中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

跨模LDA融合的多模态数据主题分析方法

赵越, 郝琨, 时彩云, 解胜震, 王之琼, 信俊昌

引用本文:
赵越,郝琨,时彩云,解胜震,王之琼,信俊昌. 跨模LDA融合的多模态数据主题分析方法[J]. 控制与决策, 2024, 39(4): 1325−1332.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1277

跨模LDA融合的多模态数据主题分析方法

赵 越[1], 郝 琨[1], 时彩云[2], 解胜震[2], 王之琼[1], 信俊昌[2†]

(1. 东北大学 医学与生物信息工程学院，沈阳 110169；2. 东北大学 计算机科学与工程学院，沈阳 110169)

摘 要: 随着互联网的高速发展,社会大众可以通过网络对医疗事件以及医患关系自由地发表个人意见和观点言论,这对于引导公众正确的价值导向有着重大研究意义. 然而,仅考虑单模态数据的主题分析算法不能精准地把握整个舆情事件的真相,存在主题提取不准确、个人情感先入为主等问题. 提出一种基于LDA的多模态数据主题分析算法MD_LDA(multimodal data topic analysis based on LDA). 通过对各模态主题分析结果进行决策级融合来计算多模态的主题分析结果,进而解决传统方法对多模态数据考虑不全面的缺陷. 实验结果表明,针对多模态舆情事件,在主题词的提取效果上,所提出的MD_LDA算法优于单一模态数据进行主题分析的算法. 而相对于传统的关键词提取算法TF_IDF与TextRank和MD_LDA算法的准确率以及主题词提取效率均有所提高,验证了结合多模态数据进行主题分析的MD_LDA算法的有效性.

关键词: 主题分析；多模态；LDA主题模型；网络舆情

中图分类号: TP311 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1277

引用格式: 赵越,郝琨,时彩云,等. 跨模LDA融合的多模态数据主题分析方法[J]. 控制与决策, 2024, 39(4): 1325-1332.

Multimodal data topic analysis method based on cross-modal LDA fusion

ZHAO Yue[1], *HAO Kun*[1], *SHI Cai-yun*[2], *XIE Sheng-zhen*[2], *WANG Zhi-qiong*[1], *XIN Jun-chang*[2†]

(1. College of Medicine and Biological Information Engineering，Northeastern University，Shenyang 110169，China；2. School of Computer Science and Engineering，Northeastern University，Shenyang 110169，China)

Abstract: With the rapid development of the Internet, the public can freely express personal opinions on medical events and doctor-patient relationships through the Internet, which are of the correct value for guiding the public. Orientation has great research significance. However, the topic analysis algorithm that only considers single-modal data cannot accurately grasp the truth of the entire public opinion event, and there are problems such as inaccurate topic extraction and preconceived personal emotions. To solve this problem, this paper proposes a LDA-based multimodal data topic analysis algorithm, named MD_LDA (multimodal data topic analysis Based on LDA). The multimodal topic analysis is calculated by the decision-level fusion of the results of each modal topic analysis. As a result, it further solves the defect that traditional methods do not fully consider multimodal data. The experimental results show that for multimodal public opinion events, the proposed MD_LDA algorithm is better than the algorithm for topic analysis of single-modal data in terms of the extraction effect of topic words. Compared with the traditional keyword extraction algorithms TF_IDF and TextRank, the accuracy of the MD_LDA algorithm and the extraction efficiency of subject words are improved, which proves the effectiveness of the MD_LDA algorithm for subject analysis combined with multimodal data.

Keywords: topic analysis；multimodality；LDA topic model；network public opinion

0 引 言

随着互联网的快速发展,社会大众可以通过网络发布对医疗事件以及医患关系的个人意见和观点言论. 舆情分析广泛应用于医疗、体育、政要,尤其是在医疗领域的舆情分析更具深远意义. 然而,仅考虑单模态数据的主题分析算法不能精准地把握整个舆情事件的真相,比如一篇关于医患关系的文字报道描述的是医生的不当之处,这时候对于读者而言更多的选择斥责医生的行为,但多方面考虑到监控、录音、照片这些其他模态数据时,可能会发现另有真相. 因此,在

收稿日期: 2022-07-18；**录用日期:** 2022-12-30.

基金项目: 国家自然科学基金项目(62072089)；中央高校基本科研业务费专项资金项目(N2116016, N2104001, N2019007)；东软集团股份有限公司开放课题项目(NCBETOP2102).

