departments according to those characteristics.

Keywords temporal analysis, topical evolution pattern, Ebola outbreak, microblog, public health emergency

1 引言

突发公共卫生事件由于其影响范围广、危害人们的身体健康甚至生命等因素一直受到政府部门和公众的高度关注。随着社交媒体的快速发展,各种社交平台例如美国的 Twitter 和中国的新浪微博已经成为公众获取和发布包括突发公共卫生事件在内的各种重大事件相关信息与观点的重要渠道。自 2014 年 2 月起,埃博拉病毒在西非肆虐,据世界卫生组织统计,截至 2015 年 7 月 12 日埃博拉病毒已造成超过 11 000 人死亡[1]。各微博平台上也产生了数十万条有关埃博拉爆发的微博。据统计,每条有关埃博拉病毒的新闻视频都会衍生出几万条微博和互联网搜索[2]。自 2015 年 5 月起,塞卡病毒在美洲蔓延,截至目前已有 34 个国家报告发现塞

* 本文系国家自然科学基金青年项目"突发公共卫生事件社交媒体信息主题演化与影响力建模"(编号:**71603189**)、国家自然科学基金面上项目"大数据环境下基于领域知识获取与对齐的观点检索研究"(编号:**71373286**)的研究成果之一。

卡病毒原地传播病例[3],新浪微博上亦生成 8100 余条相关微博,后续还可能继续增长。

为了探索微博平台上与突发公共卫生事件相关的微博主题类别的时序特征,本文以 2014 年 2 月以来有关西非埃博拉爆发的微博作为调查对象,其研究目的包括:(1)探测与突发公共卫生事件相关的微博在不同时期的热点主题;(2)勾勒与突发公共卫生事件相关的微博主题演化模式和时序发展趋势;(3)揭示比较突发公共卫生事件背景下中英文微博平台的使用模式。其研究发现有助于突发公共卫生事件应急响应部门更好地理解突发公共卫生事件的发展轨迹以及公众在不同阶段的关注点,以便在类似的情形下有效地采取措施对抗传染病疫情。

2 相关研究

2.1 微博主题分析

微博的主题分析分为各类主题的微博分析与特定主题的微博分析,其研究已取得了丰硕的成果,常见的研究方法包括统计分析和手工分类。当研究的微博涉及各类主题时,例如研究不同国家与地区的用户所发布微博的主题相关性[4],通常采用较为粗略的主题分类。当研究的微博限定在某一特定主题时,则识别更加细粒度的主题类别,例如识别公众对埃博拉病毒的主要关注方面的主题[5]、将韩国与心血管护理相关的 Facebook 发帖分为若干类别[6],然而,对主题的手工分类限制了被调查的微博数量。

由于突发事件常常会引发大量的微博消息,研究人员通常抽取其中的一小部分作为样本开展研究[7]。为了自动分析大规模微博的主题,主题建模是一个不可缺少的步骤。潜在狄利克雷(Latent Dirichlet Allocation,简称 LDA)和词频向量是普遍采用的主题建模方法[8]。Anantharam 等人[9]还开发了其他的技巧,从主题一致的微博中分离出主题异常的微博。然而,许多研究仍然对突发事件的微博主题以静态的方式展开研究,或是分析其在数量上的动态变化,而主题内容的演化则探索不足。

2.2 微博时序分析

为了揭示微博的时序特性,研究人员提出了一些新颖的主题模型、框架和方法,例如趋势敏感的潜在狄利克雷(TS-LDA)[10]、随时间变化的主题模型[11]、多分面主题建模的统一框架[12]及其他基于 LDA 的方法[13],此外还有学者提出基于 tf-idf 的方法,在不同的控制集中确定主题的稳定性[14]。

在实验研究中,学者们利用主题模型识别阿拉伯语和英语博客的主题,勾勒战争主题的时间线[15],探测与苹果设备相关的微博中新兴与演化主题[13]。还有学者通过调查来自电视、广播、互联网和报纸上关于 SARS、RVF 和 VEE 等突发公共卫生事件的新闻与报道,来验证指示与预警分级模型[16]。一项最近的研究还调查了与飓风桑迪相关的微博每日数量以及这些微博的榜首词汇[17]。

由此可见,当前的实验研究并未充分地探索当突发公共卫生事件发生时微博主题的演化,其调查通常局限于一个国家或一种语言的单一微博平台,很少涉及不同语言或者多个国家的多个微博平台上与突发公共卫生事件相关的微博主题及其演化的比较分析。本文调查了中美两大知名微博平台上关于埃博拉病毒的微博主题时序特征,并总结其各自的主题演化模式与趋势。

2.3 微博可视化分析

由于微博的数量庞大,研究者们开发了若干可视化分析工具、系统或者框架,从而高效直观地分析微博主题或事件及其演化规律,例如 WeiboEvents[18]、有交互界面的可缩放计算框架[19]、epSpread[20]和 SocialHelix[21]。

自组织映射(Self-Organizing Map,简称 SOM)和树图(Treemap)等典型的信息可视化技术也应用于微博探索中。SOM 是一种无监督的人工神经网络方法,能够将高维输入数据显示在低维度空间中,具有保持输入数据拓扑结构的优点[22]。SOM 输出由若干网状的方格构成,通过竞争学习,属性相似的输入数据被映射到相邻的 SOM 结点,而属性差异较大的输入数据则被映射到距离较远的结点。U-matrix[23]是一种常见的 SOM 显示方式,U-matrix 中的每一个元素的值等于对应的 SOM 结点的权向量与直接相邻结点的权向量之间的欧几里德距离之和除以所出现的最大值。将 U-matrix 的值转换成不同的颜色,应用于 SOM 输出的背景颜色,用户就可以直观地观察到 SOM 输出中输入数据映射的位置及其背景颜色,来理解输入数据的分布特点。由于 SOM 的诸多优点,该方法被广泛应用于众多领域[24-26]。SOM 在微博分析中的应用很少,主要集中于电影评论的调查[27-28],很少有研究者运用 SOM 方法分析微博的主题演化,尤其是涉及突发公共卫生事件的微博主题分析则更少。使用树图来分析微博的例子包括盒子里的群组元布局[29],然而,该可视化主要是用于分析社区成员和社区之间的关系,而不涉及微博主题的分析。

