

基于在线媒体的新冠疫情社会舆情 多视角分析*

黄晓辉^{1,2} 卢焱¹ 唐锡晋^{1,2}

(1. 中国科学院数学与系统科学研究院, 北京 100190; 2. 中国科学院大学, 北京 100049)

摘要 2020年初爆发的新冠肺炎疫情是近年来最大公共卫生事件, 挖掘期间关于医疗物资社会舆情的性质和演变趋势, 有助于我们了解事件演化和社会应对机制. 文章从话题演化视角和事件结构视角展开研究. 文章首先基于“口罩”相关的新闻语料库, 利用 LDA 主题模型挖掘疫情期间不同时间窗口下的新闻热点话题内容, 分析主要话题社会疫情、社会防疫、口罩生产与监管和口罩进出口的演化关系和演化路径. 针对于口罩进出口中的海关查获事件, 文章依据事件中关键词构建共现关键词网络, 获悉查获事件中的主要四种类型, 并通过构建海关 - 关键词二模网得到海关与查获事件关键词的关系, 揭示查获事件类型在对应海关的分布.

关键词 LDA, 话题演化, 社会网络分析.

MR(2000) 主题分类号 91D30, 05C90, 68T50

Multi-Perspective Analysis of Public Opinion Related to COVID-19 Based on Online Media

HUANG Xiaohui^{1,2} LU Yan¹ TANG Xijin^{1,2}

(1. *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190;*

2. University of Chinese Academy of Sciences, Beijing 100190)

Abstract COVID-19 which outbreaked in early 2020, is the most significant public health emergency in recent years. Excavating the characteristic and evolution tendency of social public opinion about medical supplies during the epidemic period will help us understand the evolution of this emergency and the social response mechanism. This study conducts research from the perspective of topic evolution and event structure. Based on the “mask” related news corpus, this study uses the LDA topic

* 国家重点研发计划基金项目 (2016YFB1000902), 国家自然科学基金 (71731002, 71971190) 资助课题.

收稿日期: 2021-01-25, 收到修改稿日期: 2021-05-28.

编委: 李振鹏.

model to dig some hot topics from the news under different periods during the epidemic. Therefore, the relationship and evolution path of some main topics, which are about the social epidemic situation, social epidemic prevention, mask production and supervision, and the import and export of mask, are analyzed. In view of the customs seizures in the import and export of masks, this study builds a co-occurrence keywords network based on the keywords in the events, learns the main four types of seizure events, and obtains the relationship between the customs and the seizure event keywords through the construction of the bipartite networks of custom-keywords.

Keywords LDA, topic evolution, social network analysis.

1 引言

随着 2020 年初新冠肺炎疫情在我国爆发, 对于该重大突发公共卫生事件的紧急应对是对政府的一大考验. 从最初的武汉封城, 到家家户户“留守”家中携手抗疫, 再到国内疫情稳定后的复工复产, “新冠肺炎疫情”这个词是过去一年公众关注的最热点话题. 而网络媒体作为如今最重要的消息渠道, 在疫情期间起着相当重要的作用. 在疫情期间, 国家采取什么样的措施控制疫情、公众生活所受到的影响和渴望解决的诉求, 均在新闻媒体有所体现. 在抗疫过程中, 医疗物资无疑是最关键的一点. 为探究疫情期间公众对于医疗物资的热点关注内容, 揭示关注对象的演化和事件性质, 助力于未来类似公共卫生事件的应对措施, 本文拟通过以“口罩”为关键词的新闻语料库展开分析. 如何利用高效的方法从新闻媒体中挖掘得到有用的文本信息, 并分析文本特征的演变及结构, 是本文研究的重点内容.

在文本话题内容的视角, Blei 等^[1]提出的 LDA 话题模型, 能高效率地挖掘文本数据中隐含的话题分布. 该方法结合词袋模型, 以非监督机器学习的方式从大量文本中识别文档中的话题及话题内容, 这能高效率地处理无标注的新闻媒体信息. 长期以来有大量研究人员利用该方法研究文本中话题性质及其演化. Xu 等^[2]通过 LDA 话题模型抽取网络新闻文本中话题内容, 与基于 k-means 文本挖掘的方法比较后发现 LDA 话题识别的结果更准确, 并且能上时间流上跟踪新闻话题演化的趋势. 曹丽娜和唐锡晋^[3]利用 LDA 话题模型和动态话题模型 (DTM) 对天涯论坛上的帖子进行分析, 探究各类热点话题的热度在时间上的变化. 张亚茹和唐锡晋^[4]结合 LDA 话题模型和 Bert 向量抽取天涯杂谈“红会贴”中热点事件的发展故事线, 可视化该热点事件的演化过程, 该故事线披露出了红会的公众信任危机. Yan 和 Tang^[5]利用 LDA 话题模型和 Kleinberg 模型从“安全帽事件”中抽取和筛选出事件的突发话题, 利用 Jaccard 系数计算时序上话题间的演化关系, 构建出话题演化的桑基图, 得到“安全帽事件”话题的演化路径. 李振鹏和黄帅^[6]基于天涯杂谈语料库, 利用 LDA 主题模型对帖子分类, 并通过分析热帖的主题比率, 探究天涯杂谈版块的网络舆情方向和网民态度. 可见, 利用 LDA 主题模型捕捉文档的主题演化已经成为一种有效且热门的方法.

在事件关系视角, 社会网络分析的方法可有效用于诠释文本内容之间的关系. 郝强和唐锡晋^[7]利用社会网络分析方法构建第一届“系统科学大会”投稿文章的关键词和作者关系相关网络, 分析系统科学领域当前研究热点、研究人员间合作关系等信息. 钱芸芸等^[8]提出融合主题相似度权重的主题社区发现模型, 结合文本信息抽取和社区模块化方法, 挖掘主题的社区化结构. L'Huillier 和 Mrvar^[9]结合 LDA 文本挖掘和社会网络分析方法, 挖掘论坛用户的主题内容并构建用户间关系网络, 分析论坛用户中极端组织的行为及影响. 社会网络

分析方法也可以应用到新闻文本中. Batagelj 等^[10]曾通过构建 9·11 恐怖袭击新闻数据集中新闻话题间的关系网络, 识别 9·11 恐袭新闻中的主要话题和次要话题, 分析关于该新闻公众的热点关注内容. Moessner 等^[11]结合 LDA 话题模型和社会网络分析方法, 分析在线饮食论坛中的帖子数据, 探究拥有不良饮食习惯用户之间的特点, 借此寻找改善公众饮食习惯的方法. 结合 LDA 话题模型和社会网络分析方法, 可以有效地分析话题动态趋势以及话题内容之间的联系.