责任编委: 胡清华.

†通讯作者. E-mail: xinjunchang@mail.neu.edu.cn.

考虑多模态数据的情况下,多方面综合处理舆情数据更能准确地还原事件的真相.

潜在语义分析LSA模型[1-2]对奇异值进行分解,可以消除处于同义词和多义词之间的复杂信息. Hofmann提出了概率潜在语义PLSA主题模型[3-4],有效解决了同义词与多义词相混合的问题. 文献[5]提出了LDA主题模型,通过引入一层Dirichlet的先验分布对PLSA模型进行了扩展. 文献[6]基于LDA模型设计出Twitter-LDA模型,该模型同时在用户和博文层面进行主题建模,提高了博文层面的数据分析能力.

文献[7]提出了多模态LDA模型,可以有效地表示视觉特征与文本特征之间的相关度. 文献[8]提出了多模态主题MMLDA模型,分别对位于多模态事件数据集中的文本词汇和视觉词汇进行计算. 文献[9]提出了文本建模方法,该方法捕捉到SMNN文本特征和CNN视觉特征之间的互补效应,从而改进图像分类效果. 文献[10]提出了一种基于语义的视频分类方法. 文献[11]设计了一种结构化的主题表示方法,使用基于SWC的聚类抽样方法检测主题.

本文主要工作如下:

1)根据各模态自身的特征实现了相应的图像主题分析算法、音频主题分析算法和视频主题分析算法;

2)实现了基于LDA的多模态数据主题分析算法MD_LDA,将各模态数据融入舆情主题分析过程中,通过决策层融合策略实现各模态数据主题分析结果的融合;

3)通过大量实验验证了MD_LDA算法的高效性和有效性.

1 舆情数据主题分析

本文面向多模态数据的主题分析展开研究. 具体实现是先进行基于LDA主题模型的各单模态主题分析,然后在此基础上进行决策级的融合,从而设计并实现面向多模态舆情数据的主题分析算法.

1.1 单模态的主题分析

本节分别从图像、音频、视频的角度出发,将LDA主题模型理论作为图像、音频以及视频进行主题分析的先验结果,提出并开展了面向图像的主题分析研究、面向音频的主题分析研究以及面向视频的主题分析研究.

1.1.1 面向图像的主题分析

算法的基本思想如下:

1)对舆情数据集中的图像和文本进行预处理,并提取图像底层特征;

2)基于预处理后的文本训练构建LDA主题模型;

3)基于图像底层SIFT特征,利用K-Means聚类算法构建视觉词袋,完成图像数据到视觉矢量的转换;

4)结合舆情训练文本的主题概率分布与利用概率分布表示的图像矢量数据,构建文本主题-视觉词相似度矩阵;

5)利用视觉词袋以及文本主题-视觉词相似度矩阵得到待分析图像的主题概率分布.

给定舆情新闻集合M,文本数据d,隐含主题数目m,图像数据I,视觉词汇数目k,舆情新闻集可表示为$M=\{I,d\}$. 在进行视觉主题学习之前,利用舆情新闻集中的舆情文本d构建LDA主题模型,进而通过m个主题上的概率分布对每篇舆情文本进行表示,即

$$d_i=p(_1|d_i),p(_2|d_i),\ldots,p(_m|d_i). \tag{1}$$

接下来,提取图像集中图像I的底层SIFT特征构建视觉词袋BoIW (bag of image words)模型,继而可通过BoIW模型中所包含视觉词的概率分布对图像I_i的语义进行描述,即

$$I_i=p(_1|I_i),p(_2|I_i),\ldots,p(_k|I_i). \tag{2}$$

然后,通过视觉主题学习构建文本主题-视觉词关系矩阵Img_topic_dic,用以表示文本主题与视觉词之间的相关性.通过计算视觉词与文本主题的相关性,得到图像与文本主题的关联度.