可见大部分微博的可视化研究旨在开发可视化工具来分析微博中的主题或事件,这些工具的分析功能包括转推路径、时空特征以及主题或事件的探测与追踪等,很少有学者研究关于突发公共卫生事件的微博主题的演化模式。而这类研究,例如关于埃博拉爆发的微博主题演化模式的研究将会揭示许多有用的规律,协助突发公共卫生管理部门在类似情况下采取合适的决策与措施。

3 突发公共卫生事件的微博主题演化模式和时序趋势的方法设计

本文的数据收集来自于 weibo.com 和 twitter.com,两者分别为中美知名的微博平台。运用 Metaseeker[30]爬

取 2014 年 2 月 1 日至 10 月 31 日两大平台上包含“Ebola“或者“埃博拉“词条的所有微博条目的内容、发布日期、发布者等字段。使用汉语词频统计工具[31]、Word Frequency Counter[32]和 Phrase Frequency Counter[33]分别从中英文微博条目的内容中抽取词和短语,剔除停用词。

按照如下步骤,分别针对 Twitter 和新浪微博的数据构造 SOM 输入矩阵,合并同一天发布的微博内容,生成日期-术语矩阵 M1,如公式(1)所示。

$$M1=\begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ & \cdots & & \\ d_{p1} & d_{p2} & \cdots & d_{pn} \end{pmatrix} \quad (1)$$

其中,p 表示微博数据集的总天数,n 表示术语的总数,$d_{kj}(k=1,2,\cdots,p;\ j=1,2,\cdots,n)$表示在第 k 天术语 j 的出现次数。

每个数据集所生成的 M1 矩阵均采用 SOM 算法进行训练,计算 U-matrix 值并应用于 SOM 输出的背景颜色,按照微博条目的发布日期给 SOM 结点加上标签。SOM 输出中,映射到相同或者邻近 SOM 结点中且 U-matrix 值较小的日期被认为是具有相似的主题。此外,我们尝试给 SOM 结点添加主题内容的标签,然而,本文所收集的微博涉及的术语数量十分庞大,且许多术语是彼此相关的,为了提高效率,本文采用一种新的聚合竞争式 SOM 结点主题标注策略,即为每个 SOM 结点添加相应日期中最突出的主题,其步骤如下。

首先,采用 LDA 模型分别探测两个微博平台上的埃博拉微博的主题。LDA 模型是一种生成式主题模型[34],认为文档包含若干主题,而每个单词以某种概率属于某个主题。对于文档 d,选择主题分布 θd,它遵循 α 的狄利克雷分布,这里 α 为每篇文档的主题分布的狄利克雷先验参数。从 β 的狄利克雷分布中选择 φk,这里 β 是每个主题中单词分布的狄利克雷先验参数。对于文档 d 中每个单词位置 i,从 θd 的多项式分布中选择主题 zi,从 φzi 的多项式分布中选择单词 wi,其中 φzi 是主题 zi 的单词分布。针对每篇文档中的每个单词,重复将单词分配到主题的过程,并循环整个文档集许多次,例如 1000 次。为了推断文档所包含的主题以及每个主题涉及某个单词的概率,通常采用吉布斯采样[35]和期望扩散[36]方法,最终生成若干主题,而每个主题涉及一定数量的单词。假设探测到 k 个主题,每个主题由若干术语构成,记为 termr1,termr2,…,termrs。已知每个 SOM 结点都与一个权向量相关联,第 i 个 SOM 结点的权向量元素用 $wi1, wi2, \cdots, win$ 表示,n 为属性(术语)的数量。为了发现相应日期中最突出的主题,将每个 SOM 结点的权向量元素根据由 LDA 发现的主题进行聚合。假设主题 $j(j=1,2,\cdots,k)$由 termr1,termr2,…,termrs 构成,对于第 i 个 SOM 结点,聚合其权向量元素的计算方法如公式(2)所示。

$$u_{ij}=\sum_{t=r1}^{rs} w_{it} \quad (2)$$

寻找 $u_{i1},u_{i2},\cdots,u_{ik}$ 中的最大,例如,如果 u_{ijv} 是最大值,那么主题 j_v 将用于给第 i 个 SOM 结点添加标签。

4 实验过程与结果分析

4.1 数据描述和预处理

本文有两个数据集,第一个数据集包括 Twitter 平台上共计 271 天的 228 992 条微博;第二个数据集包括新浪微博平台上共计 246 天的 230 274 条微博。为了提高数据处理的效率,分别从两个数据集中抽取前 4000 个高频单词和短语来构造日期-术语矩阵 M1。

4.2 英文埃博拉微博的主题分析

利用 Twitter 数据构造输入矩阵 M1,共有 271 行和 4000 列。为了避免取值范围较大的属性在 SOM 输出中占据主导地位,首先用“var”方法[34]将输入矩阵 M1 中各属性的方差标准化为 1。为了避免“边缘效应”,采用超环面的 SOM 输出[21]。

鉴于相关研究显示,线性初始化和批学习算法的组合所产生的最终量化误差比其他初始化与学习算法组合所产生的最终量化误差要小[24],本文采用线性初始化和批学习算法对输入矩阵 M1 进行训练,将 U-matrix 的值作为 SOM 输出的背景颜色,如图 1 所示。右边的颜色条表示每种颜色的 U-matrix 值。SOM 输出上的日期标签表示该日期发布的 Twitter 内容所映射到的 SOM 结点,例如 2-1 表示 2014 年 2 月 1 日。