对于新冠疫情社交媒体信息, 许多研究也在利用文本挖掘和社交网络的方法分析文本的话题内容. 李沿江等^[12]利用 LDA 话题模型挖掘新冠疫情期间微博话题“中医新冠肺炎”的关注热点, 肯定社会公众对疫情中中医医治能力的认可程度, 同时也发现未来中医传承的困境和公众对中医药的信任危机. 杨建梁等^[13]基于人民网《领导留言板》的数据, 通过 LDA 话题模型挖掘民众诉求的主要内容, 分析各类话题的时空演变特征与防疫政策的关系, 探究疫情期间公众的主要诉求以及如何制定相应的防疫政策. 王袁欣和刘德寰^[14]采用 LDA 主题模型对《纽约时报》和《卫报》的新冠肺炎疫情新闻报道进行了横向比较, 揭示了两个媒体在媒体框架选择上的区别, 并探讨大众媒体是如何建构公众对于流行病的认知. Mutanga 等^[15]使用 LDA 主题模型挖掘 Twitter 关于 COVID 话题中的潜在主题, 探究公众对于疫情相关主题, 如每日统计数据、疫苗阴谋论等的态度和看法, 分析政府政策的影响和人们的反应. Ahmed 和 Abayomi^[16]结合社交网络分析和文本分析的方法, 研究 twitter 中参与 COVID-19 疫情阴谋论话题的用户分布和内容, 探究阴谋论传播的规律特点和应采取的隔断手段. 对于新冠疫情期间各类文本的主题内容, 国内外大量研究表明 LDA 主题模型和社交网络分析可有效用于探讨文本主题的演化和文本结构.

从多视角对事件进行分析可有效展示事件的发展脉络和事件本质. 为探究新冠疫情期间人们对于“口罩”话题的关注热点及演化, 本文以疫情期间口罩相关新闻语料库为例, 利用 LDA 话题模型挖掘相应时间窗口上新闻文本的热点话题, 并根据话题演化关系构建话题演化图. 为进一步探究话题演化中出现的海关查获事件话题性质, 本文利用社交网络分析的方法, 构建事件关键词与事件主体之间的二模网络, 进而对查获事件性质做进一步展开分析. 文章研究框架如图 1 所示.

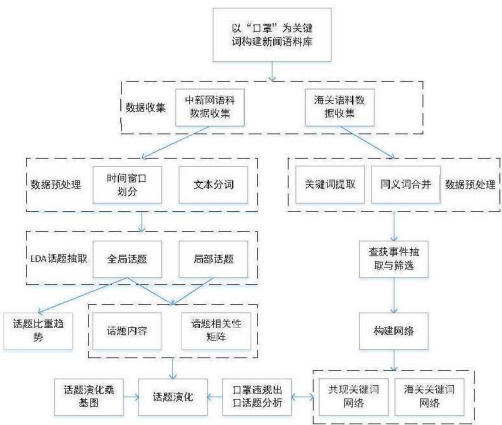


图 1 研究框架

(Figure 1 Research framework)

2 数据预处理

本文基于中新网新闻语料库及中国海关总署新闻语料库展开分析. 中新网新闻报道具有时效性, 对中新网新闻语料库做话题分析可准确获悉各阶段关于新冠肺炎疫情的新闻内容, 有助于挖掘时间上话题演化情况. 而中国海关总署官网中关于口罩的新闻报道, 能针对性的反应口罩进出口的真实情况, 有助于我们揭示口罩进出口事件间的关系. 采用以上两种不同的数据源, 能分别从时间和结构上探索新冠肺炎疫情话题演化和口罩进出口事件的性质, 能真实反映显示情况.

2.1 中新网数据收集与预处理

由中国新闻社主办的中国新闻网¹, 是亚洲上网最早的中文媒体之一, 并且是中央级重点、权威的网络媒体, 其新闻内容具有很高的可信度和参考价值. 为准确抓取以“口罩”为重点讨论对象的新闻, 本文以“口罩”为关键词, 在中国新闻网上搜索新闻标题中包含该关键词的所有新闻, 爬取结构为〈标题, 发布时间, 正文〉的文本数据. 数据样本时间从 2020 年 1 月 1 日至 12 月 31 日, 共获取 3012 条原始数据, 去除缺失值后共余 2623 条样本数据.

数据集中每日新闻数如图 2 所示. 新闻数量为先升后降的趋势, 1 月下旬至 3 月初为新闻发布的高峰期, 新闻发布比较集中, 每日新闻数量在 25 条以上. 随着国内疫情的稳定控制, 有关口罩的每日新闻数逐渐下降, 至后续 8 到 12 月每日新闻数约为 2 条. 因此我们依据实际新闻数量, 构建 7 个合适的时间窗口, 每个时间窗口对应的新闻数量如表 1 所示. 按照时间窗口构建 7 个子数据集. 除 2-4 月新闻数较多外, 其他时间窗口新闻数量较为接近, 保证了后续使用 LDA 模型时文本的一致性.

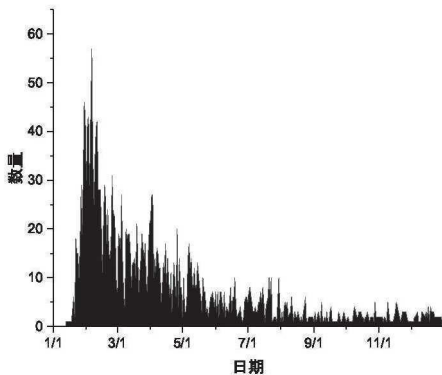


图 2 中新网每日新闻分布
(Figure 2 Number of China News per day)

表 1 各时间窗口新闻数
(Table 1 Number of news in each time window)

时间窗口	1 月	2 月	3 月	4 月	5 月	6-7 月	8-12 月
新闻数量	256	807	448	364	216	263	283

1. <https://www.chinanews.com/>.

2.2 海关数据收集与预处理

在后续章节的中新网话题演化分析中我们发现疫情前中期有大量围绕口罩海关进出口的新闻,关于海关查获违规进出口的事件也持续了较长视角.为能准确获取对应新闻中的事件内容,分析事件的本质,本文从中国海关总署官网²中,搜索并爬取新闻标题包含“口罩”的所有新闻,爬取结构为〈标题,发布时间,正文〉的文本数据.数据样本时间从2020年1月1日至12月31日,共获取564条数据,去除重复或缺失值数据后共有506条有效数据.

为获取与口罩进出口直接相关的新闻,我们根据数据的文本内容,采用人工的方式将数据分成查获、进出口和其他情况三类,并根据查获和进出口新闻提取相应的时间内容.查获指的是关于口罩进出口违法的新闻,正常进出口指的是口罩出口援助他国或受到海外捐赠的情况.最终共获得查获相关新闻51条,进出口事件新闻96条.新闻在时间上的分布如图3所示.由图3中可知此类新闻出现时间为2月至6月期间,其中进出口事件主要在前半部分,而查获事件分布在后半部分,这与前面话题演化情况相一致.

为进一步抽取查获和进出口新闻中的相关事件,本文同样采用人工处理的方式,从新闻中抽取〈对象,关键词〉的二元信息.以查获事件新闻为例,从对应新闻中抽取结构为〈杭州海关;伪报、瞒报、申报不符和逃避商品检验〉的事件信息,其中杭州海关为查获该违法事件的对象,相应的违规关键词为伪报、瞒报等.此外,考虑到不用报道对于同一类术语的面描述可能会存在区别,将意思相近的关键词合并,如关键词“申报不实”与“申报不符”,我们将统一为“申报不实”.