1.1.2 面向音频的主题分析

算法的基本思想如下:

1)对舆情数据集中的音频和文本进行预处理,并提取音频底层特征;

2)基于预处理后的文本训练构建LDA主题模型;

3)基于音频底层MFCC特征[12],利用K-Means聚类算法构建听觉词袋,完成音频数据到听觉矢量的转换;

4)结合舆情训练文本的主题概率分布以及利用概率分布表示的音频矢量数据,经迭代学习构建文本主题-听觉词相似度矩阵;

5)对待分析的音频数据,基于百度api提供的技术实现音频转文本,提取待分析音频的显性语义;

6)基于预训练的LDA主题模型得到待分析音频的显性主题概率分布;

7)利用BoAW模型以及文本主题-听觉词相似度

矩阵得到待分析音频的隐性主题概率分布;

8) 结合待分析音频的显性与隐性主题概率分布,利用求和规则得到最终的主题概率分布.

在音频特征抽取后,通过构建听觉词袋BoAW (bag of audio words)模型描述音频语义,取一定数量的MFCC特征向量聚成k个类[12]作为BoAW中的听觉词汇. 不同的音频存在不同的词汇频率[13],对于一段语音数据,在对其进行分段之后,利用训练好的BoAW模型将各小段分配到这k个类别上,并对其进行归一化,实现音频数据到听觉矢量数据的转换.

给定舆情新闻集M,文本数据d,隐含主题数目m,音频数据A,利用BoAW模型进行语义描述后形成的听觉词汇数目k,舆情新闻集可表示为$M=\{A,d\}$. 利用式(1)将舆情文本d构建LDA主题模型,可通过m个主题上的概率分布对每篇舆情文本进行表示. 接下来,提取音频A的底层MFCC特征构建BoAW模型,通过BoAW将音频A_i描述为k个听觉词的概率分布

$$A_i = p(_1|A_i), p(_2|A_i), \ldots, p(_k|A_i). \tag{3}$$

然后,通过听觉主题学习来构建文本主题-听觉词关系矩阵Aud_topic_dic,用以表示文本主题与听觉词之间的相关性. 通过计算听觉词与文本主题的相关性,便可得到音频与文本主题的关联度.

1.1.3　面向视频的主题分析

算法的具体研究思路如下:

1) 对视频数据进行预处理,按照固定的频率提取视频帧,并对帧图像进行灰度处理以及切分提取视频字幕数据;

2) 基于预训练的LDA主题模型得到字幕数据的主题概率分布;

3) 利用基于帧间差分的算法提取待分析视频关键帧,通过提出的融合图像语义的视觉主题分析算法得到关键帧数据的主题概率分布;

4) 抽取视频数据中包含的音频数据并提取其底层特征;

5) 利用提出的音频主题分析算法得到音频数据的主题概率分布;

6) 基于待分析视频中提取得到的各文本、图像、音频模态数据的主题概率分布,利用求和规则得到待分析视频数据的主题概率分布.

对于视频中蕴含的文本模态信息,通过视频数据包含的字幕信息对其进行表示. 在此按照固定频率的思想提取视频帧数据,通过对各帧图像进行切分得到字幕图像,并对各字幕图像进行灰度处理保证截取的字幕图像更加清晰,更有利于视频中包含文本特征的提取. 在此基础上,可通过利用关键帧集描述视频数据中蕴含的视觉信息,采用基于帧间差分的算法进行视频关键帧的提取[14],即当视频数据中某一帧与前一帧图像内容产生了较明显的变化时,判断此帧为关键帧,并将其提取出来作为视频的视觉信息描述.

1.2　多模态的主题分析

舆情事件包含由不同种模态特征信息混合构成的多模态数据,仅考虑某一模态的信息只能从局部侧面反映出事件内容的信息[15]. 本节设计一种多模态数据主题分析框架,如图1所示. 在此基础上,设计并实现了基于LDA的多模态数据主题分析MD_LDA (multimodal data topic analysis based on LDA)算法,全面考虑舆情事件中包含的文字、图像、音频以及视频等模态数据内容,融合各模态数据信息完成多模态数据的主题分析.