根据 SOM 的原理,映射到相同或邻近结点且 U-matrix 值较小的 SOM 结点中的日期发布的微博内容具有相似的术语。由于图 1 采用超环面输出,因此“上边缘”和“下边缘”、“左边缘”和“右边缘”实际上是相连的。图 1 显示,邻近的日期大多映射到邻近的 SOM 结点中,这意味着在一段时间内,例如一个月或者连续几天的 Twitter 内容倾向于具有集中的主题,随着时间的推移,这些焦点主题也会随之改变。

为了探索埃博拉微博的主题特征,采用 LDA 模型来识别这些微博的主题。前期研究表明,50 个主题的困惑度低于 10 个到 40 个主题的困惑度。由于太多的主题会降低分析效率,因此选择 50 个主题运行 LDA 算法,每个主题选取前 20 个概率值较高的术语。狄利克雷先验参数 α 设置为 0.5,β 设置为 0.1,学习过程迭代 1000 次。

如前所述,每个 SOM 结点的权向量分量按照 LDA 识别的主题进行聚合,而每个 SOM 结点用最突出的主题进行标注,即第 i 个 SOM 结点由公式(2)中的 uij 最大值所代表的主题来添加标签,如 48 页图 2 所示。

图 1 和图 2 的对比结果揭示了每个主题及其主导的时间段。例如,3 月 20 日和 21 日最突出的主题为第 19 个主题。总共有 21 个主题主导 Twitter 内容的时间

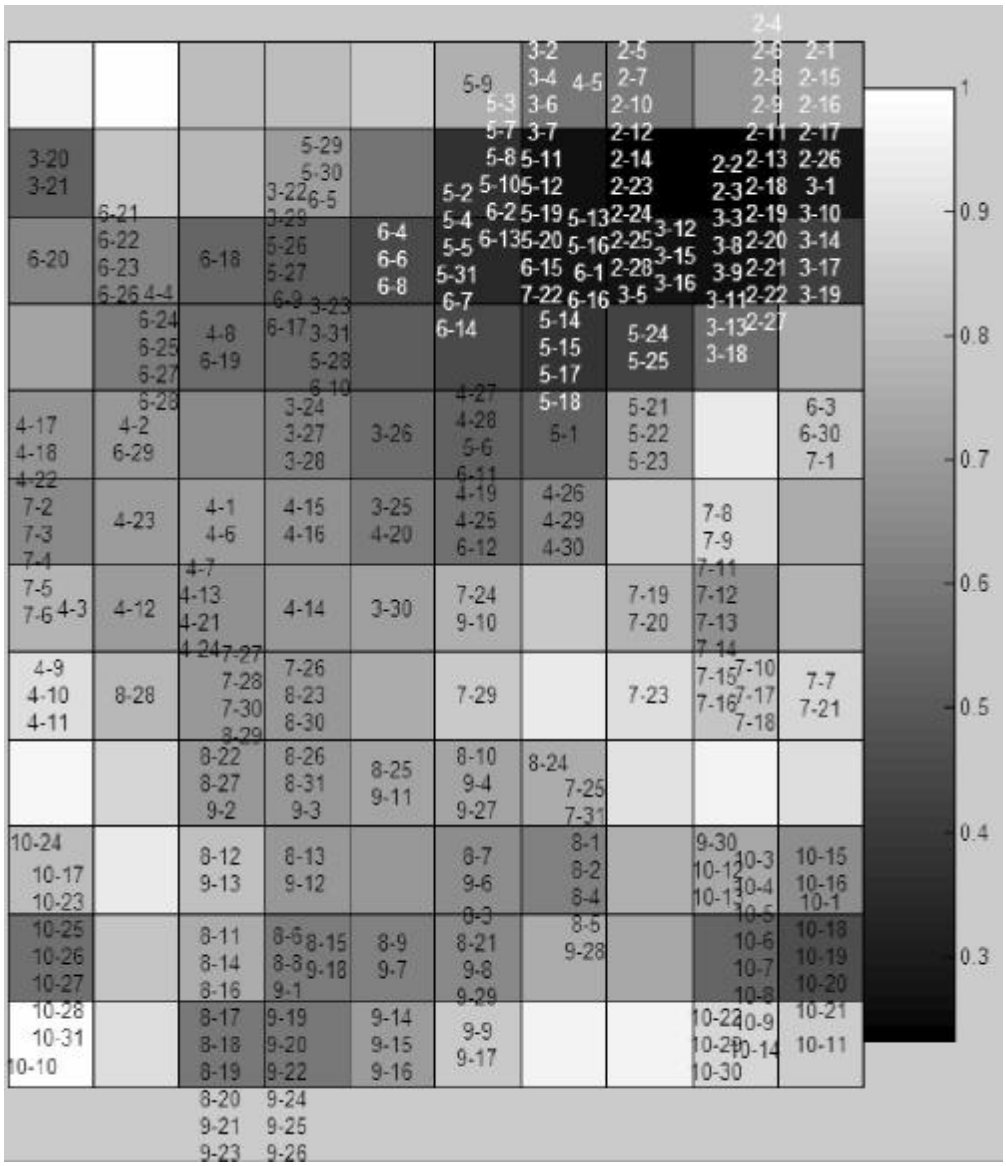

图 1 用日期标注的 SOM 输出

超过一天,详见附录 I(http://u.163.com/dRwhyBIY 提取码: RccEDoKq)。其中,第 19 个主题主导的时间最短,仅有两天;而第 1 个主题主导的时间最长,长达 50 天。我们结合查看对应的微博内容,对所有主题进行了详细分析,将其概要归纳至附录 I,发现这些主题可以归纳为 11 个类别,将每个月每个类别主导的天数归纳至表 1。

