

基于概率潜在语义分析的群体情绪演进分析

林江豪^{1*}, 周咏梅^{1,2}, 阳爱民^{1,2}, 陈昱宏¹, 陈晓帆¹

(1. 广东外语外贸大学 思科信息学院, 广州 510006; 2. 广东外语外贸大学 语言工程与计算实验室, 广州 510006)

(* 通信作者电子邮箱 lin_hao@foxmail.com)

摘要:针对群体情绪演进分析中话题内容挖掘及其对应群体情绪分析两个层面的难题,提出了一种基于概率潜在语义分析(PLSA)模型的群体情绪演进分析方法。该方法首先利用 PLSA 模型抽取时间序列上的子话题,挖掘话题内容随时间的演进规律;再利用句法关系和情感本体库,抽取与话题内容相匹配群体情绪单元,计算情绪单元的强度,形成情绪特征向量;最后,对各子话题下的情绪强度进行求和,细粒度分析子话题和事件的整体群体情绪,深入挖掘群体情绪演进规律,并将群体情绪量化和可视化。在话题情绪单元抽取过程中,引入了句法规则和情感本体库,更细粒度地抽取情绪单元,并提高了话题内容与情绪单元匹配的准确性。实验结果表明,该模型能够实现话题内容及其群体情绪按时序特征的演进分析,验证了所提方法的有效性。

关键词:群体情绪;概率潜在语义分析模型;话题挖掘;情绪演进;情绪分析

中图分类号: TP391.1 **文献标志码:** A

Analysis of public emotion evolution based on probabilistic latent semantic analysis

LIN Jianghao^{1*}, ZHOU Yongmei^{1,2}, YANG Aimin^{1,2}, CHEN Yuhong¹, CHEN Xiaofan¹

(1. Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou Guangdong 510006, China;

2. Laboratory for Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou Guangdong 510006, China)

Abstract: Concerning the problem of topics mining and its corresponding public emotion analysis, an analytical method for public emotion evolution was proposed based on Probabilistic Latent Semantic Analysis (PLSA) model. In order to find out the evolutionary patterns of the topics, the method started with extracting the subtopics on time series by making use of PLSA model. Then, emotion feature vectors represented by emotion units and their weights which matched with the topic context were established via parsing and ontology lexicon. Next, the strength of public emotion was computed via a fine-grained dimension and the holistic public emotion of the issue. In this case, the method has a deep mining into the evolutionary patterns of public emotion which were finally quantified and visualized. The advantage of the method is highlighted by introducing grammatical rules and ontology lexicon in the process of extracting emotion units, which was conducted in a fine-grained dimension to improve the accuracy of extraction. The experimental results show that this method can gain good performance on the evolutionary analysis of topics and public emotion on time series and thus proves the positive effect of the method.

Key words: public emotion; Probabilistic Latent Semantic Analysis (PLSA) model; topic mining; emotion evolution; emotion analysis

0 引言

群体情绪是指人们对社会生活各种情境的知觉,通过群体成员之间相互影响、相互作用而形成的较为复杂而又相对稳定的态度体验,这种知觉和体验对个体或群体产生指导性和动力性影响。群体情绪演进分析,主要是挖掘话题内容及其对应的群体情绪按时序特征的演进情况。该工作对网络舆情的监测与引导有重要的意义,同时模型对商业口碑分析等商业应用领域也具有价值。

目前,国内外对群体情绪演进分析研究中,针对话题内容演进分析, TDT (Topic Detection and Tracking) 是最早用于话题

检测与跟踪研究的方法^[1]。采用的关键技术主要有文本聚类的方法^[2-3]、基于概率潜在语义分析 (Probabilistic Latent Semantic Analysis, PLSA)^[4]、潜在狄利克雷分配 (Latent Dirichlet Allocation, LDA)^[5] 等主题模型的方法。如文献[2]基于小波变换提出了改进 K-SC 的算法,能显著降低聚类时间复杂度,并能应用于大量高维热点话题的模式分析;文献[3]提出了概率时空模型 (Probabilistic Spatiotemporal Model, PSM) 用于分析 Twitter 中关于地震事件的事件主体和发生地;文献[4]基于 PLSA 模型提出了上下文概率潜在语义分析 (Contextual PLSA, CPLSA) 方法,通过文本的核心解释资料 (如:作者、出版社和发表时间等) 进行分析,从而判断两个文

收稿日期: 2015-05-27; **修回日期:** 2015-07-05。 **基金项目:** 国家社会科学基金资助项目 (12BYY045); 教育部新世纪优秀人才支持计划项目 (NCET-12-0939); 教育部人文社会科学研究项目 (14YJA740011); 广东省教育厅科技创新项目 (2013KJCX0067); 2015 年广州市哲学社会科学“十二五”规划课题资助项目 (15Q16); 广东外语外贸大学校级项目 (14Q3); 广东外语外贸大学研究生科研创新项目 (14GWCXXM-36)。