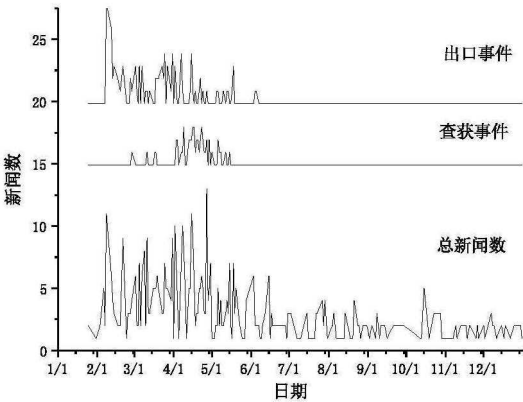


图 3 海关新闻分布
(Figure 3 Number of custom news per day)

3 基于 LDA 的中新网话题分析

为探究过去 12 个月里人们对于医疗物资的关注热点及其演化关系,本节利用 LDA 话题模型挖掘相应新闻语料库中不同时间窗口下的话题,并构建演化关系图.

2. <http://www.customs.gov.cn/>.

3.1 LDA 模型挖掘

LDA 话题模型^[1]是一种基于贝叶斯统计思想的文本挖掘模型, 可用来推测文档的话题分布. 它认为文档集中每篇文档的话题服从一定的概率分布, 话题中的词也服从某个概率分布, 并且词与词之间没有先后顺序的关系. 通过分析文档集的内容反推其话题分布, 并可根据结果进行话题聚类或文本分类. 文档中不同话题数的设定会导致不一样的话题分布结果, 本文对 7 个局部时间窗口和全局时间窗口做一致性检验, 根据最高一致性选择最合适的话题数, 一致性系数计算根据 Newman 等^[17-19]所提出的互信息计算方法. 对于每个窗口, 分别计算各个预设话题数情况下该窗口文本的一致性系数, 根据一致性系数变化曲线, 一致性系数最大的点对应的话题数则为该窗口的最佳主题数. 图 4 为最高一致性及其对应的话题数目, 其中全局话题数为 4 个, 局部时间窗口话题数为 4-6 个.

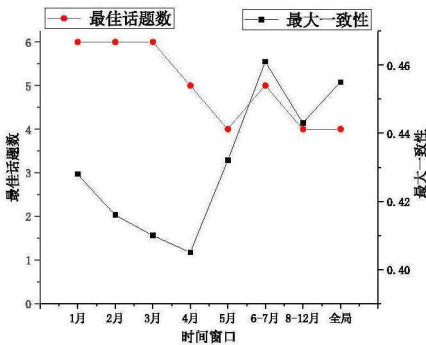


图 4 各时间窗口下最大一致性和最佳话题数

(Figure 4 Maximum consistency and optimal number of topics for each time window)

视所有数据样本为全局数据集, 根据所划分的 7 个时间窗口分别构建局部数据集, 利用 Python 的 Jieba 分词对各个文档做分词处理, 再根据各个时间窗口对应的话题数, 调用 gensim 包中的 LDA 话题模型获取各数据集的话题, 并对每个话题取前 50 个词代表该话题. 部分话题及其内容如表 2 所示.

表 2 部分 LDA 话题结果

(Table 2 Part of the LDA topic results)

话题	话题词	类型
生产与监管	价格, 生产, 市场, 企业, 监管局, 物资, 产能, 哄抬	全局话题
口罩进出口	物资, 海关, 捐赠, 通关, 假冒, 进口, 供应, 中国	全局话题
社会疫情	车站, 乘客, 旅客, 消毒, 体温, 地铁, 传播, 武汉	全局话题
社会防控	佩戴, 冠状病毒, 感染, 防护, 确诊, 医院, 预防, N95	全局话题
口罩价格	价格, 元, 涨价, 监管局, 药店, 药品, 哄抬, 违法行为	局部话题
国外疫情	意大利, 美国, 解封, 场所, 隔离, 法令, 民众, 聚集	局部话题
高考防疫	考生, 考场, 考试, 公共场所, 考前, 居民, 健康状况	局部话题
海关查获	医用, 犯罪, 物资, 嫌疑人, 假冒伪劣, 警方, 依法	局部话题

根据一致性检验的结果, 对整体数据集选取全局话题 4 个, 根据 LDA 结果分别为命名

“生产与监管”、“口罩进出口”、“社会疫情”和“社会防控”4个话题.因而可知在整体语料库中,新闻的话题要是跟这4个话题相关.全局话题分析如下

- 1) “社会疫情”:主要对应的新闻为社会中与口罩有关的疫情情况,如关于口罩佩戴情况,话题词主要为人们日常生活的相关词汇,如车站、乘客等.
- 2) “社会防疫”:主要对应于口罩相关的防疫措施,如出入相关场所佩戴口罩的要求、违反佩戴口罩规定等新闻事件,话题词主要为佩戴、防护等行为词汇.
- 3) “生产与监管”:对应于疫情期间口罩生产制造情况,如口罩紧缺时期一些口罩厂复工生产口罩、部分商家哄抬口罩价格等,话题中主要词汇为价格、生产、市场、企业等.
- 4) “口罩进出口”:对应于疫情初期国外对国内口罩捐赠和中期我国口罩对外出口,话题中主要词汇为物资、海关、捐赠等.

与之相似,表2中还列举了部分时间窗口中的话题,如“口罩价格”,主要词汇为价格、涨价、药店、哄抬,对应事件为疫情初发时期口罩储备不足而口罩需求大,部分黑心商家存在哄抬口罩价格行为.此外还有国外疫情、高考防疫、海关查获等,话题中词汇特点较为明显,与其他话题区分度较大.

根据全局话题和局部话题挖掘结果,利用话题词袋与新闻内容匹配,得到新闻对应于话题的概率分布,取最大概率的话题作为该新闻对应的话题类型,统计后得各个时间窗口下话题新闻数比例趋势图图5.其中根据话题间相关关系,为突出话题特点,将社会疫情分成国内疫情与国外疫情两个部分,从图5中可知关于社会疫情和社会防疫的新闻为整体新闻的主要部分,并且到了疫情后期所占比例越来越高.关于口罩生产与监管和口罩进出口的新闻主要出现在前中期,随着疫情的稳定控制逐渐减少.

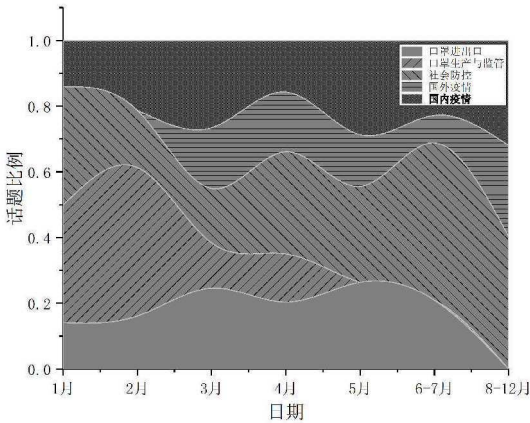


图 5 各类型新闻比例趋势

(Figure 5 Trends in the proportion of various types of news)

3.2 话题相关性分析

Niwattanakul 等人^[20]发现 Jaccard 系数可有效表现两个关键词数据集间的相关性.因此我们利用 Jaccard 系数计算各个话题之间的相关性.由于 LDA 所获得的的话题结果是词

的集合, 因此使用 Jaccard 系数可以有效地表示两个话题之间的相似性. 计算所得各个话题间相关性矩阵如图 6 所示.