表 1 中每个单元格中的数字表示该主题类别在当月主导 Twitter 内容的天数。如果某单元格为空,并不意味着该月没有出现与该主题类别相关的微博,实际上相关微博内容可能仍然存在,只是并没有占据主导地位。

表 1 揭示了埃博拉 Twitter 微博主题的时序发展过程。在西非埃博拉爆发之初,人们对于埃博拉爆发的可能原因感到好奇,许多人将矛头指向移民问题,并预测人口可能下降的后果。几天之后,注意力转向一些非理性的公众行为,同时这也属于事件。关于各种事件的微博持续了大部分时间,并随时间推移逐渐加剧。至第二个月,世界卫生组织(WHO)、美国国立卫生研究院(NIH)及西非国家政府部门意识到采取适当措施,抗击埃博拉病毒的必要性和重要性,例如埃博拉病毒研究等长期措施以及关闭边境和学校等短期措施。随着埃博拉的蔓延与肆虐,统计与状态描述的微博日益增多,例如死亡人数和新感染地区等。同时,ISIS 的涌现和乌克兰停火等其他同时出现的新闻也和埃博拉病毒一同被提及,当然该热点仅持续了很短时间。在调查时间段的中期,人们着重于将西非埃博拉爆发评价为前所未有的最大挑战。在调查的最后阶段,人们对于埃博拉的愤怒和憎恨到达顶峰,第 41 个主题(甚至包含脏话)占据了整个十月份的微博热点,这表示许多人情绪失控,且感到沮丧。

表 1 Twitter 微博中各主题类别在各月的主导时间

	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	总计
埃博拉爆发的可能原因	19	14	3	12	2					50
预测	19	18	4	12	3					56
公众行为	9	5	1	9	3	1				28
事件	9	5	1	9	3	6	18	9		60
组织行为		2	2		1		6	4		15
措施和响应		2	3		1		11	18		35
统计描述		4	10	10	14					38
状态描述		2	15	1	11	4	14	18		65
外部环境			5		2	6				13
评价			3			2				5
公众情感								1	30	31

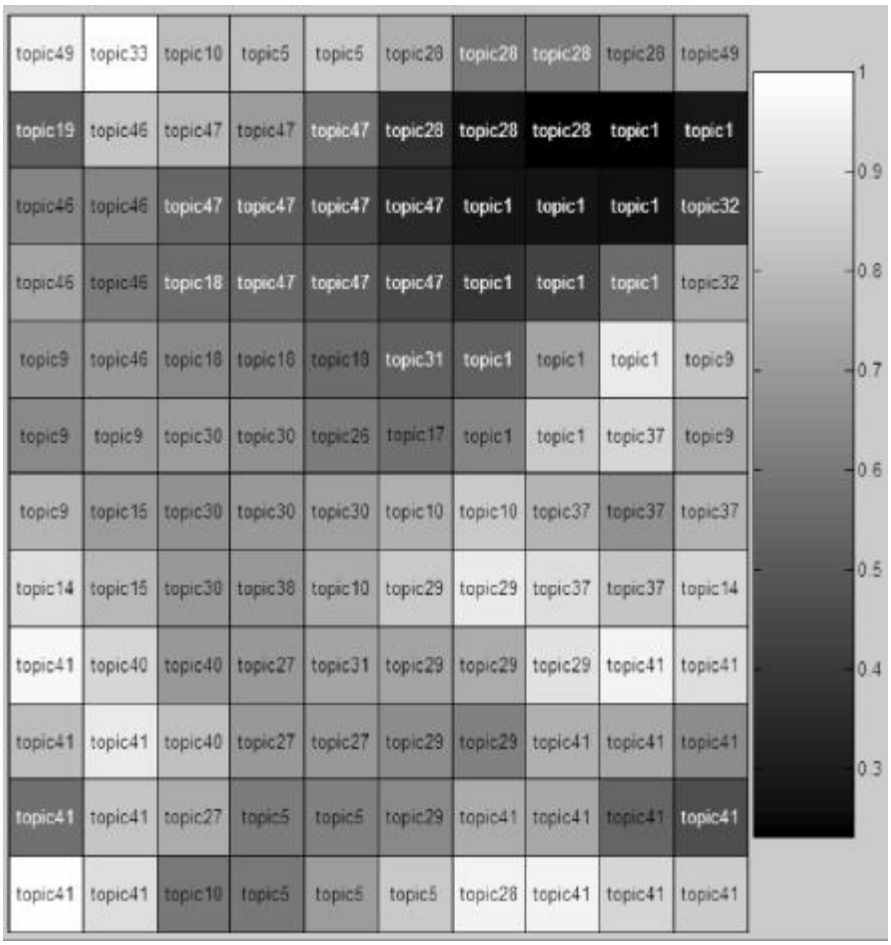

图 2 用最突出的主题标注的 SOM 输出

表 1 所识别的微博主题类别与吴媛[37]提出的甲型 H1N1 流感的报纸新闻主题框架类目与报导对象类目有相似之处,如后者所构建的定义事件、人情趣味、评价分析、疫情及影响、国家防疫政策与权威发、具体防控工作、国际情况等主题类目分别对应于表 1 中的事件、公众情感、预测、状态描述、组织行为、措施和响应、外部环境等主题类别。

