doi:10.3969/j.issn.1001-4616.2023.03.014

多源区域民生话题演化技术研究

张晓明,申　晴,王　芳,赵培森,于占鲁

(北京石油化工学院信息工程学院,北京 102617)

[**摘要**]　民生一直是社会重点话题,近两年的疫情防控又为话题聚焦和演化注入了新的内容. 本文基于大量区域化民生数据进行 LDA 模型的困惑度分析,证明多源文本话题比单源文本更全面. 并进一步提出了民生话题演化技术框架,创新设计了热度演化率和关键词演化率的计算方法和实现算法. 基于 HTDI 模型和关键词演化率,综合设计了民生话题演化指数 LTEI. 实验数据采集于北京大兴区的官方微博和百度贴吧. 实验结果表明,TF-IDF 模型比 TextRank 模型更合适计算关键词演化率;与 HTDI 指数相比,LTEI 指数与实际话题演化趋势更加贴合,更适合用于区域民生话题演化分析.

[**关键词**]　话题演化,民生,演化率,演化指数,新冠肺炎疫情
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2023)03-0105-07

Research on the Evolution Technology of Livelihood Topics in District Areas from Multiple Data Resources

Zhang Xiaoming, Shen Qing, Wang Fang, Zhao Peisen, Yu Zhanlu

(College of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing 102617, China)

Abstract: The topic of people's livelihood has always been a key social issue. The epidemic prevention and control in the past two years has injected new content into the focus and evolution of the topic. Based on a large number of collected regional livelihood topic data, the perplexities as LDA model are analyzed to show that the LDA topics from the multiple source data are more comprehensive than the individuals. Then, a kind of technique framework of livelihood topic evolution is put forward firstly. Some new ideas of heat evolution rate (ER) and keyword ER are created with detail definition and concrete algorithms. Furthermore, based on the HTDI model and keyword ER, the comprehensive model as livelihood topic evolution index (LTEI) is designed for topic evolution process. The data set is collected online from official Weibo, Baidu Tieba mainly in Daxing District of Beijing. The experimental results show that the TD-IDF model is more suitable for keyword ER than TextRank model. Compared with HTDI, the LTEI is more consistent with the evolution trend of actual topics and is more suitable for the evolution of regional livelihood topics.

Key words: topic evolution, livelihood, evolution rate, evolution index, new crown pneumonia epidemic

"民生"涵盖了人的生命、健康、权利、尊严等全部内容,并随着社会变迁和环境变化而呈现出鲜明的时代性. 政府可以通过民生话题的演化趋势,来判断某民生问题是否得以解决:话题热度升高,即是民众关注度高,该民生问题未能解决. 话题演化的定义为话题随时间的变化,包括话题强度和话题内容的变化[1],衡量的是同一话题随时间推移表现出的动态性、发展性和差异性. 在话题内容的演化评估上,可使用话题在每个时间点上的绝对长度和正则化长度来评测话题内容随时间的演化,或者将关键词在一个时间段内出现的次数作为演化评估标准. 在话题强度的演化评估上,可使用话题的平均得分和累积得分来评测话题和子话题的强度随时间的变化. 也有使用话题在某一时间间隔中所占的比重作为衡量标准,来衡量话题强度的演化[2]. 钱莉等[3]通过对 74 篇话题演化研究文献的研读分析,认为话题强度、话题状态、话题内容以及演化路径是话题演化分析的主要维度. 在话题强度演化过程中,应该考虑特征词或主题词对话题的贡献. 已有的话题演化趋势预测相关工作多是预测话题强度,很少对话题内容演化趋势进行预测.