根据图 6 (a) 中结果, 可以发现 4 个全局话题没有明显的相关性, 说明话题之间差异较为明显, LDA 结果较好, 所得话题能全面反映语料库中内容. “社会疫情”与“社会防控”两个话题分别对应社会中新疫情情况以及相应的防疫工作, 从图 6 (a) 中可以看到这个话题之间的相似性比另外两个话题更大, 与实际情况相符合. 此外, “口罩生产与监管”和“口罩进出口”的相关性也较大, 而在实际中口罩出口跟国内口罩生产相关. 因此这进一步能说明 LDA 话题模型的结果能较好反映实际情况, 话题之间特点鲜明且与现实相符.

对于不同时间窗口所获得的局部话题, 同样构造局部主体间的关系热图, 如图 6 (b) 中所示. 图中数字对应的按时间序列所获得的话题, 时间起点为第一个时间窗口至最后一个时间窗口. 从图 6 (b) 中可以看到邻近的话题之间相似性较大, 推测前后两个时间窗口中的话题间可能存在演化的关系. 此外, 还发现当两个话题的时间间隔变大, 话题之间的相关性会降低, 推测随着时间推移, 不同时间窗口的话题间存在演化关系, 并且后续话题的特征与初始话题相差较大. 也就是说, 从疫情爆发、应对到稳定控制的过程中, 人们对于口罩的关注热点会产生变化, 新闻报道的话题内容会发生转变.

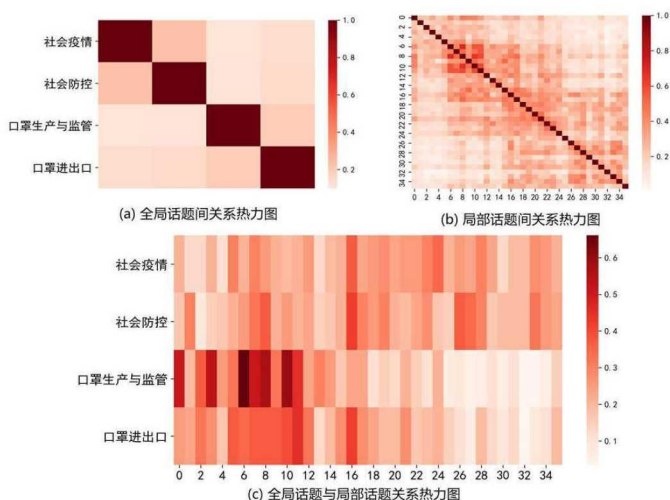


图 6 各话题间关系热力图

(Figure 6 Heat maps of the relationships between topics)

为验证这个推测, 本文构造全局话题与局部话题之间的相关性, 结果如图 6 (c) 所示. 从关系矩阵中可以看到, 所有局部话题都可以找到一个全局话题与之对应, 因此可以依据全局话题对局部话题进行分类, 根据分类后局部话题的分布情况可以分析全局话题在话题演化过程中的变化. 从图 6 (c) 中可以看到, 当疫情刚爆发时, 前 3 个时间窗口的话题主要是围绕“口罩生产与监管”和“口罩进出口”, 此时话题的相关性特别高, 为这两个话题的爆发期. 而到了疫情的后期, 跟这两个全局话题相关的局部话题逐渐消失, 说明此时人们关注的热点不再围绕“口罩生产与监管”和“口罩进出口”, 而对于“社会疫情”和“社会防疫”两个话题, 在

整个时间段都有出现,而分布比较均匀,到了后期局部话题基本是跟这两个话题相关.由图6(c)可以清楚地看到在随着时间推移,人们对于“口罩”相关新闻的关注话题存在明显的演化关系.

3.3 话题演化关系桑基图

根据实际中话题时序演化^[5]的情况,可以分为图7中话题产生、话题消失、话题继承、话题分裂和话题合并这五种关系.通过计算得到话题间 Jaccard 系数,并根据所设定的阈值,从而确定临近话题间的演化关系.设第 i 个时间窗口的话题集为 Z_i , z_{ij} 为第 i 个月的话题集中第 j 个话题.对于前后两个时间窗口生成的话题 $z_{ij} \in Z_i$ 和 $z_{i+1,j} \in Z_{i+1}$,使用 Jaccard 系数计算话题之间的相似性 $J(z_{ij}, z_{i+1,k})$.同时设定继承阈值,当 $J(z_{ij}, z_{i+1,k}) > \tau$,则前后话题存在继承关系; $J(z_{ij}, z_{i+1,k}) < \tau$ 时,则不存在继承关系,若 z_{ij} 与 $i+1$ 个时间窗口中所有话题均无继承关系,则话题演化消失.

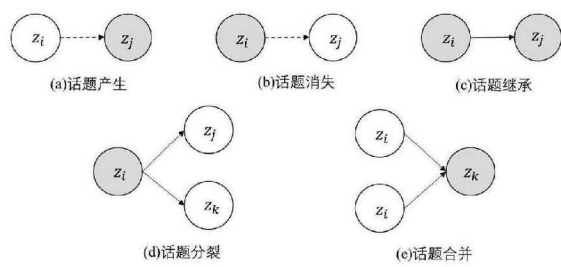


图7 话题演化关系^[5]
(Figure 7 Topical evolutionary relation^[5])

由前面分析我们可以推测,随着时间的进行,关于“口罩”新闻的话题存在演化关系.分析新闻中话题间的演化关系,可以得到不同阶段国家对于疫情的防控工作和社会关注重点内容.为进一步分析话题的演化情况,根据 Jaccard 系数确定临近时间窗口下话题间的演化关系,构建话题演化桑基图^[5, 24].设定继承阈值 $\tau = 0.15$,对每个话题取前 50 个词代表该话题,利用 Jaccard 系数计算得话题之间的演化关系,构造得桑基如图8所示.

图8中不同纹路的标签对应着不同的全局话题,其中“社会疫情”由浅灰色和带横线浅灰色部分组成,分别代表国内社会疫情和国外社会疫情.相同类型的话题间存在话题继承和分裂的情况,即存在话题演化路径.对不同时期的分析如下

疫情初发时期 1-2 月

“口罩生产与监管”和“口罩进出口”话题所占比重比较大,新闻报道主要关于口罩生产和口罩价格,相关话题为“口罩生产”、“口罩监管”和“口罩价格”.此时为国内口罩紧缺的时候,多家工厂加急生产口罩,同时政府发布相关规定控制口罩价格,遏制哄抬价格等现象.当国内口罩紧缺时在口罩进出口方面主要为海外对国内的口罩捐赠,同时还出现一些口罩进出口违规的情况,相关话题为“口罩捐赠”和“海关查获”.除此之外,还有关于新冠疫情爆发时社会疫情情况和防疫工作,相关话题为“社会疫情”和口罩防疫.