表 1 显示,在所有主题中,状态描述和事件是最热门的主题类别,而评价和外部环境则热度最低,这与吴媛[37]发现的《人民日报》中关于甲型 H1N1 流感的疫情及影响、定义事件的篇数最多,评价分析、国际情况的篇数较少相吻合。周婕[38]同样发现《人民日报》中关于甲流与非典的疫情动态的报导最多,国际援助的报导较少,将所调查的时间段分为三个区间,分别为 2014 年 2 月到 4 月、5 月到 7 月以及 8 月到 10 月,第一阶段最热门的主题类别是预测和埃博拉爆发的可能原因;第二阶段最热门的主题类别是统计描述和事件;第三阶段最热门的主题类别是状态描述和公众情感。该研究发现与左莹莹[39]的研究结果有相似之处,后者发现《人民日报》和《文汇报》对 H7N9 的报导在事件上升阶段主要包括疫情通报、疫情分析等主题,这与图 3 中的第一、二阶段的埃博拉爆发的可能原因与统计描述这两个主题类别较为吻合。实际上,第二阶段对应的 2014 年 5~7 月仍然是埃博拉的上升阶段,而 8~10 月属于埃博拉的维持阶段。在维持阶段,《文汇报》的主题主要涉及患者情况、疫情的治疗情况、民众反应以及社会影响等,这与图 3 中第三阶段的状态描述与公众情感这两个主题类别较为吻合。图 3 显示了埃博拉微博的主题演化过程。

关于各主题类别的发展趋势,措施和响应、统计描述以及公众情感这三个主题类别的热度都呈上升趋势;埃博拉爆发的可能原因、公众行为、预测和评价这四个主题的热度则呈下降趋势;事件、状态描述、外部环境和组织行为这四个主题的热度此起彼伏,分别于 2014 年 8 月、9 月、7 月和 8 月达到顶峰。图 4 演示了英文埃博拉微博的主题时序发展。这样的研究发现能够为突发公共卫生事件应急响应部门提供有用的决策依据。在突发公共卫生事件的早期,管理部门应当提供充分的有关传染病起因和合理后果的准确信息。随着疾病的蔓延,管理部门需要及时通报关于措施与响应、统计描述等信息。在后续阶段,管理部门需要平复公众可能高涨的情绪。

4.3 中文埃博拉微博的主题分析

以相同步骤处理中文埃博拉微博。分别用日期和主题标注 SOM 输出,其结果与图 1 和图 2 类似,在此省略。附录 II (http://u.163.com/dRwhyBIY 提取码: RccEDoKq)提供~~了每个主题及其主导日期~~。通过对主题及其概要的详细审查发现,中文微博的主题可以归纳为 12 个类别,如表 2 所示。

表 2 揭示了中文埃博拉微博的主题时序发展过

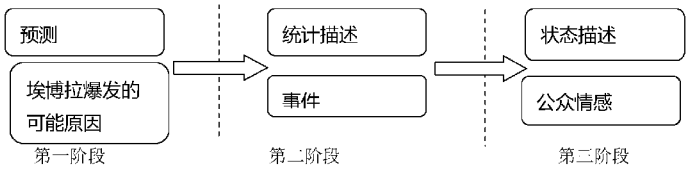

图 3 英文埃博拉微博的主题演化

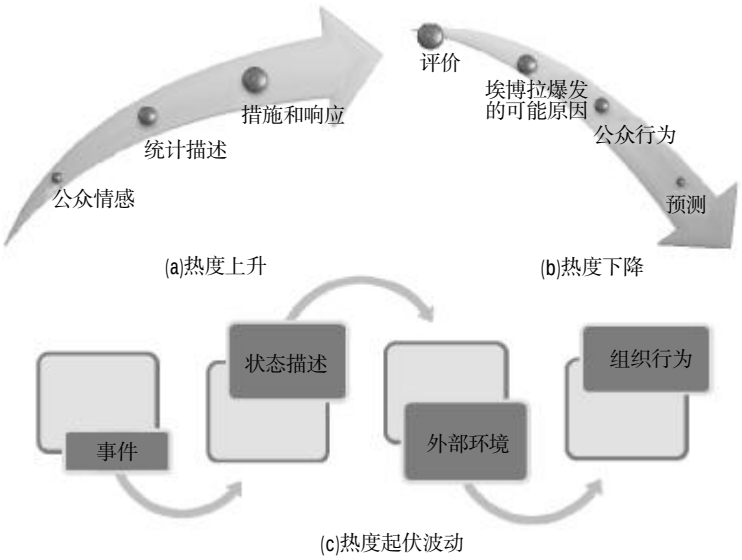

图 4 英文埃博拉微博主题的时序发展

表 2 中文微博中最热门主题类别在各月的主导时间

	2月	3月	4月	5月	6月	7月	8月	9月	10月	合计
埃博拉背景知识	8	25	30	26	29	23	1			142
广告		1			1	4	7			13
措施和响应				3		2	5	10	3	23
组织行为				3		2	7	13	12	37
预测						1	1			2
状态描述						1	7		1	9
统计描述						1	14	2	1	18
事件							4	5	8	17
公众情感							4			4
谣言								2		2
公众行为								4	3	7
外部环境									1	1

注：每个单元格中的数字表示该主题类别作为最热门主题类别的持续天数。

程。自 2014 年西非埃博拉爆发开始到调查中期阶段，国内用户一直在传播埃博拉背景知识。值得注意的是，从第二个月开始，一些代购商利用埃博拉热点优势发布了众多与埃博拉无关的广告。中文微博中措施和响应、组织行为成为最热门主题类别的时间比英文微博晚了近两个月。例如，早在 2014 年 3 月底，许多 Twitter 用户就注意到 NIH 五年内投入了 2800 万美元用以抗击致命埃博拉病毒这一举措上。然而直到 2014 年 5 月中旬，国内用户才将注意力集中于世界卫生组织及其他部门派遣专家组前往扎伊尔这一措施与响应之上。

通过表 1 和表 2 的比较，发现中英文微博有九个共同的主题类别，例如措施和响应、组织行为和预测等。埃博拉背景知识、广告和谣言这三个主题类别仅存在于中文微博中；而埃博拉爆发的可能原因和评价这两个主题则仅存在于英文微博中。