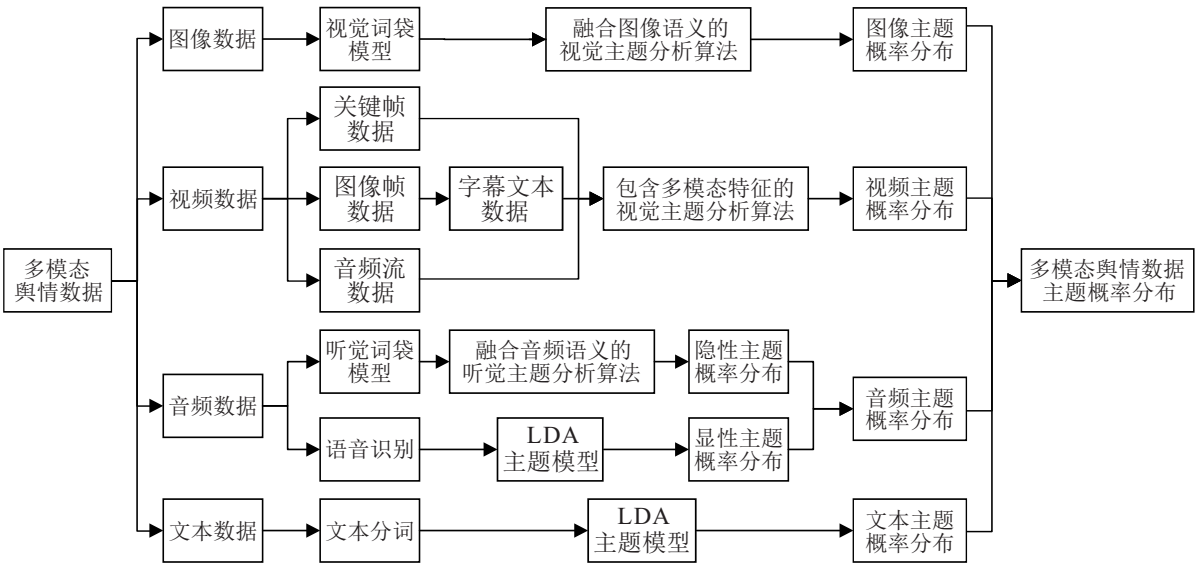

图1　多模态数据主题分析框架

算法的具体研究思路如下:

1) 分别提取舆情数据集中文本、图像、音频以及视频4种模态数据;

2) 基于文本集,预训练构建LDA主题模型;

3) 利用融合图像语义的视觉主题分析算法构建文本主题-视觉词矩阵;

4) 利用融合音频语义的听觉主题分析算法构建文本主题-听觉词矩阵;

5) 提取待分析舆情事件中文本、图像、音频以及视频模态数据;

6) 基于预训练的LDA主题模型得到待分析舆情事件中文本主题概率分布;

7) 利用融合图像语义的视觉主题分析算法以及音频主题分析算法分别得到待分析舆情事件中图像以及音频的主题概率分布;

8) 利用视频主题分析算法得到待分析舆情事件中视频的主题概率分布;

9) 利用求积规则融合各模态主题概率得到待分析舆情事件主题概率分布.

在各模态数据主题分析完成后,将文本、图像、音频和视频的主题分布根据一定的数学规则进行融合,得出最终的舆情新闻主题分析结果.在进行决策层融合[16]时,大多采用求和规则与求积规则[17],下面将对这两种数学规则进行介绍.

1.2.1 求和规则

假设经LDA主题模型训练后得到的文本主题分析结果为$P(\theta_j|T)(j=1,2,\dots,m)$,而其对应的各模态图像、音频以及视频主题分析结果分别为$P(\theta_j|I)(j=1,2,\dots,m)$,$P(\theta_j|A)(j=1,2,\dots,m)$,$P(\theta_j|V)(j=1,2,\dots,m)$. 经求和规则,得到最终归属各主题分析的概率

$$\begin{cases} P_j = aP(\theta_j|T)+bP(\theta_j|I)+ \\ \qquad cP(\theta_j|A)+dP(\theta_j|V), \\ 1\leqslant j\leqslant m,\ a+b+c+d=1. \end{cases} \tag{4}$$

其中:m代表主题数目;a、b、c、d分别代表舆情事件各模态文本、图像、音频以及视频主题的权重;P_j表示融合各模态主题分析结果后主题j的概率,选择$P_1,P_2,\dots,P_m$中概率最大值所对应的主题作为最终舆情主题分析结果.

1.2.2 求积规则

假设经LDA主题模型训练后得到文本主题分析结果为$P(\theta_j|T)(j=1,2,\dots,m)$,对应的各模态图像、音频及视频主题分析结果分别为$P(\theta_j|I)(j=1,2,\dots,m)$,$P(\theta_j|A)(j=1,2,\dots,m)$,$P(\theta_j|V)(j=1,2,\dots,m)$. 经求积规则,得到最终主题分析概率$P_j$,选择$P_1,P_2,\cdots,P_m$中概率最大值所对应的主题作为最终舆情主题分析结果,即

$$P_j = P(\theta_j|T)P(\theta_j|I)P(\theta_j|A)P(\theta_j|V),\ 1\leqslant j\leqslant m. \tag{5}$$

其中求和规则需要选定各模态数据主题分析的结果权重,权重选择影响主题分析结果,因此MD_LDA算法选择求积规则完成各模态数据主题分析结果的融合,从而实现更为准确有效的舆情主题分析研究.