收稿日期:2022-08-08.
基金项目:北京市优秀人才项目(ZZB2019005)、北京市科技计划一般项目(KM202010017011).
通讯作者:张晓明,博士,教授,研究方向:网络信息隐藏、大数据技术与智能计算. E-mail:zhangxiaoming@bipt.edu.cn

近年来,疫情传播和防控成为了热点的民生话题之一. 刘怡君等[4]从网民、信息、心理和观点四个研究维度展开分析,从人民网舆情频道梳理出代表性的 164 件非常规突发事件,并分为政府信用、社会管理、征地拆迁、民生问题等八类. 通过时序分析表明,民生问题类和重大事故类突发事件观点较独立,“民生问题”中民众具有从众、从亲与从利的心理倾向分布. 唐丽等[5]针对 2020 年疫情防控过程的注意力配置,研究了疫情阻断、民生保障、经济发展以及政府监管 4 个向度,结果表明民生保障向度的政策在高峰期和平稳期都有较高的配置占比. Bai 等[6]认为有三种典型方法计算话题一致性分别是 UCI、UMass 和 NPMI. 从 Kaggle 数据集获得来源于 CBC 的 COVID-19 相关消息文章 3534 篇,并采用热力图表示 7 个时间片段之间 20 个话题的演化情况. 龚晓康等[7]则提出了一种文本邻近度判定模型 PDRBL,并基于此提出 TETP 话题追踪方法用于分析社交媒体信息,还给出了“疫情”话题的演化案例.

针对话题热度分析问题,一些研究方法包括直接将话题的帖子数或点击数作为话题热度,以及综合考虑时间、关注度、转发数、用户等多种因素来定义话题热度的方法. 裴可锋等[8]结合表示文本重要性的基于文本外在特征的热度因素和基于内容的热度因素计算方法,定义了话题热度表示公式. 唐晓波等[9]结合转发数和评论数,以信息量的思想提出了计算微博热度的公式,但未考虑时间因素的影响. 陈兴蜀等[10]提出了基于热点话题演化与跟踪的 HTOLDA 模型,该模型对各个时间片论坛数据的建模能力均优于 OLDA 模型,而且能有效地对论坛中热点话题的内容与强度进行演化跟踪. Zhu 等[11]提出新模型 CTH,采用了分布相似度增强的贝叶斯玫瑰树,更能反映文本数据之间的真实关联关系,结果展示了一组持续 1 周的新浪微博热门话题的演化过程.

在话题演化模型及其评价方面,Mei 等[12]研究时态文本挖掘技术,提出了主题演化图、主题跨度、演化传递等相关概念. 研究自动获取演化主题模式,其中采用 KL 散度来度量 2 个主题跨度的演化距离,定义主题生命周期来表示在整个周期中每个主题的强度分布. 针对 2 个数据集进行案例分析,包括 2004 年亚洲海啸灾难事件的 8 万篇新闻文章、KDD 会议自 1999 到 2004 年的论文摘要,能够从时间角度有效地分析和总结话题演化结构. BLEM 等[13]提出动态主题模型 DTM(dynamic topic models). 该模型引入了时间窗口,关注主题词分布和文档主题分布随时间的演化,并利用非参数小波回归和卡尔曼滤波改进变分推断算法. Wang 等[14]提出 TOT(topics over time)模型,在 LDA 模型的基础上,引入了服从 Beta 分布的时间变量,能够有效地预测主题分布随时间的演化. Patrick 等[15]在定量评价方面,采用 PMI 计分方法测试话题的一致性,并通过 2 个实例展示话题的年度演化过程. 李纲等[16]的研究是以大气灾害类型的热点事件为代表的案例,揭示和比较了两类热点事件民众的关注度及其在宏观层面和微观层面的话题变化及其演化规律. 黄微等[17]以“高以翔去世”事件作为数据样本,提出 HTDI 指数以表征话题演化情况,对该事件的爆发期以及蔓延期进行了研究. 张佩瑶等[18]通过计算相邻时间片的主题相似度分析微博的话题演化. 以新浪微博“勒索病毒”话题为例,设置主题之间相似度阈值为 0.5,确定相邻时间片主题之间的演化关系,从主题内容方面进行演化分析. 该方法能利用词向量充分挖掘词语间的语义关系,提高话题间的区分度,但仍不能给出话题演化的衡量标准.

本文针对当前民生话题研究少、话题演化评价标准缺乏的问题,从困惑度分析入手,基于民生多源数据的采集和分析,提出了新的演化计算架构和民生话题演化计算方法. 在此基础上,基于 HTDI 指数[17]的不足,设计新的演化模型. 最后,对北京市大兴区疫情话题的演化情况进行实证分析,取得了令人满意的效果.