在中文埃博拉微博中，埃博拉背景知识和组织行为是最热门的两个主题类别；外部环境、预测和谣言则是热度最低的三个主题类别。图 5 显示了中文埃博拉微博的主题演化模式，埃博拉背景知识这一主题类别在第一和第二阶段均为最热门的主题。广告、措施和响应、组织行为的突出性次之，因此在图 5 中用虚线框显示。组织行为、措施和响应在第三阶段为最热门的主题类别，该研究发现与左莹莹[39]和左馨[40]的研究结果之间存在相似之处。左莹莹发现《人民日报》和《文汇报》在事件的上升与维持阶段的报导都涉及政府措施，这与图 5 中的第二、三阶段的措施和响应、组织行为这两个主题类别较为吻合。《文汇报》在上升阶段主要是介绍疫情与专家的权威观点，这与图 5 中第一、二阶段的埃博拉背景知识较为吻合。左馨发现《S 商报》、《晶报》等六家报纸在登革热的早期与中期都较多刊登了关于风险提示、知识普及与政府举措的文章，分别对应于图 5 中的埃博拉背景知识、措施和响应以及组织行为。

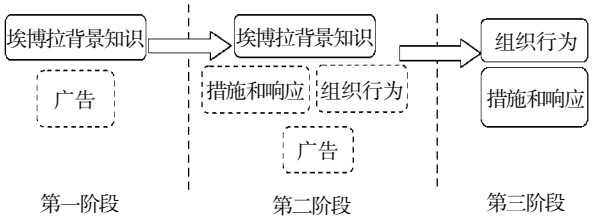

图 5 中文埃博拉微博的主题演化模式

图 5 表明国内用户倾向于将微博平台作为一个普及传染病医学知识，向公众通告组织行为的渠道，许多内容都与抗击致命疾病的措施相关。他们倾向于接受现状并采取务实的态度，例如了解埃博拉病毒及其抗击方式，而 Twitter 用户则热衷于探究埃博拉的起因与结果。

值得注意的是，与中文微博相比，英文微博中埃博拉背景知识这一主题类别的热度并不明显，措施和响应这一主题类别也不够受重视。正如相关研究[4]所发现的，公众实际上高度关注埃博拉的症状(与埃博拉背景知识相关)、安全旅行以及埃博拉的防护(与措施和响应相关)，研究者建议告知公众关于埃博拉的知识、旅行中被感染的风险以及预防埃博拉的措施。

在各主题类别的发展趋势方面，组织行为、事件和广告这三个主题类别的热度始终呈上升趋势；埃博拉背景知识和公众行为的热度则呈下降趋势；措施和响应、现状描述、统计与描述的热度则起伏波动；预测、公众情绪、谣言和外部环境的热度转瞬即逝，如昙花一现。下页图 6 显示中文埃博拉微博主题的时序发展趋势。

在预测、现状描述、统计与描述、事件、公众行为和

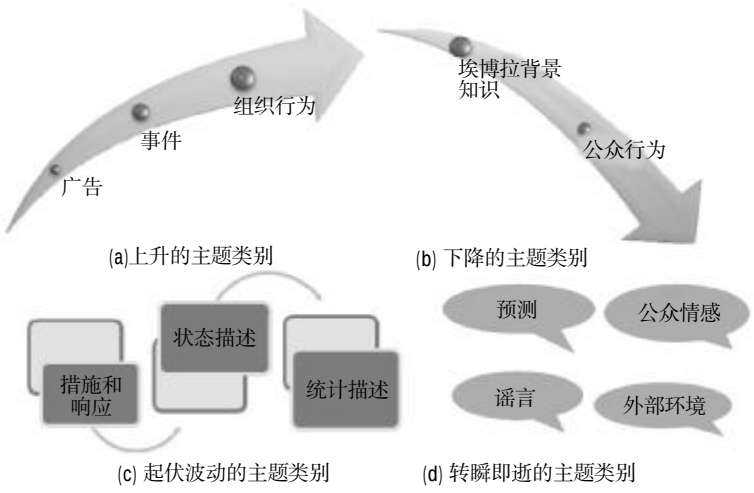

图 6 中文埃博拉微博主题的时序发展趋势

公众情感这几个主题类别上，英文微博明显强于中文微博。其中，公众情感在中文微博中大多表现为对医务工作者的感激和尊敬之情，但是在英文微博中该主题类别则大部分表现为对埃博拉的愤怒和憎恨之情。

令人遗憾的是，谣言唯独在中文微博中作为热点持续时间长达两天，主要内容为鲑鱼是否被埃博拉感染以及能否传播该疾病，许多国内用户都相信此内容，直到两天后谣言被驳斥。在相同时间段内全面检索英文微博仅发现了一条微博，实际是关于中国的这一现象："You can get Ebola from atlantic salmon? That what many in China believe. Market op for NZ king salmon?http://www.stuff.co.nz/business/10490846/Ebola -a -boost -for -NZ -salmon …"(你会从大西洋鲑鱼中感染埃博拉？这是许多中国人所相信的。新西兰帝王鲑鱼的营销操作?)为什么这类谣言仅出现在中文微博中？这种反常现象值得深思，我们也有必要提醒国内用户在对待可疑且未经验证的信息时要保持冷静，不要转发或评论。

在外部环境方面，Twitter 用户主要是谈论 ISIS 的涌现和乌克兰停火，它们时常同时出现在新闻微博中。新浪微博用户则谈论生物医学股票的价格，似乎相信这些公司与埃博拉相关，虽然并没有相关线索表明这一点。