作者简介: 林江豪 (1985 -), 男, 广东揭阳人, 助理工程师, 硕士, CCF 会员, 主要研究方向: 自然语言处理、文本情感分析; 周咏梅 (1971 -), 女, 湖南永州人, 教授, CCF 高级会员, 主要研究方向: 文本情感分析、舆情发现; 阳爱民 (1970 -), 男, 湖南永州人, 教授, 博士, CCF 高级会员, 主要研究方向: 文本倾向性分析; 陈昱宏 (1993 -), 男, 广东潮州人, 主要研究方向: 文本情感分析; 陈晓帆 (1993 -), 牙买加人, 男, 主要研究方向: 文本情感分析。

本主题是否相同,对解释资料的依赖性比较强;文献[5]基于LDA模型提出双通道在线主题演化(Bi-Path Evolution Online-LDA, BPE-OLDA)模型,在下一时间片生成文本时考虑文本的内容遗传和强度遗传,很好地模拟了人在生成时效性较强的文本时的特征。这些方法虽能有效分析话题内容的演进情况,但并不能满足群体情绪演进分析中话题及其对应的群体情绪分析。在现有研究中,针对话题和情绪的研究主要是对演进过程的统计特性进行分析,对话题及其群体情绪的分析不够细化。如:文献[6]选取典型事件为研究对象,研究分析网络情绪的演进基本规律;文献[7]提出一种自适应网络舆情演化建模方法(Adaptive Evolution Modeling method of Internet Public Opinions, AEMPO),对网络舆情的平稳性、周期性和自相似性进行演化分析;文献[8]将演化过程划分“触发-集聚-热议-升华”四个关口进行突发事件网络舆情的生成演化规律研究;文献[9]提出用“信息熵”来表示社会情绪的稳定性,同时需要进一步的实例验证。文献[6-9]主要从舆情演进过程的统计特性出发,均缺乏对舆情的内容及其所体现的群体情绪的分析。针对这些不足,文献[10]利用PLSA模型对不同时间段上的网络舆情话题进行子话题抽取和情感词表构建,综合考虑修饰词对情感词的影响以及情感词对子话题的贡献程度,最终得到一个时间序列上各个子话题的情感倾向值以及整个话题的情感变化趋势。但通过扫描情感词表获取情感词,缺乏考虑主观表达以及句法结构对情感词抽取的影响,同时仅从正、负两个维度分析群体的情绪,情绪粒度不够细化,达不到舆情分析的目标。

情感分析(sentiment analysis)是情绪分析的核心工作,主要分析主观性文本中隐藏的私有状态(private state),如意见、态度、情绪等。目前国内外对文本情感分析的研究已逐步成熟,文献[11-12]对文本情感分析关键技术进行了综述,文本情感分析方法主要有基于情感词典^[13]、机器学习^[14]、句法分析^[15]和语义分析^[16]等方法。本文采用基于句法分析和情感词典相结合的方法来实现情绪单元的抽取和群体情绪的量化。

在现有的研究中,学者们已经从话题这一粒度深入挖掘网民关于社会事件的情绪演进规律,通过在时间序列上抽取子话题,量子化话题的正负向情感强度来分析事件的发展与情绪的演进,提升了群体情绪演进分析的效能。然而,现有文献只分析了情绪的正负向,没有深层次地挖掘话题的情绪,缺乏对话题情绪的细粒度分析,无法达到舆情监控的要求。本文提出一种基于PLSA的群体情绪演进分析方法,该方法通过获取话题随时间的演化规律,结合句法分析和情感本体库,深入挖掘话题的群体情绪,实现了群体情绪演进的细粒度分析。

1 基于PLSA的群体情绪演进分析方法

本文采用如图1所示的群体情绪演进分析方法。首先将采集的新闻评论集进行文本预处理,分词后过滤掉停用词和无用词,进行词频统计,获得“文档-词”矩阵;接着,利用PLSA模型抽取贡献度较高的词汇,构建子话题词表,结合情感本体库和句法规则集提取群体情绪特征向量;最后计算子话题群体情绪,对子话题群体情绪加总得到整体群体情绪。