2 实验分析

首先对实验环境进行介绍,然后开展各模态数据主题分析实验并分析实验结果.

2.1 实验设置

算法采用Python实现,利用其提供的gensim库、sklearn库等进行模型的构建,表1展示了开展实验所需要的实验环境.实验选取的数据集主要来自于新浪网、凤凰网等时事新闻网站,采集了自2021年10月至2022年3月内医疗、体育、社会等领域舆情突发事件的相关报道,共涉及14 999篇文本数据及其对应的图像、音频与视频数据等. 为了便于对算法实验效果进行衡量,在此采用精确率(precision)、召回率(recall)以及F值(F-measure)作为对舆情新闻主题词抽取效果的评判标准[18].

表1 实验环境表

配置	规格
CPU	Intel(R) Xeon(R) CPU E7-4809 v3
内存	64 GB
硬盘	512 GB
网卡	1 Gb/s 自适应以太网卡
操作系统	Windows 10

2.2 最优主题数选择

随着主题数目的不断递增,LDA模型的困惑度bound值表现出逐渐递增的趋势,当主题数目增长到13时,模型困惑度bound值趋于下降. 因此,选取主题数目为1～13之间的LDA主题模型来继续进行下一步的模型最优主题数目选择,模型困惑度bound值曲线如图2所示.

随着主题数目的不断递增,LDA主题模型与上一主题数目模型的一致性数值表现出曲折变化的趋势,当主题数目增长到10时,主题模型一致性数值达到峰值. 主题数目为10时,该LDA主题模型效果最优,模型一致性数值曲线如图3所示.

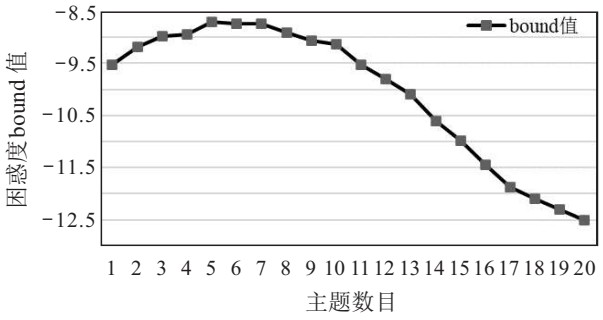

图 2　困惑度bound score曲线

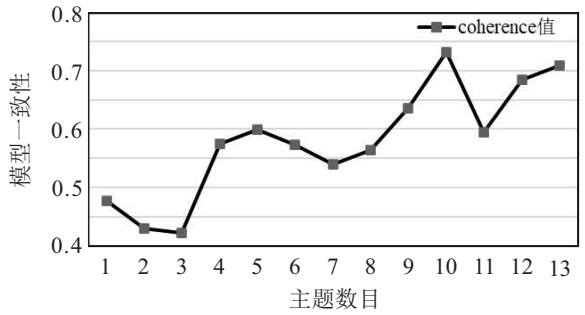

图 3　主题一致性曲线

对于主题模型效果的度量,主题结构相似度(corre)将基于密度的聚类思想集成到最优主题数目选择的计算中. 在LDA的主题结构中认为一个主题相当于一个语义群,群内相似度越大表明该群能更明确地表示含义,而群内相似度越小则表明该主题结构越稳定. 在基于主题密度的LDA中,基于密度聚类的思想符合LDA中选择最佳主题结构的标准,其保证在群内相似度尽量高,而在群间相似度尽量低. 基于此,文献[19]提出并证明了主题结构之间的平均相似度最小时所对应模型为最优模型这一定理. 该定理用β矩阵中主题在U维词空间中的分布对主题矢量进行表达,并将主题矢量之间的矢量余弦距离作为主题之间的相关度,即

$$\mathrm{corre}(Z_i, Z_j) = \mathrm{corre}(\beta_i, \beta_j) = \sum_{v=0}^{V} \beta_{iv} \circ \beta_{jv} \Big/ \sqrt{\sum_{v=0}^{V} (\beta_{iv})^2 \sum_{v=0}^{V} (\beta_{jv})^2}. \tag{6}$$