1　多源文本民生话题的有效性分析

模型泛化能力是衡量模型对未观测到的数据的预测能力. 比较公认的判断方法是衡量模型的困惑度. 困惑度越小,表示模型的泛化能力越强.

通过采集 2020 年所有月份的数据,包括微博和百度贴吧数据,分别表示为 S_α、S_β. 分析三种文本 S_α、S_β 和组合数据 $S_{(\alpha+\beta)}$ 的最低困惑度的曲线图,如图 1 所示. 从图 1 可以得知,合成的多源文本 $S_{(\alpha+\beta)}$ 的最低困惑度值,高于单源文本 S_α 和 S_β 的最低困惑度值. 根据困惑度的原理可知,这是因为多源文本 $S_{(\alpha+\beta)}$ 是由多种单源文本 S_α 和 S_β 合成的,由于融合了新的文本内容,其内容繁杂程度高于单个的单源文本,所以在同一月份下,多源文本 $S_{(\alpha+\beta)}$ 的最低困惑度值高于单源文本 S_α 和 S_β 是合理的.

接着,关于三种文本的每月最低困惑度值,即 S_α、S_β 和 $S_{(\alpha+\beta)}$ 对应的最佳话题数如图 2 所示. 从图中可

见,2020 年 2 月和 4 月,多源文本 $S_{(\alpha+\beta)}$ 的最佳话题数低于单源文本 S_α、S_β 的最佳话题数,而剩余月份的 $S_{(\alpha+\beta)}$ 的最佳话题数趋势高于 S_α、S_β. 这个结果表明,多源文本在进行实验分析时优于单个的单源文本,多个单源文本的融合增加了话题,使语料更加丰富. 因此,经过 LDA(Latent Dirichlet Allocation)主题模型聚类后,多源文本的最佳话题数优于各个单源文本符合常理. 所以,LDA 主题模型可以应用于多源文本分析,我们将在后续实验中使用多源文本对大兴区民生区域话题进行演化分析.

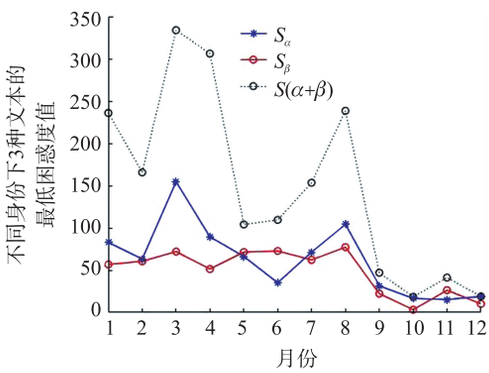

图 1　三种文本每月的最低困惑度

Fig. 1　Minimum perplexity of three texts per month

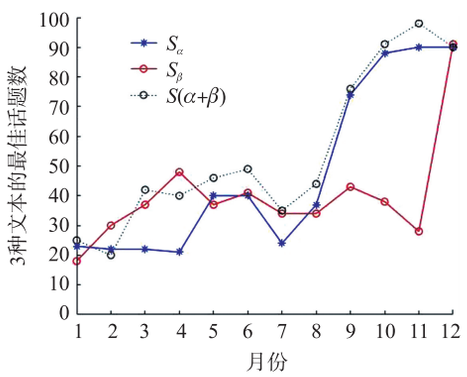

图 2　三种文本每月的最佳话题数

Fig. 2　The best number of topics for three texts per month

2　多源民生话题演化技术框架设计

2.1　HDTI 的局限性分析

黄微等[17]以困惑度为突破点,根据不同时间段的困惑度得到最佳话题数,同时使用 LDA 模型计算出话题-词的矩阵,通过前后时间差的关键特征词的权重差来得到话题漂移指数 HTDI,以此得到话题的演化趋势. 具体的 HTDI 指数公式如式(1)所示[17]:

$$\mathrm{HTDI}=\frac{\sum_{i=1}^{K} w_{Ti}-\sum_{i=1}^{k} w_{ti}}{T-t} \tag{1}$$

可见,HTDI 仅通过单一话题的特征词总权重来定义漂移指数过于片面. 同时,LDA 模型只是适用于挖掘词汇在文档级别上的共现关系,且多次主题聚类的次序不定,词权重具有数值浮动性. 因此,需要增加多源信息,并融入词汇在话题上的重要性.