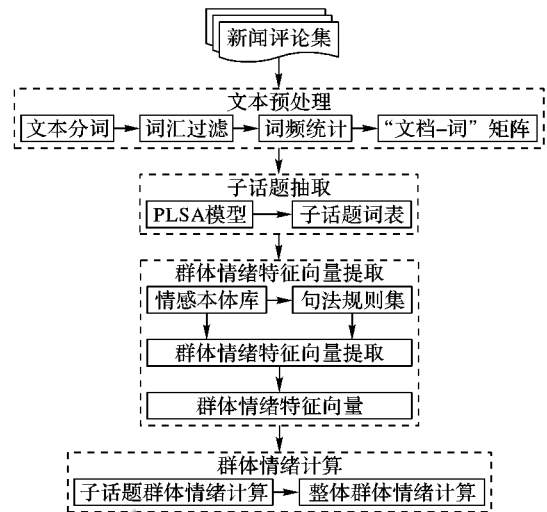


图1 群体情绪演进分析过程

用二元组 (Z_t^k, E_t^k) 表示 t 时刻群体情绪演进情况,其中:
 Z_t^k 表示 t 时刻共抽取到 k 个子话题,话题采用关键词表表示,即话题集 $Z_t^k = \{[w_{d1}, w_{d2}, \dots, w_{dw}]^1, [w_{d1}, w_{d2}, \dots, w_{dw}]^2, \dots, [w_{d1}, w_{d2}, \dots, w_{dw}]^k\}$;情绪 E_t^k 表示 t 时刻 k 个子话题对应的情绪向量,情绪维度为 n ,即有 $E_t^k = \{[e_{d1}, e_{d2}, \dots, e_{dn}]^1, [e_{d1}, e_{d2}, \dots, e_{dn}]^2, \dots, [e_{d1}, e_{d2}, \dots, e_{dn}]^k\}$ 。因此, t 时刻模型的输出为 k 个 $\{[w_{d1}, w_{d2}, \dots, w_{dw}], [e_{d1}, e_{d2}, \dots, e_{dn}]\}$,当 $\|E_t^j\| = 0 (j \leq k)$ 时,表示子话题 j 为中性。则基于PLSA的群体情绪演进算法如下。

算法1 基于PLSA的群体情绪演进算法。

输入 t 时刻的语料集 $Data_set_t$ 。

输出 (Z_t^k, E_t^k) 。

第1步 对 $Data_set_t$ 进行预处理,包括分词、词频统计等,获得“文档-词频”矩阵 M_t 。

第2步 计算 $PLSA(M_t) \rightarrow$ “主题-词语”矩阵 $[M_{word-topic}]_{m \times k}$ 。

第3步 逐列对 $[M_{word-topic}]_{m \times k}$ 进行排序,获取每个子话题 Z_t^k 中概率较高的关键词,得到 $Z_t^k = \{[w_{d1}, w_{d2}, \dots, w_{dw}]^1, [w_{d1}, w_{d2}, \dots, w_{dw}]^2, \dots, [w_{d1}, w_{d2}, \dots, w_{dw}]^k\}$ 。

第4步 利用句法关系,基于情感本体库,考虑否定词及程度副词的作用,抽取修饰 Z_t^k 的情绪单元 eu ,并计算 eu 的权重 uw ,形成情绪特征向量 V_t^k 。

第5步 对 V_t^k 进行量化计算,得到 E_t^k 。

结束 输出 (Z_t^k, E_t^k) 。

模型的输出二元组 (Z_t^k, E_t^k) 中, E_t^k 用于话题的群体情绪分析, Z_t^k 可用于子话题抽取,本文采用基于PLSA的话题抽取方法。

2 基于PLSA的话题抽取方法

2.1 PLSA模型

PLSA模型是由Hofmann在1999年提出的,首先给定文档集 $D = \{d_1, d_2, \dots, d_n\}$ 和词集 $W = \{w_1, w_2, \dots, w_m\}$,用 $freq(d_i, w_j)$ 表示词 w_j 在文档 d_i 中出现的概率,则“文档-词语”共现矩阵 $M_{D-W} = [freq(d_i, w_j)]$ 。假设主题类别 $Z = \{z_1, z_2, \dots, z_k\}$, k 为主题个数。PLSA模型假设词与文档之间、话题

与文档或者词之间的概率服从条件独立,由此得到相应的联合分布概率为:

$$P(d_i, z_k, w_j) = P(d_i)P(z_k | d_i)P(w_j | z_k) \quad (1)$$

其中: $P(d_i)$ 表示选择文档 d_i 的概率, $P(z_k | d_i)$ 表示某个主题 z_k 在给定文档 d_i 下出现的概率, $P(w_j | z_k)$ 表示词 w_j 在给定主题 z_k 下出现的概率。本文基于该“词语-主题”的概率分布获取事件 Ev_i , 根据贝叶斯法则可得到:

$$P(d_i, w_j) = P(d_i) \sum_{k=1}^K P(z_k | d_i) P(w_j | z_k) \quad (2)$$

采用最大期望 (Expectation Maximization, EM) 算法对潜在语义模型进行拟合^[17]。用随机数初始化之后, 交替执行 Expectation 步骤和 Maximization 步骤进行迭代计算。Expectation 步骤计算 (d_i, w_j) 所产生的潜在语义 z_k 的先验概率为:

$$P(z_k | d_i, w_j) = \frac{P(z_k | d_i)P(w_j | z_k)}{\sum_{i=1}^K P(z_i | d_i)P(w_j | z_i)} \quad (3)$$

在 Maximization 步骤中, 根据 $P(z | d, w)$ 对 $P(w | z)$ 和 $P(z | d)$ 矩阵重新估计:

$$P(w_j | d_i) = \frac{\sum_{k=1}^K \text{freq}(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{m=1}^M \sum_{i=1}^N \text{freq}(d_i, w_m) P(z_k | d_i, w_m)} \quad (4)$$

$$P(z_k | d_i) = \frac{\sum_{j=1}^M \text{freq}(d_i, w_j) P(z_k | d_i, w_j)}{\text{freq}(d_i)} \quad (5)$$

似然函数的对数如式(6)所示:

$$L = \sum_{i=1}^N \sum_{j=1}^M \text{freq}(d_i, w_j) \log(P(d_i, w_j)) \propto \sum_{i=1}^N \sum_{j=1}^M \text{freq}(d_i, w_j) \log\left(\sum_{k=1}^K P(z_k | d_i) P(w_j | z_k)\right) \quad (6)$$