$\mathrm{corre}(Z_i, Z_j)$越小,主题间的独立性就越强. 利用所有主题之间的平均相关度来衡量主题结构的稳定性,即

$$\mathrm{avg}_{\mathrm{corre(structure)}} = \frac{\sum_{i=0}^{K-1} \sum_{j=i+1}^{K} \mathrm{corre}(Z_i, Z_j)}{K \dfrac{(K-1)}{2}}. \tag{7}$$

根据主题一致性曲线变化趋势初步确定主题数目为10的模型为最优LDA主题模型,在此利用模型主题结构相似度再次对最优主题数目进行验证. 选取主题数目为2～10之间的模型计算主题结构相似度,得到平均相似度变化曲线如图4所示.

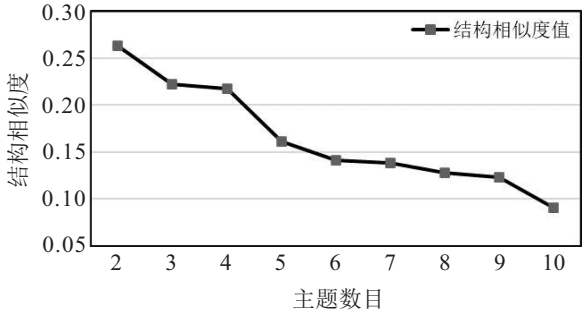

图 4　各主题平均相似度曲线

由于主题结构之间平均相似度最低时模型最优,可见随着主题数目的不断递增,当主题数目为10时,该LDA主题模型效果最优,进一步验证了主题数目为10的主题模型为最优LDA主题模型.

2.3　实验结果

首先分别对数据集中包含的文本、图像、音频以及视频各单模态数据进行主题分析实验; 然后,用所提出的MD_LDA算法对多模态数据进行主题分析实验; 接着,分别与图像、音频、视频数据相结合进行主题分析实验,开展各组合特征主题分析实验; 最后,为了验证MD_LDA算法的有效性,将其与传统经典关键词提取算法TF-IDF、TextRank进行对比分析.

2.3.1　文本数据概率分布

基于主题分析所得到的主题概率分布,分别计算测试文本内的分词与LDA主题模型中各主题词的余弦距离,得到测试文本内各分词与主题词的相似度,取10个相似度最高的词语作为最终主题词,前3个主题的抽取结果如表2所示.

表 2　文本数据概率分布结果

主题1	主题2	主题3
病例(0.968 925 9)	国足(0.967 364 6)	乌克兰(0.966 531)
核酸(0.968 925 9)	李铁(0.967 364 6)	北约(0.966 531)
感染者(0.968 925 9)	世界杯(0.967 364 6)	顿巴斯(0.966 531)
全市(0.968 925 9)	阿曼(0.967 364 6)	军事行动(0.966 531)
肺炎(0.968 925 8)	国脚(0.967 364 6)	官员(0.966 531)
社区(0.968 925 6)	足球(0.967 364 5)	军队(0.966 531 1)
通报(0.968 925 3)	杜绝(0.372 839 6)	美国(0.966 530 2)
病毒(0.968 925 3)	激情(0.372 839 6)	地区(0.966 525 8)
本土(0.968 924 9)	杜威(0.372 839 6)	命令(0.322 675 1)
轨迹(0.968 924 9)	高昂(0.372 839 6)	策划(0.322 675 1)

2.3.2　图像数据概率分布

经视觉主题学习后,利用新闻测试数据集中的图像数据集对融合图像语义的视觉主题分析算法进行验证. 将测试图像与生成的主题-视觉词矩阵进行迭代运算,得到该图像在不同主题上的概率分布. 基于

上述实验结果利用求积规则得到联合文本与图像数据的多模态主题概率分布,基于该分布分别计算测试文本内分词与LDA主题模型各主题词的余弦距离,得到测试文本内各分词与主题词的相似度,取10个相似度最高的词语作为最终主题词,前3个主题的抽取结果如表3所示.