总之，国内用户在埃博拉某些方面的焦点，如状态描述、统计与描述、事件和公众行为等相较于 Twitter 用户滞后了大约 4 到 7 个月，且强度更弱。国内用户更倾向于接受现状，而不是试图探究传染病的原因及其后果。相反，他们非常务实，将微博平台作为普及传染病背景知识的渠道，并且密切关注抗击致命疾病的有效措施，公众情感大多是对奋战在疫区的医疗工作者表达感激和尊敬之情。一些国内用户利用埃博拉病毒这一社会热点发布一些与之不相关的广告，许多人容易受到谣言的影响，因此，国内公共卫生管理部门应当在突发公共卫生事件爆发时及时识别并驳斥谣言和误导性消息，微博平台还需要严控某些广告，将其联系到不相关的热点。

与之相对比，Twitter 用户比国内用户更关注埃博拉爆发的动态，并长期密切关注埃博拉病毒的多个方面，如事件和状态描述；他们热衷于讨论埃博拉的起因和后果，并定期评价疫情爆发的严重程度；在英文微博中没有发现广告或谣言成为热点；埃博拉背景知识的热度也明显弱于国内；公众情感主要是对埃博拉病毒的愤怒和憎恨之情，并在调查接近尾声时达到顶峰。鉴于此，美国公共卫生管理部门应当充分向公众普及有关传染病的背景知识，从而更好地满足公众希望了解疫情起因和后果的需求，并在必要的时候安抚公众情绪。

5 结论与展望

微博的主题可视化时序分析能够揭示突发公共卫生事件时微博的主题演化模式，其研究发现有助于突发事件响应部门更好地了解重大公共卫生事件的发展轨迹以及公众在各阶段的关注点，从而在类似事件中采取有效的措施来抗击传染性疾病。本文演示了如何利用一种有效的可视化方法——自组织映射(SOM)与潜在狄利克雷分配(LDA)模型相结合，分析传染性疾病的微博主题时序分布，并概述当疾病爆发时相关微博的主题演化模式。我们调查了两大知名微博平台上 2014 年 2 月至 10 月共计 45 万多条关于西非埃博拉爆发的中英文微博，提出一种新的聚集竞争式 SOM 结点标注方法，即每个 SOM 结点由权向量元素之和最大值所对应的最突出的主题来标注。我们探索了各阶段的中英文微博的主题，并将其归纳为 14 个类别，即埃博拉爆发的可能原因、预测、公众行为、事件、组织行为、措施和响应、统计描述、状态描述、外部环境、评价、公众情感、埃博拉背景知识、广告与谣言。其中公众行为、事件等 9 个主题类别为两个平台所共有的主题类别，而埃博拉爆发的可能原因与评价为 Twitter 平台所独有的热点主题类别，广告与谣言则是新浪微博所独有的热点主题类别。

研究发现新浪微博与 Twitter 平台具有不同的主题演化模式和时序变化趋势。新浪微博用户主要将微博平台作为普及传染病背景知识的渠道，并且密切关注抗击致命疾病的有效措施，而 Twitter 用户则更关注埃博拉爆发的动态，并长期密切关注埃博拉病毒的多个方面，如事件和状态描述。Twitter 平台的主题类别可划分为三种不同的时序变化趋势，即上升、下降与波动

的主题类别，而新浪微博则增加了转瞬即逝的主题类别。两者的共同之处在于,公众行为在两个平台上均为热度下降的主题类别，而状态描述均为波动的主题类别。其研究发现有助于理解重大公共卫生事件的发展轨迹、公众和相关利益者对传染性疾病的关注点以及中英文微博平台的主题差异。该研究发现为突发公共卫生事件应急响应部门在处理类似事件，如目前寨卡病毒蔓延提供了一定的借鉴，例如利用微博来传播传染病的背景知识,消除谣言与无关广告,安抚公众的不安情绪,及时通报响应措施等。本文所构建的研究方法也可应用于与突发事件相关的其他社交媒体分析。后续我们将以其他突发事件为例,探究这 14 个主题类别是否适用于相关微博，以及是否能发现类似的时序发展模式用于微博的主题预测。

参考文献

[1] World Health Organization. Ebola Situation Report - 15 July 2015 [OL]. [2015-07-18]. http://apps.who.int/ebola/current-situation/ebola-situation-report-15-july-2015.

[2] Towers S, Afzal S, Bernal G, et al. Mass media and the contagion of fear: the case of Ebola in America[J].Plos One, 2015,10(6):1-13.

[3] 世卫：已有 34 国报告发现寨卡病毒原地传播病例 [OL]. [2016-02-17]. http://world.huanqiu.com/hot/2016-02/8554557.html.

[4] Yun H. Analysis of similarity of twitter topic categories among regions [J]. Journal of Information and Communication Convergence Engineering, 2012,10(1):27-32.

[5] Lazard A J, Scheinfeld E, Bernhardt J M, et al. Detecting Themes of Public Concern: A Text Mining Analysis of the Centers for Disease Control and Prevention's Ebola Live Twitter Chat [OL]. [2015-07-18]. http://www.sciencedirect.com/science/article/pii/S0196655315006148.

[6] Kim C, Kang B S, Choi H J, et al. Nationwide online social networking for cardiovascular care in Korea using Facebook [J]. Journal of the American Medical Informatics Association, 2014, 21(1):17-22.

[7] Qu Y, Huang C, Zhang P, 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, Hangzhou, China, March 19-23, 2011. New York, NY: ACM, 2011: 25-34.

[8] Schaal M, O'Donovan J, Smyth B. An analysis of topical proximity in the twitter social graph [J]. Lecture Notes in Computer Science, 2012,7710: 232-245.

[9] Anantharam P, Thirunarayan K, Sheth A. Topical anomaly detection from twitter stream [C]. Proceedings of the 3rd Annual ACM Web Science Conference, Evanston, IL, USA, June 22-24. New York, NY: ACM, 2012: 11-14.

[10] Yang M C, Rim H C. Identifying interesting Twitter contents using topical analysis [J]. Expert Systems with Applications, 2014, 41(9): 4330-4336.