当似然函数 L 期望值的增加量小于阈值时, 迭代终止。此时得到一个最优解 $P(w | z) = [P(w_j | z_k)]_{m \times k}$ 和 $P(z | d) = [P(z_k | d_i)]_{k \times n}$ 。

2.2 子话题抽取

子话题可由一组语义上相关的词及与话题相关的权重的向量表示^[18]。则 t 时刻的话题集 $Z_t^k = \{[(w_{i1}, P(w_{i1} | z_t^j)), (w_{i2}, P(w_{i2} | z_t^j)), \dots, (w_{im}, P(w_{im} | z_t^j))]^j\}$, $j \in [1, k]$ 。

首先, $P(w_{ii} | z_t^j) = 0$ 表示 w_{ii} 对 t 时刻的主题 z_t^j 没有贡献度, 直接过滤掉; 然后将 $P(w_{ii} | z_t^j)$ 由大到小排序, 提取 $P(w_{ii} | z_t^j) \geq \alpha$ 的词 w_{ii} 作为话题词表; 最终, 获得子话题内容为话题词表和词语在话题中的概率。

3 群体情绪量化

情感词作为情绪的载体, 情感词典是情绪识别与分析的重要工具。情感本体库 OL^[19] 共含有情感词 27 466 个, 分 7 类情绪: 乐、好、怒、哀、惧、恶、惊; 强度分为 1, 3, 5, 7, 9 五档, 9 表示强度最强, 1 为强度最弱。因此, 本文利用情感本体库抽取话题集 Z_t^k 的情绪特征向量。

由于评论具有内容短、评价对象多样化等特点, 直接扫描评论中出现的情感词来获取情感词表, 容易造成情感词与话

题 Z_t^k 不对称; 另外, 否定副词、程度副词的出现, 都对情绪分析的准确性造成影响, 也是情绪分析的难题。利用词语之间的句法关系, 可有效解决该难题。因此, 本文利用句法分析工具^[20], 采用以下算法获取情绪特征向量。

算法 2 情绪特征向量获取。

输入 话题集 Z_t^k , 情感本体库 OL, 否定副词表 NWL, 程度副词表 DAL。

输出 情绪特征向量 V_t^k 。

第 1 步 抽取修饰 Z_t^k 的情感词 $ew (ew \in OL)$, 同时获得 ew 的强度 wt 和情绪类别 C 。其中: 强度 $wt \in \{1, 3, 5, 7, 9\}$, 类别 $C \in \{\text{乐, 好, 怒, 哀, 惧, 恶, 惊}\}$ 。

第 2 步 参考文献[21] 的模板, 获取并计算每个情感单元 eu 的情感权重 uw 。

第 3 步 获得子话题 z_t^j 的 r 个情感特征, 则特征向量可表示为 $v_t^j = \{(eu_1, uw_1, C), (eu_2, uw_2, C), \dots, (eu_r, uw_r, C)\}$ 。

结束 输出 t 时刻话题集 Z_t^k 的情绪特征向量 $V_t^k = \{v_t^1, v_t^2, \dots, v_t^k\}$ 。

根据情绪特征向量 V_t^k , 对 t 时刻的群体情绪进行量化计算。则 t 时刻子话题 z_t^j 的群体情绪类别“乐”的情绪值采用式(7)来进行量化, 其他情绪类别的计算方法相同。

$$e_t^j (C = \text{乐}) = \sum_{i=1}^r uw_i; C = \text{乐} \quad (7)$$

则 t 时刻话题集 Z_t^k 的群体情绪类别“乐”可采用式(8)来计算, 其他类别的计算方法相同。

$$E_t (C = \text{乐}) = \sum_{j=1}^k e_t^j (C = \text{乐}); C = \text{乐} \quad (8)$$

4 实验结果及分析

4.1 实验数据采集

实验采集了凤凰网 2014 年 8 月 22 日新闻“湖南: 逼医生向尸体下跪事件引静坐抗议”的评论, 共 5 871 条。如图 2 为 22 日 6:00—9:00 的评论数, 曲线呈现先增后减, 逐步趋向平缓的走向, 与网络舆情的发展规律趋势相符。本文截取 22 日 7:00—11:00 这段时间的评论作为群体情绪演进分析的对象: 一方面在这 4 h 内产生了 4 464 条评论, 占评论总数的 76%; 另一方面, 每个时间段内的评论数量均大于 500 条, 群体情绪特征比较明显。

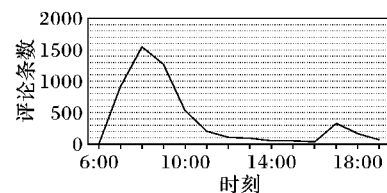


图2 “评论数-时间”分布图

4.2 PLSA 子话题抽取

按时间顺序, 设置时间间隔为 1 h, 将评论分为数据集 DS7、DS8、DS9、DS10。对数据进行分词处理, 去掉停用词和无用词, 统计词频得到“文档-词汇”的共现矩阵 $M7$ 、 $M8$ 、 $M9$ 、 $M10$ 。将矩阵输入到 PLSA, 得到每小时的子话题, 以及概率矩阵 $[P(w | z)]_{m \times k}$ 和 $[P(z | d)]_{k \times n}$ 。逐列排序矩阵 $[P(w | z)]_{m \times k}$, 取概率大于 0.8 的词作为话题词, 部分话题词如表 1 所示。