表 3 图像数据概率分布结果

主题1	主题2	主题3
通报 (0.963 165 7)	国足 (0.968 837 6)	乌克兰 (0.965 875 1)
人员 (0.963 165 4)	李铁 (0.968 837 6)	北约 (0.965 875 1)
病例 (0.963 165 2)	世界杯 (0.968 837 6)	顿巴斯 (0.965 875 1)
核酸 (0.963 165 2)	阿曼 (0.968 837 6)	军事行动 (0.965 875 1)
感染者 (0.963 165 2)	国脚 (0.968 837 6)	官员 (0.965 875 1)
全市 (0.963 165 2)	足球 (0.968 837 5)	军队 (0.965 874 6)
肺炎 (0.963 165 1)	杜绝 (0.353 898 7)	美国 (0.965 873 7)
社区 (0.963 164 9)	激情 (0.353 898 7)	地区 (0.965 869 2)
病毒 (0.963 164 6)	杜威 (0.353 898 7)	命令 (0.320 271 7)
部门 (0.963 164 6)	高昂 (0.353 898 7)	策划 (0.320 271 7)

2.3.3 音频数据概率分布

利用新闻测试数据集中的音频集对融合音频语义的听觉主题分析算法进行验证,将测试音频与生成的主题-听觉词矩阵进行计算,得到该音频在不同主题上的概率分布,结合音频显性主题概率分布利用求和规则得到最终音频主题概率分布. 结合音频与文本主题概率分布结果,利用求积规则得到联合文本与音频数据的多模态主题概率分布,基于该分布分别计算测试文本内分词与LDA主题模型各主题词的余弦距离,得到测试文本内各分词与主题词的相似度,取10个相似度最高的词语作为最终主题词,前3个主题的抽取结果如表4所示.

表 4 音频数据概率分布结果

主题1	主题2	主题3
通报 (0.963 896 9)	国足 (0.967 364 6)	乌克兰 (0.965 704 2)
人员 (0.963 896 6)	李铁 (0.967 364 6)	北约 (0.965 704 2)
病例 (0.963 896 5)	世界杯 (0.967 364 6)	顿巴斯 (0.965 704 2)
核酸 (0.963 896 5)	阿曼 (0.967 364 6)	军事行动 (0.965 704 2)
感染者 (0.963 896 5)	国脚 (0.967 364 6)	官员 (0.965 704 1)
全市 (0.963 896 5)	足球 (0.967 364 5)	军队 (0.965 703 6)
肺炎 (0.963 896 4)	杜绝 (0.372 839 6)	美国 (0.965 702 7)
社区 (0.963 896 2)	激情 (0.372 839 6)	地区 (0.965 698 2)
病毒 (0.963 895 9)	杜威 (0.372 839 6)	命令 (0.319 662 4)
部门 (0.963 895 9)	高昂 (0.372 839 6)	策划 (0.319 662 4)

2.3.4 文本结合视频数据

本小节对联合文本与视频数据进行主题分析的效果进行实验,旨在联合文本与视频数据进行主题分析,基于测试数据集中的视频集,分别提取该集合中的视频帧集、关键帧集和音频集开展主题分析实验. 结合视频与文本主题概率分布结果,利用求积规则得到联合文本数据与视频数据的多模态主题概率分布,基于该分布分别计算测试文本内分词与LDA主题模型各主题词的余弦距离,得到测试文本内各分词与主题词的相似度,取10个相似度最高的词语作为最终主题词,前3个主题的抽取结果如表5所示.

表 5 视频数据概率分布结果表

主题1	主题2	主题3
人员 (0.936 705 1)	国足 (0.967 686 3)	乌克兰 (0.965 346 8)
通报 (0.936 704 0)	李铁 (0.967 686 3)	北约 (0.965 346 8)
部门 (0.936 702 8)	世界杯 (0.967 686 3)	顿巴斯 (0.965 346 8)
病例 (0.936 702 2)	阿曼 (0.967 686 3)	军事行动 (0.965 346 8)
核酸 (0.936 702 2)	国脚 (0.967 686 3)	官员 (0.965 346 7)
感染者 (0.936 702 2)	足球 (0.967 686 2)	军队 (0.965 346 2)
全市 (0.936 702 2)	杜绝 (0.328 704 0)	美国 (0.965 346 2)
报告 (0.936 702 2)	激情 (0.328 704 0)	地区 (0.965 340 8)
肺炎 (0.936 702 1)	杜威 (0.328 704 0)	命令 (0.318 349 1)
市民 (0.936 702 1)	高昂 (0.328 704 0)	策划 (0.318 349 1)