[11] Wang X, McCallum A. Topics over time: a nonmarkov continuous time model of topical trends [C]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, Pennsylvania, USA, August 20-23, 2006. New York: ACM, 2006: 424-433.

[12] Vosecky J, Jiang D, Leung K W, et al. Integrating social and auxiliary semantics for multifaceted topic modeling in twitter [J]. ACM Transactions on Internet Technology, 2014, 14 (4): 1-24.

[13] Dey L, Khurdiya A, Mahajan D. Topical evolution and regional affinity of tweets [C]. Proceedings of 2013 International Symposium on Computational and Business Intelligence, New Delhi, India, August 24-26. Los Alamitos, CA: IEEE, 2013: 297-300.

[14] Lai V, Rand W. Does love change on twitter? The dynamics of topical conversations in microblogging [C]. Proceedings of 2013 ASE/IEEE International Conference on Social Computing, Washington, DC, USA, September 8-14, 2013: 81-86.

[15] Mark G, Bagdouri M, Palen L, et al. Blogs as a collective war diary [C]. Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, Seattle, Washington, USA, February 11-15. New York: NY: ACM, 2012: 37-46.

[16] Wilson J M, Polyak M G, Blake J W, et al. A heuristic indication and warning staging model for detection and assessment of biological events [J]. Journal of the American Medical Informatics Association, 2008, 15(2): 158-171.

[17] Wang H, Hovy E, Dredze M. The hurricane sandy twitter corpus [C]. Proceeding of AAAI Workshops at the 29th AAAI Conference on Artificial Intelligence, Austin Texas, USA, January 25-30, 2015: 20-24.

[18] Ren D, Zhang X, Wang Z H, et al. Weibo events: a crowd sourcing weibo visual analytic system [C]. Proceedings of IEEE Pacific Visualization Symposium, Yokohama, Kanagawa, Japan, March 4-7, 2014: 330-334.

[19] Cao G, Wang S, Hwang M, et al. A scalable framework for spatiotemporal analysis of location-based social media data [J]. Computers, Environment and Urban Systems, 2015, 51: 70-82.

[20] Walker R, Cenydd L, Pop S, et al. Storyboarding for visual analytics [J]. Information Visualization, 2015, 14(1): 27-50.

[21] Cao N, Lu L, Lin Y R, et al. SocialHelix: visual analysis of sentiment divergence in social media [J]. Journal of Visualization, 2015, 18(2): 221-235.

[22] Kohonen T. Self-Organizing Maps[M]. (3rd ed.) Berlin: Springer, 2001.

[23] Ultsch A. Self-organizing neural networks for visualization and classification [C]. Proceedings of Conference of Society for Information and Classification, Dortmund, Germany, 1992.

[24] An L, Yu C, Li G. Visual topical analysis of Chinese and American library and information science research institutions [J]. Journal of Informetrics, 2014, 8(1): 217-233.

[25] An L,Zhang J,Yu C.The visual subject analysis of library and information science journals with self-organizing map [J]. Knowledge Organization,2011,38(4):299-320.

[26] Zhang J, An L, Tang T,et al. Visual health subject directory

analysis based on users' traversal activities [J]. Journal of the American Society for Information Science and Technology, 2009, 60(10): 1977-1994.

[27] Hao M C, Rohrdantz C, Janetzko H, et al. Visual sentiment analysis of customer feedback streams using geo-temporal term associations [J].Information Visualization, 2013, 12 (3-4):273-290.

[28] Claster W B, Dinh Q H, Shanmuganathan S. Unsupervised artificial neural nets for modeling movie sentiment [C] .Proceedings of the 2nd International Conference on Computational Intelligence Communication Systems and Networks, Liverpool, United Kingdom, July 28-30. IEEE, 2010: 349-354.

[29] Chaturvedi S, Dunne C, Ashktorab Z. et al. Group-in-a-box meta-layouts for topological clusters and attribute-based groups: space-efficient visualizations of network communities and their ties [J]. Computer Graphics Forum, 2014, 33(8): 52-68.

[30] Metaseeker [OL]. [2014-11-10]. http://www.gooseeker.com/cn/node/product/front.

[31] Chinese Word Frequency Statistical Tool [OL]. [2015-01-15]. http://nlp.blcu.edu.cn/downloads/download-tools/26.html#ecms.

[32] Word Frequency Counter [OL]. [2015-01-16]. http://www.writewords.org.uk/word_count.asp.

[33] Phrase Frequency Counter [OL]. [2015-01-17]. http://www.writewords.org.uk/phrase_count.asp.

[34] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3 (4-5): 993-1022.

[35] Griffiths T L , Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences, 2004, 101 (Suppl. 1): 5228-5235.

[36] Minka T , Lafferty J. Expectation-propagation for the generative aspect model [C]. Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, San Francisco, CA: Morgan Kaufmann, 2002.

[37] 吴 媛.突发公共卫生事件的报纸新闻框架研究[D].长沙:湖南大学,2010.

[38] 周 婕.《人民日报》"非典"与"甲流"报道比较研究[D].长沙:湖南大学,2014.

[39] 左莹莹.主流媒体突发公共事件报道的比较分析——以《人民日报》和香港《文汇报》对 H7N9 的报道为例[J].东南传播,2015(12):61-64.

[40] 左 馨. 健康传播视域下 S 市疾控中心 2014 年媒体应用与风险评估研究[D].广州:暨南大学, 2015.

[作者简介]安 璐,女,1979 年生,武汉大学信息管理学院副教授,硕士生导师。

杜廷尧,男,1991 年生,武汉大学信息管理学院硕士研究生。

余传明,男,1978 年生,中南财经政法大学副教授,硕士生导师。

周利琴,女,1992 年生,武汉大学信息管理学院博士研究生。

李 纲,男,1966 年生,武汉大学信息管理学院教授,博士生导师,教育部长江学者特聘教授(本文通讯作者)。

收稿日期:2016-05-03