疫情中期 3-5 月

4 个全局话题比重相对平衡,此时国内疫情得到稳定控制,国内口罩供应充足,关于社会

疫情和防疫的新闻增多. 随着春节假期结束和疫情稳定控制, 对于复工复产后防疫工作的新闻变多, 对应为“复工复产”和“国内防控”话题. 而此时国外疫情爆发, 关于国外疫情和国外口罩生产的报道出现, 口罩进出口方面的报道集中为口罩对外出口的情况. 到了5月, 国内口罩供应充足, 关于“口罩生产与监管”的话题消失.

疫情后期 6-12月

主要话题为“社会疫情”和“社会防疫”, 国内疫情得到控制, 医疗物资供应充足, 基本没有关于“口罩生产与监管”话题的新闻, 而与“口罩进出口”相关新闻也逐渐减少, 到了8-12月话题消失. 此期间主要报道为国内外的疫情和防疫情况, 国内防疫主要为6-7月关于高考的防疫工作, 而国外则为相关颁布的法令.

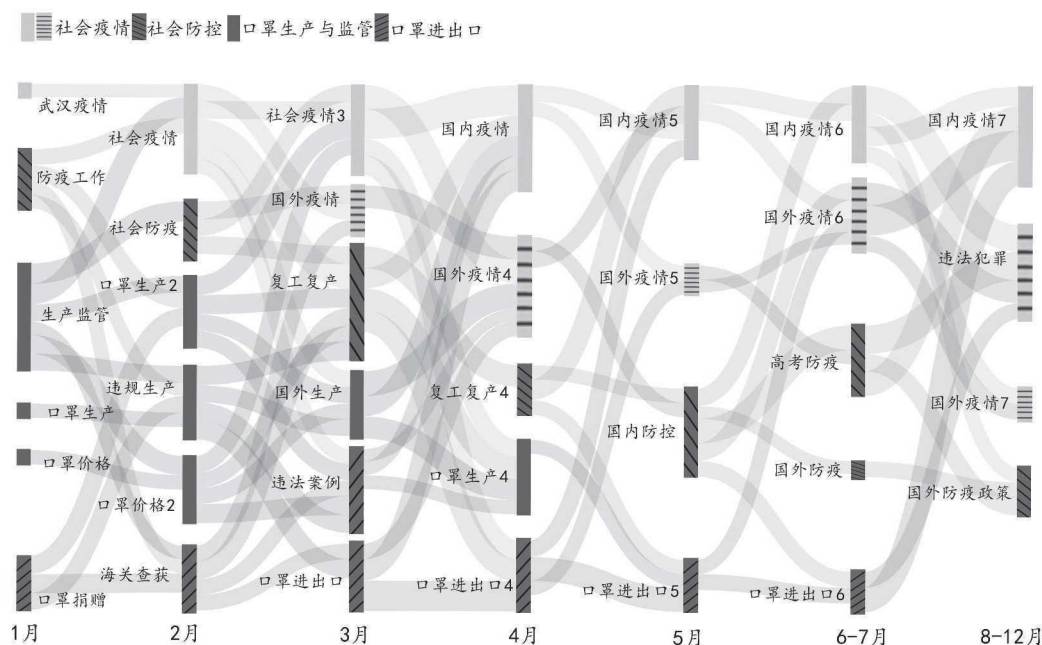


图8 中新网话题演化桑基图

(Figure 8 China news Topic Evolution San key Chart)

为进一步探究话题间演化的关系, 以1月到3月关于口罩进出口的全局话题为例, 分析前后话题间共现的词语(如图9所示). 在相同全局话题的演化路径中, 话题间共现的词语会更相似, 如口罩捐赠-海关查获, 海关查到-违法案例和口罩进出口3这几个继承关系中, “海关”、“捐赠”和“进出口”这几个与进出口相关的词出现的频率更高, 说明前后话题演化存在关联性. 而从海关查获-违法案例和口罩进出口3, 与海关查获-复工复产3和国外疫情的比较中可发现, 不同全局话题间传递的词语会存在区别, 如“警方”、“嫌疑人”和“诈骗”这几个词会出现在违法案例和查获事件中, 但在关于社会疫情和防疫的话题中没有出现. 这说明不同全局话题的演化存在差异性.

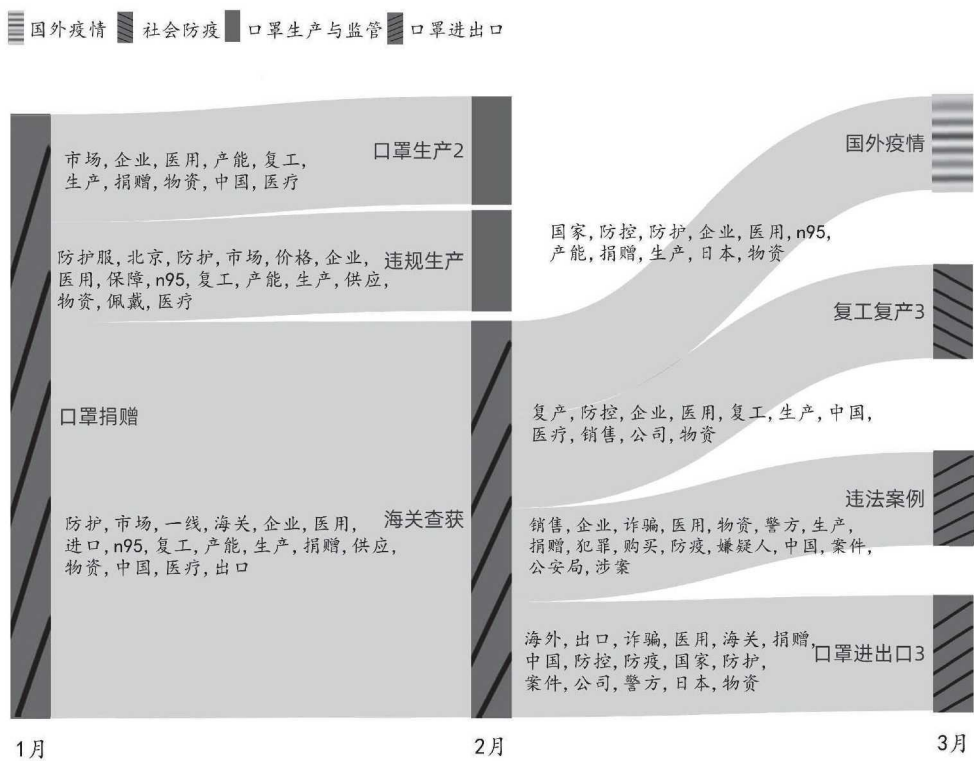


图 9 口罩进出口话题演化情况示例

(Figure 9 Examples of topic evolution of mask import and export)

本节利用 LDA 主题结果构造话题演化桑基图, 从图 9 中可知在疫情初期中新网主要新闻报道话题为“口罩生产和监管”和“口罩进出口”, 随着国内疫情的稳定控制, 这两个话题逐渐减少并消失, 最后演化到“社会疫情”和“社会防控”. 局部话题也由国内演化到国外, 如“口罩进出口”和“社会疫情”包含的局部话题. 相同全局话题间的演化存在关联性, 不同全局话题的演化存在差异性. 话题演化桑基图所分析的结果与实际情况相一致, 我们基于 LDA 话题模型构造的话题演化桑基图能揭示随着疫情的发展人们在不同阶段对于口罩医疗物资的关注热点, 也能侧面反映对新冠疫情的关注点.