2.2　多源民生话题演化框架设计

针对民生话题的聚焦和演化需求,应该基于宏观和微观相结合的思想. 宏观上,采用 LDA 模型在文档上的生成话题结果和漂移指数;微观上,利用民生话题中共生作用明显的关键词.

由此,我们设计了多源数据的区域化民生话题演化框架,如图 3 所示. 选择一个县市类的区域,其关注的话题比较集中,时效性高,话题演化的研究价值对政府的指导作用更明显. 通过互联网主要社交媒体

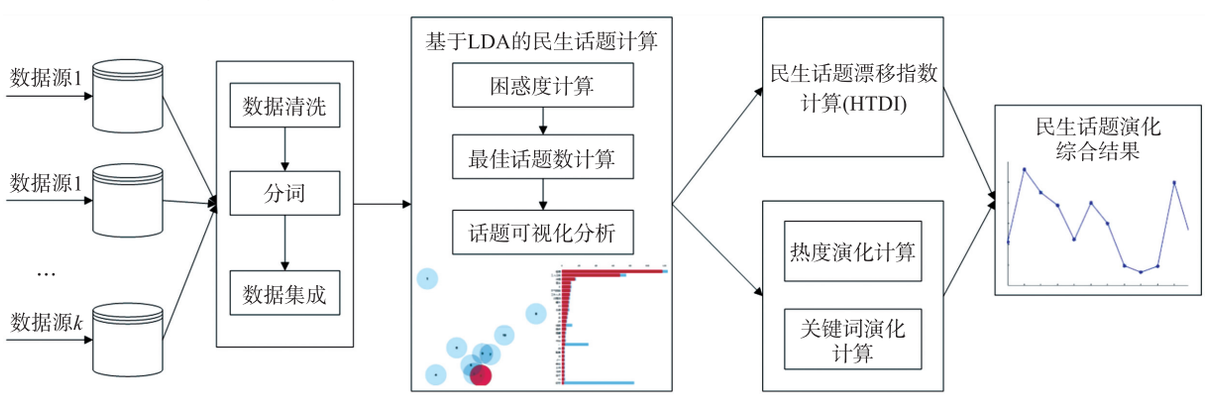

图 3　多源区域民生话题演化技术框架

Fig. 3　Framework of topic evolution for district livelihood

交流渠道,如微博、贴吧、新闻网等多种类数据源,每天采集对应的民生文本数据,并按天保存到数据集中. 经过数据清洗、中文分词等处理后,按时间和类型保存,并开展数据集成. 由于演化过程需要长时间积累,因此需要构造大规模的数据集. 随后,采用 LDA 开展话题分析和关键词权重计算,并通过困惑度优化计算,得到最佳话题数结果,获得某时段需要关注的民生话题. 在此基础上开展话题演化分析,一是基于话题中关键词权重矩阵值,计算式(1)值,二是按照指定的一类民生话题中的关键词集,开展关键词在时域上的热度计算和演化计算. 最后通过两者的综合,求得民生话题的综合演化结果.

3 民生话题演化率设计

为了便于精准描述话题的演化过程,需要针对特定话题的关键词进行跟踪.

3.1 热度计算分析

使用 TextRank 和 TF-IDF 两种算法提取关键词,通过分析比较,选出一个最优方法.

定义 1 热度. 指文本数中的特定话题文本数占比,以求得特定话题要点,其计算方法如下:

$$H_i=\frac{Y_i}{C_i} \tag{2}$$

式中,Y_i 特定话题文本数,C_i 为文本数,i 为月份,即 1 到 12 个月.