表 1 PLSA 子话题抽取结果示例

语料	子话题	评论条数	(话题词,概率)
DS7	责备家属无理	635	(手术,0.9306),(死者,0.9291)(医闹,0.9143),(家属,0.9139),(尊重,0.8792),(法律,0.8734),(生命,0.8521),(患者,0.8448),(医疗,0.8253),(检查,0.8157)
	医生态度恶劣	476	(素质,0.9416),(亲属,0.9365),(态度,0.9315),(流氓,0.9162),(犯罪,0.8954),(官僚,0.8935),(服务,0.8695),(亲人,0.8408),(农民,0.8181),(领导,0.8161)
DS8	责备家属无理	943	(维权,0.9225),(尊严,0.9159),(合法,0.9148),(死者,0.9144),(医术,0.9010),(急诊科,0.8837),(公正,0.8638),(家属,0.8360),(手术,0.8323),(闹事,0.8199)
	乱开药,乱收费	578	(医药费,0.9494),(医保,0.9382),(药店,0.9136),(利益,0.9051),(吸血鬼,0.894),(药品,0.8415),(费用,0.8356),(金钱,0.8284),(卫生局,0.8153),(试验品,0.808)
	医德败坏	735	(风气,0.9391),(摇钱树,0.9383),(天使,0.9287),(责任心,0.9179),(良心,0.8880),(形象,0.8869),(儿子,0.8746),(白大褂,0.8344),(风气,0.8221),(责任,0.8105)
	医生态度恶劣	623	(孩子,0.9478),(教育,0.9440),(心态,0.9263),(领导,0.9171),(服务,0.9006),(眼光,0.8869),(仁心,0.8593),(情绪,0.8591),(群众,0.8476),(素质,0.8350)
DS9	责备家属无理	813	(家人,0.9100),(合法,0.9008),(政府,0.8967),(手术台,0.8730),(闹事,0.8705),(急诊科,0.8681),(死者,0.8445),(凶手,0.8440),(公正,0.8337),(闹事者,0.8216)
	医德败坏	679	(责任心,0.9394),(天使,0.9327),(白大褂,0.9243),(风气,0.9164),(儿子,0.9164),(高利贷,0.9009),(摇钱树,0.8937),(形象,0.8781),(急救,0.8629),(医德,0.8306)
DS10	责备家属无理	327	(医闹,0.9273),(检查,0.9268),(手术,0.9132),(生命,0.9064),(医疗,0.8760),(法律,0.8657),(死者,0.8584),(政府,0.8535),(医护,0.8453),(尊重,0.8388)
	医生态度恶劣	258	(反思,0.9386),(孩子,0.9243),(服务,0.9235),(流氓,0.9085),(人渣,0.8984),(犯罪,0.8734),(病患,0.8624),(素质,0.8539),(官僚,0.8457),(心态,0.8408)
	医德败坏	336	(形象,0.9310),(天使,0.9297),(高利贷,0.9296),(医德,0.9245),(急救,0.9131),(红包,0.8896),(风气,0.8730),(责任心,0.8688),(摇钱树,0.8283),(职业道德,0.8225)

由表 1 可见,子话题“责备家属无理”,在话题演进过程中,都有较强的体现,也是新闻的核心主题。其他子话题“医德败坏”“乱开药,乱收费”“医生态度恶劣”在演进过程中均有所体现。为了验证话题抽取的准确性,以“责备家属无理”“医德败坏”“乱开药,乱收费”“医生态度恶劣”4 个子话题为标注目标,如果语料无法归类到 4 个子话题则过滤掉。由于一条语料可能涉及多个主题,因此同一条语料允许标注为多个话题。同时,为了验证情绪演进分析的有效性,还标注了语料对应话题的情绪类别,每一条在某一话题下的情绪类别只能标注为乐、好、怒、哀、惧、恶、惊中的一种,如果没有情绪则标注为中性。通过统计语料集 DS7、DS8、DS9、DS10 中在 4 个子话题中的评论数发现,结果如表 1 所示的评论数,系统自动抽取的结果与实际的语料分布相符合,说明了这种话题抽取方法是有效的。