4 基于社会网络的海关事件分析

前一节从话题演化角度进行分析, 可知疫情初期主要的话题为口罩生产和口罩进出口, 而随着时间的推移该话题消失并演化到社会疫情和社会防疫. 而在话题演化中存在关于海关违规进出口事件的话题, 为挖掘海关查获事件中的潜在信息, 本小节根据前面抽取的〈海关, 关键词〉二元组, 构造相应的网络, 通过分析网络的结构和特征, 挖掘海关查获事件之间的关系.

4.1 海关查获事件共现关键词网络

查获事件 51 个二元组中共涉及 29 个不同的海关及 24 个不同的关键词, 首先我们构建

共现关键词网络. 若两个关键词在同一件查获事件中出现, 则认为这两个关键词间存在共现关系. 根据计算所有关键词之间的共现关系, 得到共现关键词矩阵, 并构建共现关键词网络, 网络中的每个节点对应于一个关键词, 关键词间的连边表示这两个关键词存在共现关系.

共现关键词网络如图 10 所示, 网络中共有 19 个节点, 52 条边, 节点大小表示该节点的度. 边的权重为关键词共现的次数. 网络中平均聚类系数为 0.711, 模块化指数为 0.373, 节点连接紧密, 共现网络中存在模块化现象. 表 3 为共现关键词网络中心性分析结果, 分别为度、中介中心性和 PageRank 最大的前 5 个关键词. 我们可以看到“伪报”的 3 个中心性结果都是最高的, 说明它作为重要节点连接着其他的关键词, 与大多数查获事件均有联系. 另外“瞒报”、“申报不实”、“假冒伪劣”和“逃避检验”这几个词在表 3 中出现了两次以上, 可推测这几个关键词在共现关键词网络中也处于重要的位置, 作为重要节点与其他关键词相连, 表现出查获事件的特点.

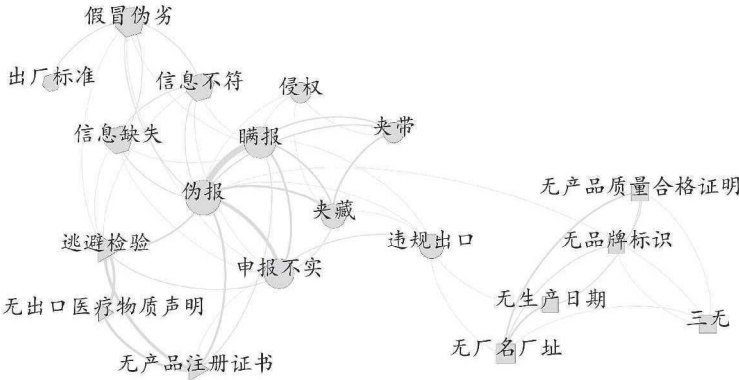


图 10 共现关键词网络
(Figure 10 Keyword co-occurrence network)

表 3 关键词中心性分析结果
(Table 3 The centrality analysis results of keywords)

关键词	degree	关键词	Betweenness centrality	关键词	Pageranks
伪报	12	伪报	56.9381	伪报	0.102911
瞒报	10	违规出口	26.6309	瞒报	0.085467
申报不实	9	三无	25.3690	申报不实	0.076946
假冒伪劣	7	假冒伪劣	17.1667	假冒伪劣	0.065876
逃避检验	7	瞒报	16.1595	逃避检验	0.063520

为找出关键词之间的聚类情况, 利用 Louvain 算法对关键词做社区化计算. Louvain 算法^[21]利用模块度 Q 表示社区内紧密程度, 模块度增量表示合并节点前后社区模块度的变化值, Q 的计算公式如下

$$Q = \frac{1}{2} \sum_{ij} \left[A_{ij} - \frac{k_i * k_j}{2m} \right] * \delta(C_i, C_j), \tag{1}$$

其中 A_{ij} 为邻接矩阵系数, k_i 和 k_j 为点的度, m 为网络总边数, C_i 和 C_j 为 i 和 j 所在社区,

如果 $C_i = C_j$, 则 $\delta(C_i, C_j) = 1$, 否则为 0. 根据 Louvain 算法对共现关键词社区化结果如下所示, 可分成以下四类关键词

- 1) 夹带私藏: 以瞒报为主, 在其他货物中夹带私藏口罩;
- 2) 逃避检验: 缺少相关证明, 而通过伪报手段逃避海关检验;
- 3) 三无产品: 三无产品, 缺少重要产品信息的违规出口;
- 4) 假冒伪劣: 信息不符或质量问题, 假冒伪劣生产的口罩.

在所有海关查获事件中出现的关键词可分成图中四类情况, 同类别的关键词有更高的概率共同出现在一起查获事件中. 图 10 中各个社区内关键词的性质相近, 均对应于某一类查获事件. 如在“三无产品”社区中的关键词为“无厂名厂址”、“无品牌标识”、“无生产日期”和“无产品质量合格证明”, 均是与三无产品质量问题有关的. 图 10 中“违规出口”作为割点连接夹带私藏和三无产品两类关键词, “伪报”和“瞒报”则为连接逃避检验、假冒伪劣和夹带私藏的主要节点.

4.2 海关 - 关键词二模网

为进一步探究各个海关与查获关键词类型之间的关系, 本节依据海关和关键词关系构建海关 - 关键词二模网, 并利用前面共现关键词网络中社区化结果对二模网进行分类. 网络中节点由海关和关键词组成, 海关和关键词的连边表示发生在该海关的查获事件中包含该关键词.

见图 11, 二模网通常可加权映射到单模网^[22, 23]然后再做社区划分. 共现关键词网络中关键词的共现关系是对于某一起查获事件而言的, 而每一起查获事件只会在一个海关中发生, 关系如图 12 所示. 每个海关可由在该海关发生的事件集合表示, 可构建关键词 - 事件 - 海关的映射关系. 我们结合关键词在海关中的共现频率, 和共现关键词社区化的结果对海关 - 关键词二模网进行聚类.