定义 2 热度演化率. 用于描述相邻两个月热度变化情况,以求得特定话题基准. 其它演化值应该以此为参照. 其计算方法如下:

$$r_H=\frac{H_{i+1}-H_i}{mon_i} \tag{3}$$

式中,H_i 是第 i 月的热度值,H_{i+1}为第(i+1)月的热度值,mon_i 为 12 个月份里各月份的具体天数:

$$mon_i=[31,29,31,30,31,30,31,31,30,31,30,31] \tag{4}$$

3.2 关键词演化计算思路

由于 TextRank 和 TF-IDF 两种算法原理不同,所以计算出的关键词词频的区间也不同,无法直接比较其优劣,所以提出关键词演化率的概念,旨在进行归一化处理. 以下的 TR 代表 TextRank 值,TI 代表 TF-IDF 值.

定义 3 关键词演化率.用于描述相邻 2 个月的关键词变化情况,对应 TextRank 和 TF-IDF 两种方法,其关键词演化率 $r_{TR,i}$和 $r_{TI,i}$分别定义为:

$$r_{TR,i}=\frac{w_{TR,i+1}-w_{TR,i}}{mon_i} \tag{5}$$

$$r_{TI,i}=\frac{w_{TI,i+1}-w_{TI,i}}{mon_i} \tag{6}$$

式中,$w_{TR,i}$ 即为使用 TextRank 方法后得到的某词的权重,i 为第 i 个月,以此类推.

最后,在特定话题的最佳关键词选取上,采用伪标准差计算方法,来选取关键词演化率. 基于 TextRank 和 TF-IDF 的伪标准差 σ_{TR}、σ_{TI}计算方法如式(7)和(8)所示:

$$\sigma_{TR}=\sqrt{\frac{\sum_1^K(w_{TR,i}-r_{H,i})^2}{K}} \tag{7}$$

$$\sigma_{TI}=\sqrt{\frac{\sum_1^K(w_{TI,i}-r_{H,i})^2}{K}} \tag{8}$$

式中,K 为关键词数量. 通过比较这 2 个值,按较小值选择关键词演化率,确定为 r_T.

3.3 基于关键词的民生话题演化算法设计

基于上述思路和定义,下面先设计该演化过程,从关键词获取开始,一直到计算出关键词演化率. 通过分析最佳民生话题数结果,选定一类关键词集合,比如围绕“疫情防控”这一特定热点话题,来分析其演化趋势.

算法 1　关键词演化率计算算法

输入:特定话题关键词集合 S_C,数据集 D,月份天数集合 Mon;
输出:关键词演化率 r_T.
①导入 D
②for each in Mon/ * 对每个时段
③　用式(3)计算 r_H/ * 计算热度演化率
④end for
⑤for each in S_C/ * 对每个词汇
⑥　for each in Mon
⑦　　按 TextRank 模型计算权重 $w_{TR,i}$
⑧　　按 TD-IDF 模型计算权重 $w_{TI,i}$
⑨　　用式(5)计算演化率 $r_{TR,i}$
⑩　　用式(6)计算演化率 $r_{TI,i}$
⑪　　用式(7)计算伪标准差 σ_{TR}
⑫　　用式(8)计算伪标准差 σ_{TI}
⑬　end for
⑭end for
⑮if($\sigma_{TR}<=\sigma_{TI}$)
⑯　$r_{T,i}=r_{TR,i}$,转第 20 步
⑰else
⑱　$r_{T,i}=r_{TI,i}$
⑲end if
⑳输出 $r_{T,i}$

3.4　民生话题演化的综合设计

针对民生话题数据来源多样和精准描述需求,需要改进 HTDI 进行话题演化分析. 以热度演化率为参考依据,对 HTDI 和关键词演化率的计算结果进行综合设计.

定义 4　民生话题演化指数. 指从多源民生数据出发,综合分析文档级和关键词级对民生话题演化的贡献,提出构建区域民生话题演化指数(livelihood topic evolution index,LTEI),来改进 HTDI 指数. LTEI 的计算方法如下:

$$LTEI=HTDI^{\varepsilon}+r_T \tag{9}$$

式中,将每相邻两个月的 HTDI 的 ε 次方与关键词演化率进行加性综合,$\varepsilon\in[1,2]$.