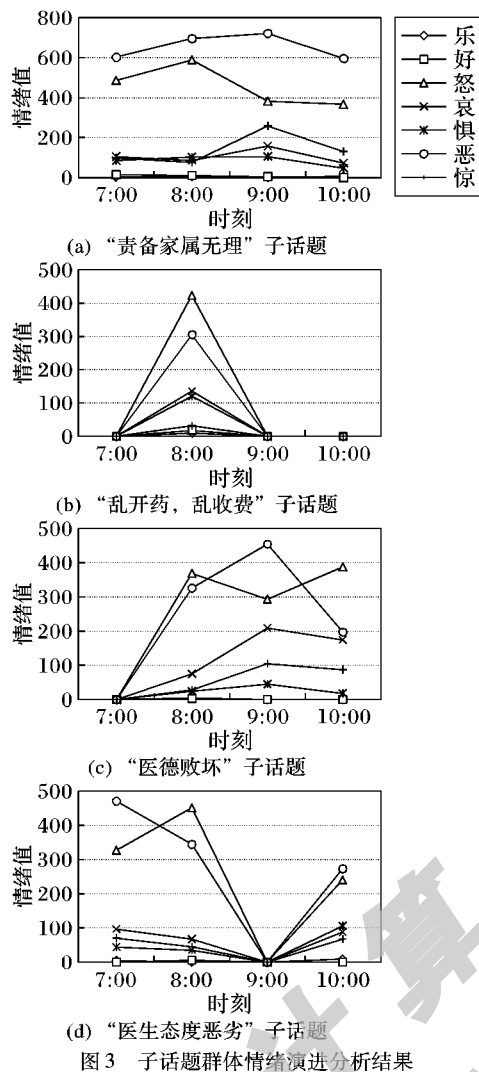
表 2 子话题“责备家属无理”的部分情绪单元

语料	情绪单元(情绪单元,情绪权重,情绪类别)
DS7	(民愤,5,怒),(真 TM,5,怒),(MD,7,怒),(完全无理,6,恶),(哇靠,5,怒),(非常蛮横,30,恶),(威胁,5,恶),(霸道,5,恶),(极嚣张,42,恶),(横蛮无理,6,恶)
DS8	(超不平,30,怒),(百分之百活该,18,怒),(惩办,7,怒),(极其激愤,42,怒),(嚣张,7,恶),(蛮横,5,恶),(绝对横蛮无理,42,恶),(完全无理,6,恶),(太混蛋,28,恶),(强词夺理,5,恶)
DS9	(真 TM,5,怒),(MD,7,怒),(你丫,5,怒),(土匪,5,恶),(强词夺理,5,恶),(混账,5,恶),(扯皮,3,恶),(莽撞,5,恶),(偏激,5,恶),(绝对败类,42,恶)
DS10	(横蛮无理,6,恶),(TNND,5,怒),(实在窝火,20,怒),(kao,5,怒),(忿怒,5,怒),(狗屁,7,恶),(扯淡,3,恶),(犯罪,5,恶),(十分流氓,18,恶),(十分恶劣,42,恶)

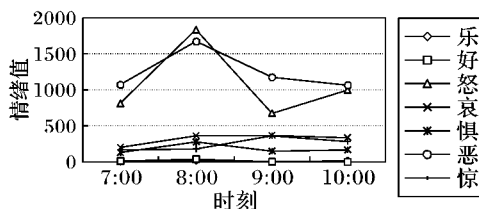
4.3 群体情绪演进分析

根据第 3 章中的群体情绪量化方法,对 22 日 7:00—11:00 的子话题进行量化计算,结果如图 3 所示,显示了 4 个子话题对应的群体情绪随着时间序列的变化,y 轴为情绪值,通过式(7)对时间窗口内各情绪类别的情绪单元的情绪权重进行累加获得。

从图 3 可看出群体对事件的情绪主体为“怒”与“恶”。观察语料,发现图 3(a)的情绪对象主要为家属的无理要求,图 3(b)~(d)主要的情绪对象集中在医院和医生。从时间角度来看,7:00—8:00 网友在“责备家属无理”的同时,也表示是否因为“医生的态度恶劣”造成了家属的医闹,并且持“怒”与“恶”的情绪;8:00—9:00 是评论的高峰,产生新的子话题,在谴责家属的同时,对医院和医生“收费”与“医德”也进行了评论,主体情绪为“怒”与“恶”;9:00—11:00,子话题数由 4 个减少到 2 个再演进为 3 个,评价对象主要围绕医生和家属,情绪重点在“怒”与“恶”。由于没有对每条语料进行情绪分类,因此,本文统计了语料的情绪标注结果,发现子话题下各类情绪的语料数量分布曲线与图 3 中的演进情况接近,说明了情绪演进的规律与实际情况相符合。观察语料发现,多数评论是网民自己或带家人到医院看病经历,如“上次带小孩去省医,随便看个感冒,都要几百块,医生太黑了,现在的药费真的伤不起”。也有周边朋友的经历,如“同事小小的感冒,医生要求拍这个片那个片的,一下子千把块就没了,现在的医生,眼里只有钱,一点医德都没有”。这些评论都跟网民的生活经历有关,网民的经历形成了“医患类”新闻事件的情绪认知,一旦看到相关新闻,立即产生刺激,并通过网络表达自己的情绪。这种“刺激-反射”的现象有较强的不确定性,是群体情绪演进分析的难点,同时也是舆情演进分析的核心。



基于式(8)计算事件的整体情绪演进情况结果如图4所示, 通过将整个事件在时间窗口内各情绪类别的情绪单元的情绪权重进行累加获得。对整个话题的群体情绪主要集中在“怒”与“恶”, 而且波动比较大; “惊”“哀”和“惧”三种情绪也有一定的表现, 情绪相对比较稳定; 而“乐”与“好”在整个事件的发展过程基本上没有体现。分析结果发现, 对于这种主题分明、是非易辨的主题, 群体情绪的表达更趋向于直接表达, 较少使用反语或者讽刺的表达方式, 有利于情绪的有效分析。