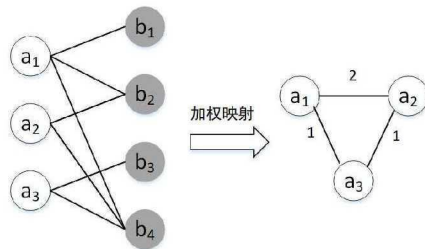


图 11 二模网投影
(Figure 11 Projection of bipartite network)

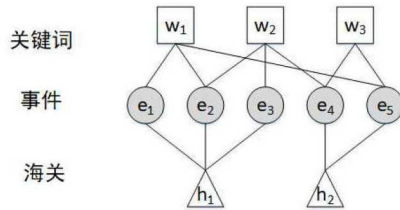


图 12 关键词、事件和海关关系
(Figure 12 The relationships of keywords, event and customs)

假设 W, E, H 分别对应于关键词、事件和海关的集合, 集合大小分别为 M, N 和 K . 设关键词 w_i 和 w_j 在所有海关查获事件集合中共现了 n 次, 即 $A_{ij} = n$ (A 为共现关键词矩阵), 其中在海关 H_k 的事件集 $E_k = \{e_1, e_2, \dots, e_j, \dots, e_{N_k}\}$ 中共现了 t_{ij}^k 次, 并且 $\sum_{k=1}^K t_{ij}^k = n$, 那么海关 H_k 中对于关键词 w_i 的共现指数可记为

$$\varphi_i^k = \frac{\sum_{e_j \in E_k} t_{ij}^k}{\sum_{w_i \in W} \sum_{e_j \in E_k} t_{ij}^k}, \quad (2)$$

即在海关 H_k 所有事件中关键词 w_i 的总共现次数除以海关 H_k 中所有共现关键词的总次数. 设 M_l 为共现关键词网络中第 l 个社区中所有关键词集合, 那么海关 H_k 对于社区 l 的社区化指数为 $\theta_{lk} = \sum_{w_i \in M_l} \varphi_i^k$, 令社区化指数最大的社区 l_0 则为海关 H_k 所对应的社区, 即 $l_0 = \operatorname{argmax}(\theta_{lk}, \forall l \in M_l)$.

根据以上方法构建海关-关键词二模网, 并依据共现关键词进行社区划分, 结果如图 13. 图 13 中共有 29 个海关和 24 个关键词, 节点间的连边表示该关键词在该海关上出现, 边的权重为出现的次数. 二模网中的社区包括海关节点和关键词节点, 社区划分最基本依据是关键词间的共现关系. 出现在同一个社区中的关键词均属于同个事件类型, 与他们处于同个社区里的海关为与他们关系最亲密的, 即在這些海关上发生的查获事件很大概率是属于该关键词事件类型的.

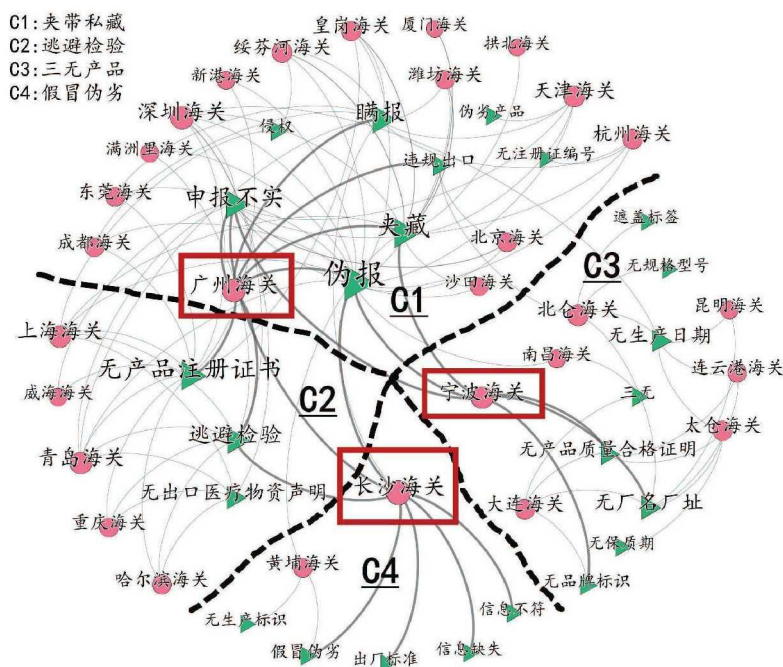


图 13 海关-关键词二模网

(Figure 13 Customs-Keywords bipartite network)

图 13 中有 4 个社区, 分别对应前面出现的夹带私藏、逃避检验、三无产品和假冒伪劣四类事件. 从图 13 中可以看到各类查获事件主要出现的海关分布, 其中广州海关、宁波海

关和长沙海关为割点,分别为连接社区对“夹带私藏-逃避检验”、“夹带私藏-三无产品”和“逃避检验-假冒伪劣”中的关键词。“假冒伪劣”、“信息不符”、“信息缺失”和“出厂标准”这几个关键词在共现关键词网络中属于同一个社区,而从二模网中可以看到这几个词只出现在黄埔海关和长沙海关,这说明二模网的社区化结果能较好的反映出关键词与海关的共现情况.另外根据分析各社区中海关的地域分布情况,发现这些海关没有在地域上的关联性,因此海关的社区划分只与在海关中发生的事件类型有关,与地域无关.

本节从社会网络分析的角度,构建共现关键词网络和海关-关键词二模网,分析关键词内部以及海关与关键词间的关系.从共现关键词网络中利用 Louvain 算法将查获事件分成夹带私藏、逃避检验、三无产品和假冒伪劣四类.随后构建海关-关键词二模网,利用共现次数加权的方式将各个海关归类到对应的关键词社区中,得到关键词与他们出现的海关之间的分布关系.我们基于单模网和二模网的分析结果揭示了疫情期间海关查获事件的类型特点,得到了海关查获事件间的聚类情况,反映出我国海关在口罩进出口中核查的重点对象和方式.

5 结束语

本文以中国新闻网和海关总署官网关于“口罩”的新闻报道为数据源,通过 LDA 模型抽取新闻话题和构建查获事件社会网络两个工作,分析疫情期间人们对于口罩医疗物资的关注热点变化趋势,以及口罩违规进出口事件的性质.在 LDA 话题抽取中,本文根据最大一致性原则确定文本中的话题数,分别生成全局话题和各个时间窗口的局部话题.通过计算话题之间的 Jaccard 系数,得到话题之间的相关性矩阵,发现各个局部话题与 4 个全局话题“社会疫情”、“社会防控”、“口罩生产与监管”和“口罩进出口”间存在对应关系,即随着时间的进行,时间窗口中的局部话题存在演化关系,如话题继承和话题消失.针对于话题演化关系,本文构建话题演化趋势桑基图,发现随着疫情稳定控制,在疫情初期出现的“口罩生产与监管”和“口罩进出口”全局主题会逐渐减少并消失,最后演化为“社会疫情”和“社会防控”两个话题.同时针对口罩违规出口事件,构造话题词演化流动桑基图,分析在同类和不同类话题演化中,话题重点词语的流动关系.为进一步探究此类海关查获事件中的性质,本文从语料库中抽取出海关查获事件的〈海关,关键词〉二元组,利用关键词在事件中的共现关系,构建共现关键词网络,并利用 Louvain 算法得到关键词间存在社区化结构,海关查获事件可根据关键词分成四类事件.随后构建海关-关键词二模网,利用关键词共现频率指数对海关和关键词归类,得到海关与关键词的分布关系,并找到连接不同查获事件的海关割点.