因此,针对某一特定民生话题,可通过求解 LTEI 来分析其演化过程.

4　实验分析

4.1　区域民生数据源

为了描述区域特征,本研究选择北京市大兴区的民生话题为背景,数据来源分为以下两个部分:大兴区官方微博和大兴区话题下的文本内容,以及百度贴吧“大兴吧”的文本帖子及其评论内容. 通过爬取 2020 年 12 个月的原始文本数据,获得数据量 6 万多条,并经过数据清洗后使用.

4.2　疫情话题的关键词演化率分析

任何一个民生话题必然会有出现期、增长期、爆发期、蔓延期、衰退期等,本研究以话题漂移的方式来表征话题演化的情况. 对每月采集好的数据绘制 TextRank 和 TF-IDF 值随月份的变化趋势,选取几率最大的与“疫情”相关的词汇. 经过分析,选取“疫情”、“肺炎”、“防控”、“新冠”、“口罩”和“病例”6 个特征词,计算其 TextRank 和 TF-IDF 权重.

接着,统计多源文本中每个月的总文本数和疫情文本数,绘制出热度值随月份的变化趋势,将其作为话题真正的演化趋势. 热度值的变化趋势如图 4 所示. 从图 4 可以看出,疫情话题在大兴区民生话题的关注热度在 2020 年 2 月份达到巅峰,5 月下降,在 8 月到 10 月讨论度降到低谷,证明此时该话题热度消

退. 并在 11 月份话题再次飙升,12 月话题热度又开始降低. 该折线图符合疫情真正的发展趋势,同时符合大兴区民生话题演化的真正趋势. 基于 TextRank 和 TF-IDF 的伪标准差对比图也如图 5 所示. 从图 5 可以看到,TF-IDF 的伪标准差数值普遍低于 TextRank 的值,证明使用 TF-IDF 的来提取关键词的方法. 因此,在后续的 LTEI 指数中,采用基于 TF-IDF 的关键词演化率.

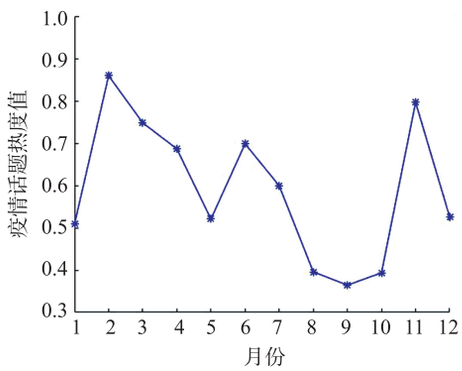

图 4　2020 年每月疫情话题热度值

Fig. 4　Monthly epidemic topic heat value in 2020

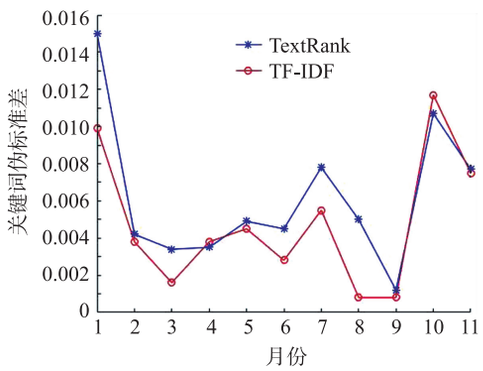

图 5　两种算法的伪标准差

Fig. 5　Pseudo standard deviation of two algorithms

4.3　区域民生话题演化的模型效果对比

根据 HTDI 话题漂移指数的计算方法,使用 LDA 算法将 12 个月的趋势描绘出来,该趋势就是疫情话题的演化情况. 先对 12 个月的疫情话题的总词权重描绘出来,如图 6 所示. 从图 6 中可以看到,7 月份疫情的话题权重最高,其他月份相对较低. 与热度图对比发现,使用该方法分析疫情话题的演化明显不合理,所以更加无法继续使用 HTDI 指数进行接下来的分析.