通过对子话题和情绪的演进分析及可视化展示, 可以观察到整个话题内容随着时间推移产生的演变, 及其对应情绪的演进情况。从更细粒度考察了话题的群体情绪, 实现群体情绪的演进规律分析, 可更有针对性地实施网络舆情的监控与引导。

5 结语

群体情绪演进分析是舆情管理的重要工作, 本文提出一种基于PLSA的群体情绪演进分析方法。实验结果表明, 这种方法能对“话题-情绪”在时间序列上的演进过程进行分析, 并从更细粒度分析群体的情绪; 同时考虑了句法关系对评价单元抽取的影响, 提升了评价单元抽取准确性, 能更好匹配评价对象与评价单元之间的对称关系, 实现了针对子话题的群体情绪分析, 很好地表达了话题内容及其群体情绪的演进规律。下一步的工作是分析情绪认知对情绪演进的影响, 通过抽取群体的情绪认知, 预测话题的可能发展方向, 达到舆情的预测效果。

参考文献:

- [1] CHEN H, WANG F, ZENG D. Intelligence and security informatics for homeland security: information, communication, and transportation[J]. IEEE Transactions on Intelligent Transportation Systems, 2004, 5(4): 329-341.
- [2] HAN Z, CHEN N, LE J, et al. An efficient and effective clustering algorithm for time series of hot topics[J]. Chinese Journal of Computers, 2012, 35(11): 2337-2347. (韩忠明, 陈妮, 乐嘉锦, 等. 面向热点话题时间序列的有效聚类算法研究[J]. 计算机学报, 2012, 35(11): 2337-2347.)
- [3] SAKAKI T, OKAZAKI M, MATSUO Y. Earthquake shakes Twitter users: real-time event detection by social sensors[C]// Proceedings of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 851-860.
- [4] XING E P, YAN R, HAUPTMANN A G. Mining associated text and images with dual-wing harmoniums[EB/OL]. [2014-10-10]. <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1968&context=compsci>.
- [5] CAO J, WANG H, XIA Y, et al. Bi-path evolution model for online topic model based on LDA[J]. Acta Automatica Sinica, 2014, 40(12): 2877-2886. (曹建平, 王晖, 夏友清, 等. 基于LDA的双通道在线主题演化模型[J]. 自动化学报, 2014, 40(12): 2877-2886.)
- [6] TANG C. Empirical research on the evolution of online emotion[J]. Journal of Intelligence, 2012(10): 48-52. (唐超. 网络情绪演进的实证研究[J]. 情报杂志, 2012, 31(10): 48-52.)
- [7] ZHOU Y, LI B. Adaptive evolution modeling method of Internet public opinions[J]. Journal of Data Acquisition and Processing, 2013, 28(1): 69-76. (周耀明, 李弼程. 一种自适应网络舆情演化建模方法[J]. 数据采集与处理, 2013, 28(1): 69-76.)
- [8] YI C, HE Z. Study on the regularities of formation and evolution of Internet public opinions caused by emergent events[J]. Journal of Xiangtan University: Philosophy and Social Sciences, 2014, 38(2): 74-78. (易臣何, 何振. 突发事件网络舆情的生成演化规律研究[J]. 湘潭大学学报: 哲学社会科学版, 2014, 38(2): 74-78.)
- [9] LI C, HONG Y. Modeling method for social emotional stability facing emergency[J]. Journal of Intelligence, 2014, 33(1): 146-151. (李从东, 洪宇翔. 面向突发事件的社会情绪稳定性建模方法研究[J]. 情报杂志, 2014, 33(1): 146-151.)
- [10] HUANG W, CHEN L, WU M. Research on sentiment evaluation of online public opinion topic[J]. Journal of Intelligence, 2014, 33(1): 102-107. (黄卫东, 陈凌云, 吴美蓉. 网络舆情话题情感演化研究[J]. 情报杂志, 2014, 33(1): 102-107.)

(下转第2756页)

证提出的方法,并将其应用到医学图像分析中。

参考文献:

- [1] GUYON I, ELISSEEFF A. An introduction to variable and feature selection [J]. *Journal of Machine Learning Research*, 2003, 3: 1157–1182.
- [2] HE X, CAI D, NIYOGI P. Laplacian score for feature selection [EB/OL]. [2014-10-10]. <http://people.cs.uchicago.edu/~niyogi/papersps/HeCaiNiyolapscore.pdf>.
- [3] YU L, LIU H. Feature selection for high-dimensional data: a fast correlation-based filter solution [EB/OL]. [2014-10-10]. <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.68.2975>.
- [4] WESTON J, GUYON I. Support vector machine-recursive feature elimination (SVM-RFE): US, US8095483 B2[P]. 2010.
- [5] TIBSHIRANI R. Regression shrinkage and selection via the LASSO: a retrospective [J]. *Journal of the Royal Statistical Society*, 2011, 73(3): 273–282.
- [6] PUDIL P, NOVOCIOVA J, KITTLER J. Floating search methods in feature selection [J]. *Pattern Recognition Letters*, 1994, 15(11): 1119–1125.
- [7] NG A Y. Feature selection, L_1 vs. L_2 regularization, and rotational invariance [J]. *International Conferences on Machine Learning*, 2004, 19(5): 379–387.
- [8] ZHOU J, LU Z, SUN J, *et al.* FeaFiner: biomarker identification from medical data through feature generalization and selection [C]// *KDD 2013: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM Press, 2013: 1034–1042.
- [9] LIU F, WEE C Y, CHEN H. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification [J]. *NeuroImage*, 2014, 84: 466–475.
- [10] ZOU H, HASTIE T. Regularization and variable selection via the elastic net [J]. *Journal of the Royal Statistical Society*, 2005, 67(2): 301–320.
- [11] TIBSHIRANI R, SAUNDERS M, ROSSET S. Sparsity and smoothness via the fused lasso [J]. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 2005, 67(1): 91–108.
- [12] YE G, XIE X. Split Bregman method for large scale fused lasso [J]. *Computational Statistics & Data Analysis*, 2011, 55(4): 1552–1569.
- [13] YAMADA M, JITKRITTUM W, SIGAL L. High-dimensional feature selection by feature-wise kernelized lasso [J]. *Neural Computation*, 2014, 26(1): 185–207.
- [14] CHEN X, PAN W, KWOK J T, *et al.* Accelerated gradient method for multi-task sparse learning problem [C]// *Proceedings of the Ninth IEEE International Conference on Data Mining*. Piscataway: IEEE Press, 2009: 746–751.
- [15] LIU J, YE J. Efficient L_1/L_q norm regularization [R]. Arizona: Arizona State University, 2009.
- [16] CAI D, HE X, ZHOU K. Locality sensitive discriminant analysis [C]// *Proceedings of the 2007 International Joint Conference on Artificial Intelligence*. [S. l.]: Morgan Kaufmann Press, 2007: 708–713.
- [17] XUE H, CHEN S, YANG Q. Discriminatively regularized least-squares classification [J]. *Pattern Recognition*, 2009, 42(1): 93–104.
- [11] ZHAO Y, QIN B, LIU T. Sentiment analysis [J]. *Journal of Software*, 2010, 21(8): 1834–1848. (赵妍妍, 秦兵, 刘挺. 文本情感分析 [J]. *软件学报*, 2010, 21(8): 1834–1848.)
- [12] YANG L, ZHU J, TANG S. Survey of text sentiment analysis [J]. *Journal of Computer Applications*, 2013, 33(6): 1574–1578. (杨立公, 朱俭, 汤世平. 文本情感分析综述 [J]. *计算机应用*, 2013, 33(6): 1574–1578.)
- [13] YANG A, LIN J, ZHOU Y, *et al.* Research on building a Chinese sentiment lexicon based on SO-PMI [J]. *Applied Mechanics and Materials*, 2013, 263/264/265/266: 1688–1693.
- [14] YANG A, ZHOU Y, LIN J. A method of Chinese texts sentiment classification based on Bayesian algorithm [J]. *Applied Mechanics and Materials*, 2013, 263/264/265/266: 2185–2190.
- [15] LU H, NIU Z, ZHANG N, *et al.* A model for sentiment classification of Chinese microblog based on parsing and theme extension [J]. *Transactions of Beijing Institute of Technology*, 2014, 34(8): 824–829. (陆浩, 牛振东, 张楠, 等. 基于句法与主题扩展的中文微博情感倾向性分析模型 [J]. *北京理工大学学报*, 2014, 34(8): 824–829.)
- [16] YANG J, YANG A, ZHOU Y. Sentiment classification method of Chinese micro-blog based on semantic analysis [J]. *Journal of Shandong University: Natural Science*, 2014(11): 1671–9352. (杨佳能, 阳爱民, 周咏梅. 基于语义分析的中文微博情感分类方法 [J]. *山东大学学报: 理学版*, 2014(11): 1671–9352.)
- [17] JIN X, ZHOU Y, MOBASHER B. A unified approach to personalization based on probabilistic latent semantic models of Web usage and content [C]// *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization*. Menlo Park: AAAI Press, 2004: 654–658.
- [18] CHU K, LI F. LDA model-based news topic evolution [J]. *Computer Applications and Software*, 2011, 28(4): 4–7, 26. (楚克明, 李芳. 基于 LDA 模型的新闻话题的演化 [J]. *计算机应用与软件*, 2011, 28(4): 4–7, 26.)
- [19] XU L, LIN H, PAN Y, *et al.* Constructing the affective lexicon ontology [J]. *Journal of the China Society for Scientific and Technical Information*, 2008, 27(2): 180–185. (徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造 [J]. *情报学报*, 2008, 27(2): 180–185.)
- [20] CHE W, LI Z, LIU T. LTP: a Chinese language technology platform [EB/OL]. [2014-10-10]. <http://ir.hit.edu.cn/~car/papers/coling10demo.pdf>.
- [21] WU J, TANG C, LI T, *et al.* Sentiment analysis on Web financial text based on semantic rules [J]. *Journal of Computer Applications*, 2014, 34(2): 481–485, 495. (吴江, 唐常杰, 李太勇, 等. 基于语义规则的 Web 金融文本情感分析 [J]. *计算机应用*, 2014, 34(2): 481–485, 495.)

(上接第 2751 页)