本文从话题演化和事件结构两个视角对新冠疫情社会舆情展开分析,利用关于“口罩”这个重点医疗物资的新闻语料库,从 LDA 话题演化的角度分析人们关注医疗物资热点的话题演化趋势,并从社会网络分析的角度分析关于医疗物资违规出口事件的内在性质.社会热点话题的网络舆情一直是人们关注的对象,而分析社会舆情的动态变化以及热点事件的内在性质是两个重要的手段.本文结合这两种方法,从“口罩”热点话题演化和“口罩”重要事件性质两个角度,有效地展示了人们对于“口罩”医疗物资的关注热点和事件性质.医疗物资作为突发公共卫生事件中重要的角色之一,如何从公众的关注热点反映出不同时期社会的需求关系,是一项值得研究的工作.未来也希望能继续结合 LDA 和社会网络的方法,用于

更多话题与事件间的分析中去,如突发事件的网络舆情、研究主题的演化与结构分析等,利用 LDA 主题挖掘找出事件的演化情况,或利用社会网络分析事件的结构。

本文仅分析了单一类型语料,未从多种类型的语料中挖掘话题内容,普遍性有待提高,今后将尝试搜集更多类型的语料,从更多角度反映人们关注的热点话题。此外,本文所使用的方法均是前人所提出的,虽然方法有效但仍存在局限性,对于方法间有效性对比和方法改进有待深入,未来将对比不同方法之间的优劣,甚至提出改进的方法,这对有效挖掘事件间的演化关系极为重要。

参 考 文 献

- [1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, **3**: 993–1022.
- [2] Xu G, Meng Y, Chen Z, et al. Research on topic detection and tracking for online news texts. *IEEE Access*, 2019, **7**: 58407–58418.
- [3] 曹丽娜, 唐锡晋. 基于主题模型的 BBS 话题演化趋势分析. 管理科学学报, 2014, **17**(11): 109–121.
(Cao L N, Tang X J. Analysis of dynamic topics evolution based on Tianya club. *Journal of Management Sciences in China*, 2014, **17**(11): 109–121.)
- [4] 张亚茹, 唐锡晋. 天涯杂谈“红会贴”热点事件分析. 系统科学与数学, 2020, **40**(10): 1723–1736.
(Zhang Y R, Tang X J. Analysis of hot events of the red cross posts from Tianya club. *Journal of Systems Science and Mathematical Sciences*, 2020, **40**(10): 1723–1736.)
- [5] Yan Z H, Tang X J. Exploring evolution of public opinions on Tianya club using dynamic topic models. *Journal of Systems Science and Information*, 2020, **8**(4): 309–324.
- [6] 李振鹏, 黄帅. 基于 LDA 主题模型的网络舆情研究. 系统科学与数学, 2020, **40**(3): 49–62.
(Li Z P, Huang S. Analysing on network public opinion based on LDA topic model. *Journal of Systems Science and Mathematical Sciences*, 2020, **40**(3): 49–62.)
- [7] 郗强, 唐锡晋. 基于社会网络分析的“系统科学大会”可视化探析. 系统科学与数学, 2018, **38**(5): 521–536.
(Xi Q, Tang X J. Visual exploration of CSSC2017 based on social network analysis. *Journal of Systems Science and Mathematical Sciences*, 2018, **38**(5): 521–536.)
- [8] 钱芸芸, 杨文忠, 姚苗, 等. 融合主题相似度权重的主题社区发现模型. 计算机工程与应用, 2021, **57**(5): 107–114.
(Qian Y Y, Yang W Z, Yao M, et al. Topic community discovery model incorporating topic similarity weight. *Computer Engineering and Applications*, 2021, **57**(5): 107–114.)
- [9] L’Huillier G, Alvarez H, Ríos S A, et al. Topic-based social network analysis for virtual communities of interests in the dark web. *ACM SIGKDD Explorations Newsletter*, 2011, **12**(2): 66–73.
- [10] Batagelj V, Mrvar A. Density based approaches to network analysis. Analysis of Reuters Terror News Network, 2001.
- [11] Moessner M, Feldhege J, Wolf M, et al. Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders*, 2018, **51**(7): 656–667.
- [12] 李沿江, 赵红霞, 苏玲霞. 新冠肺炎疫情期间微博话题“中医新冠肺炎”的关注热点分析——基于 LDA 模型的微博话题挖掘. 亚太传统医药, 2020, **16**(11): 15–17.
(Li Y J, Zhao H X, Su L X. Analysis of the hot spots of the Weibo topic “TCM COVID-19” during the epidemic of COVID-19 — Microblog topic mining based on LDA model. *Asia-Pacific Traditional Medicine*, 2020, **16**(11): 15–17.)

- [13] 杨建梁, 刘越男, 祁天娇, 等. 重大公共卫生事件中民众诉求的话题挖掘与演变透视. 图书馆论坛, 2021, 41(4): 121–131.
(Yang J L, Liu Y N, Qi T J, et al. Topic mining and evolution analysis of public demands during major public health events. *Library Tribune*, 2021, 41(4): 121–131.)
- [14] 王袁欣, 刘德寰. 框架理论视角下西方主流媒体新冠肺炎疫情报道的 LDA 主题模型分析 —— 以《纽约时报》和《卫报》为例. 广告大观 (理论版), 2020, (3): 76–89.
(Wang Y X, Liu D H. LDA topic modeling analysis for COVID-19 news of western mainstream media from the perspective of framing analysis: An analysis of New York times and guardian publications. *Advertising Panorama*, 2020, (3): 76–89.)
- [15] Mutanga M B, Abayomi A. Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach. *African Journal of Science, Technology, Innovation and Development*, 2020, <https://doi.org/10.1080/20421338.2020.1817262>.
- [16] Ahmed W, Vidal-Alaball J, Downing J, et al. Dangerous messages or satire? Analysing the conspiracy theory linking 5G to COVID-19 through social network analysis. *Journal of Medical Internet Research*, 2020, 22(5): e19458.
- [17] Newman D, Noh Y, Talley E, et al. Evaluating topic models for digital libraries. Proceedings of the 10th Annual Joint Conference on Digital Libraries, 2010, 215–224.
- [18] Stevens K, Kegelmeyer P, Andrzejewski D, et al. Exploring topic coherence over many models and many topics. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, 952–961.
- [19] Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. Proceedings of the eighth ACM International Conference on Web Search and Data Mining, 2015, 399–408.
- [20] Niwattanakul S, Singthongchai J, Naenudorn E, et al. Using of Jaccard coefficient for keywords similarity. Proceedings of the International Multiconference of Engineers and Computer Scientists, 2013, 380–384.
- [21] Vincent D B, Jean-Loup G, Renaud L, et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 10: P1008.
- [22] 张婧婧, 黄廷磊, 张银明. 基于聚类分析的二分网络社区挖掘. 计算机应用, 2015, 35(12): 3511–3514.
(Zhang Q Q, Huang T L, Zhang Y M. Detecting community in bipartite network based on cluster analysis. *Journal of Computer Applications*, 2015, 35(12): 3511–3514.)
- [23] Guimera R, Sales-Pardo M, Amaral L A N. Module identification in bipartite and directed networks. *Physical Review E*, 2007, 76(3): 036102.
- [24] Rule A, Cointet J P, Bearman P S. Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. Proceedings of the National Academy of Sciences of the United States of America, 2015, 10837–10844.