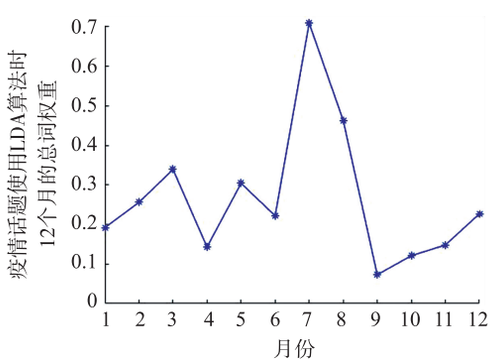

图 6　疫情话题 LDA 的总词权重

Fig. 6　Total word weight of LDA

接下来,对比 HTDI 指数与 LTEI 指数. 经过经验计算,式(9)中的 $\varepsilon=1.5$. 计算结果如表 1 所示.

表 1　HTDI、TF-IDF 总演化率、LTEI 和 r_H

Table 1　HTDI、total evolution rate of TF-IDF、LTEI and r_H

月份序号	HTDI	TF-IDF 关键词总演化率	LTEI	r_H
1	0.002 1	0.002 5	0.002 6+0.000 0i	0.011
2	0.002 9	−0.000 5	−0.000 3+0.000 0i	−0.004
3	−0.003 3	−0.000 8	−0.000 8−0.000 5i	−0.002
4	0.005 4	−0.002 4	−0.002 0+0.000 0i	−0.005
5	−0.002 6	0.001 9	0.001 9−0.000 1i	0.006
6	0.016 2	−0.000 6	0.001 5+0.000 0i	−0.003
7	−0.007 9	−0.001 5	−0.001 5−0.000 7i	−0.007
8	−0.012 6	−0.000 5	−0.000 5−0.001 4i	−0.001
9	0.001 6	0.000 5	0.000 6+0.000 0i	0.001
10	0.000 9	0.001 9	0.001 9+0.000 0i	0.013
11	0.002 6	−0.002 2	−0.002 1+0.000 0i	−0.009

表 1 汇总了疫情话题 12 个月来的 HTDI 指数、TF-IDF 关键词(6 个)总演化率以及 LTEI 指数. 可以看到由于 LTEI 公式内含有 1.5 次方,LTEI 指数计算出来之后得到了复数. 在此省略虚部,将实部的值、HTDI 指数、LTEI 指数和 r_H 值绘制到一个折线图中,观察三个指标的趋势,如图 7 所示.

从图 7 可以看到,HTDI 指数的波动较大,在 6 月~7 月期间值最高,即 6 月~7 月关于疫情的词权重突然升高,话题讨论度大幅度增加. 但对比热度值图,6 月~7 月疫情讨论的热度值实际上是下降的. 而且根据热度图显示,10—11 月话题热度上升幅度较大,但查看 HTDI 曲线可以发现,横坐标为 10 的点,其纵坐标代表的 HTDI 值大于 0 但很接近于 0. 同时对比 r_H 曲线,由于 r_H 是由文本实际数据得出的,可作为基础

对比的指标. 可以看到 LTEI 指数的幅度曲线虽然没有 r_H 大,但从同时间段曲线的升降趋势来看,指数更贴合实际情况,而多数情况下 HTDI 指数的升降与 r_H 相反,这证明使用 HTDI 指数描绘的话题漂移方法并不能很好地应用于多源文本的区域民生的话题演化.

可见,采用 LTEI 指数明显改善了 HTDI 不合理结果,与 TF-IDF 的趋势较吻合,贴合热度值的趋势.

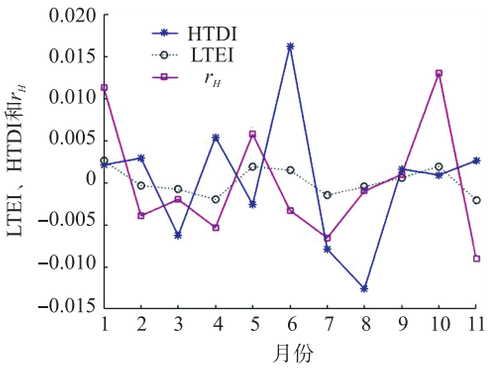

图 7　HTDI、LTEI 和 r_H 的比较

Fig. 7　Comparison of HTDI, LTEI and r_H

5　结论

在分析 LDA 模型困惑度的基础上,针对区域化民生服务需求,提出了民生话题演化技术框架,并融入多源民生数据. 为了探索时间因素影响,创新设计了热度演化率和关键词演化率,阐述了具体的实现算法,综合设计了民生话题演化 LTEI,用于表征多源民生话题的演化趋势. 在实证方面,以大兴区民生话题为例,在线采集了官方微博、网站和百度贴吧数据 6 万余条,涵盖了 2020 年 12 个月的全部文本,并重点分析新冠肺炎疫情防控话题类. 实验结果表明,基于 TF-IDF 的伪标准差数值普遍较低,所以使用 TF-IDF 算法与 HTDI 指数相结合更合理. 同时,所提出的演化指数 LTEI 相较于 HTDI,更能够与实际话题演化趋势相贴合,更适合应用于区域民生话题演化中.

[参考文献]

[1] 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.

[2] 彭敏,官宸宇,朱佳晖. 面向社交媒体文本的话题检测与追踪技术研究综述[J]. 武汉大学学报(理学版),2016,62(3):197-217.

[3] 钱莉,朱恒民,魏静. 话题演化研究综述[J]. 数字图书馆论坛,2021(11):57-64.

[4] 刘怡君,马宁,李倩倩. 非常规突发事件中社会舆论的超网络建模与态势预测[J]. 中国应急管理,2014,(7):14-21.

[5] 唐丽,甄东,李倩. 基于泊松回归模型和注意力配置理论的新冠疫情防控研究[J]. 南京师大学报(自然科学版),2021,44(1):6-12.

[6] BAI Y,JIA S L,CHEN L. Topic evolution analysis of COVID-19 news articles[C]//Journal of physics:Conference Series. New York:ACM,2020:052009.

[7] 龚晓康,应文豪,王骏. 结合 LDA 和孪生 BiLSTM 的话题演化跟踪方法[J]. 中文信息学报,2022,36(2):93-103.

[8] 裴可锋,陈永洲,马静. 基于 DTPM 模型的话题热度预测方法[J]. 情报杂志,2016,35(12):52-57.

[9] 唐晓波,向坤. 基于 LDA 模型和微博热度的热点挖掘[J]. 图书情报工作,2014,58(5):58-63.

[10] 陈兴蜀,高悦,江浩. 基于 OLDA 的热点话题演化跟踪模型[J]. 华南理工大学学报(自然科学版),2016,44(5):130-136.

[11] ZHU J H,LI X H,PENG M,et al. Coherent topic hierarchy:a strategy for topic evolutionary analysis on microblog feeds[J]. Web-age information management. 2015,9098:70-82.

[12] MEI Q,ZHAI C. Discovering evolutionary theme patterns from text-an exploration of temporal text mining[C]//KDD'05,2005.

[13] BLEM D M,LAFFERTY J D. Dynamic topic models[C]//Proceedings of the 23rd International Conference on Machine Learning. New York:ACM,2006:113-120.

[14] WANG X,MCCALLUM A. Topics over time:a non-Markov continuous-time model of topical treads[C]//Proceedings of the 12th ACM SIGKADD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2006:424-433.

[15] PATRICK K,Elaheh Momeni. Optimized tracking of topic evolution[J]. arXiv,2019.

[16] 李纲,陈思菁,毛进,等. 自然灾害事件微博热点话题的时空对比分析[J]. 数据分析与知识发现,2019,3(11):1-15.

[17] 黄微,赵江元,闫璐. 网络热点事件话题漂移指数构建与实证研究[J]. 数据分析与知识发现,2020,4(11):92-101.

[18] 张佩瑶,刘东苏. 基于词向量和 BTM 的短文本话题演化分析[J]. 数据分析与知识发现,2019,3(3):95-101.

[责任编辑:杜忆忱]