2.3.5 文本结合多模态数据

本小节结合各模态概率分布,利用求积规则完成舆情新闻数据的主题分析. 区别于传统的仅针对单一模态特征进行主题分析的方法,基于各模态舆情数据的实验,MD_LDA算法联合各模态数据信息进行主题分析,结合上述各模态主题概率分布结果,得到面向多模态数据的主题概率分布. 基于该分布计算测试文本内分词与LDA模型各主题词的余弦距离,得到测试文本内各分词与主题词的相似度,取10个相似度最高的词语作为最终主题词,前3个主题的抽取结果如表6所示.

表 6 MD_LDA算法主题词抽取结果

主题1	主题2	主题3
措施 (0.936 695 8)	国足 (0.966 802 8)	乌克兰 (0.965 417 4)
原则 (0.936 688 4)	李铁 (0.966 802 8)	北约 (0.965 417 4)
地区 (0.936 688 1)	世界杯 (0.966 802 8)	顿巴斯 (0.965 417 4)
新闻 (0.936 684 6)	阿曼 (0.966 802 8)	军事行动 (0.965 417 4)
机场 (0.936 683 9)	国脚 (0.966 802 8)	官员 (0.965 417 4)
视频 (0.936 679 6)	足球 (0.966 802 8)	军队 (0.965 416 9)
道路 (0.936 679 2)	杜绝 (0.324 509 9)	美国 (0.965 416 0)
专家 (0.418 348 2)	激情 (0.324 509 9)	地区 (0.965 411 4)
人员 (0.418 325 6)	杜威 (0.324 509 9)	命令 (0.318 658 0)
发布会 (0.418 323 0)	高昂 (0.324 509 9)	策划 (0.318 658 0)

2.3.6 组合特征对比实验

通过依次结合文本与不同单模态构成的组合特征以及MD_LDA算法所提出的组合特征进行对比实验. 各评估指标取平均值后的结果对比如图5所示,可以看出MD_LDA算法提出的组合特征在精确率、

召回率以及F值上均优于其余组合特征. 同时,表7显示了不同组合特征对测试数据集进行主题分析的精确率、召回率和F值取均值后的对比实验结果.

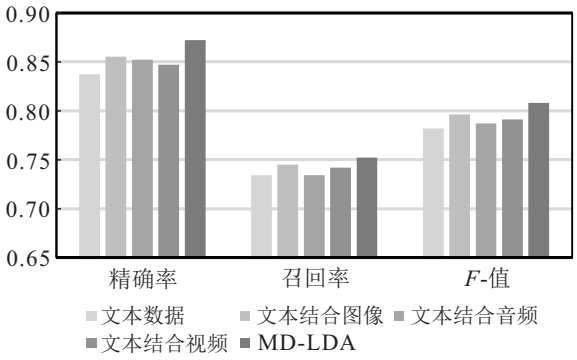

图5　特征组合结果对比

表7　各模态抽取特征组合的对比实验结果

算法类型	精确率	召回率	F-值
文本	0.837	0.734	0.782
文本＋图像	0.855	0.745	0.796
文本＋音频	0.852	0.734	0.787
文本＋视频	0.847	0.742	0.791
MD_LDA	0.872	0.752	0.808

2.3.7　经典算法对比实验

将MD_LDA算法分别与传统关键词提取算法TF-IDF、TextRank进行对比分析,通过权重迭代计算每个词汇节点的得分,根据得分情况完成关键词抽取[20]. 表8展示了不同算法的对比实验结果.

表8　各算法对比实验结果

算法类型	精确率	召回率	F-值	运行时间/min
TF-IDF	0.775	0.751	0.763	60.341
TextRank	0.751	0.724	0.737	31.125
MD_LDA	0.872	0.752	0.808	29.288

图6分别展示了在200～1 200篇不等新闻数据数目下,3种方法在主题词抽取方面的实验结果对比. 由图6(a)和图6(b)可知,MD_LDA算法较其余两种算法主题词抽取的准确率更高. 由图6(c)可知,MD_LDA算法的F值显著高于其他两种算法. 由图6(d)可知,MD_LDA算法的运行时间较其余两种算法均有所缩短,主题词提取效率更高. 由此可见,MD_LDA算法较其余两种算法主题词抽取的效果更好、内容更准确、效率更高.

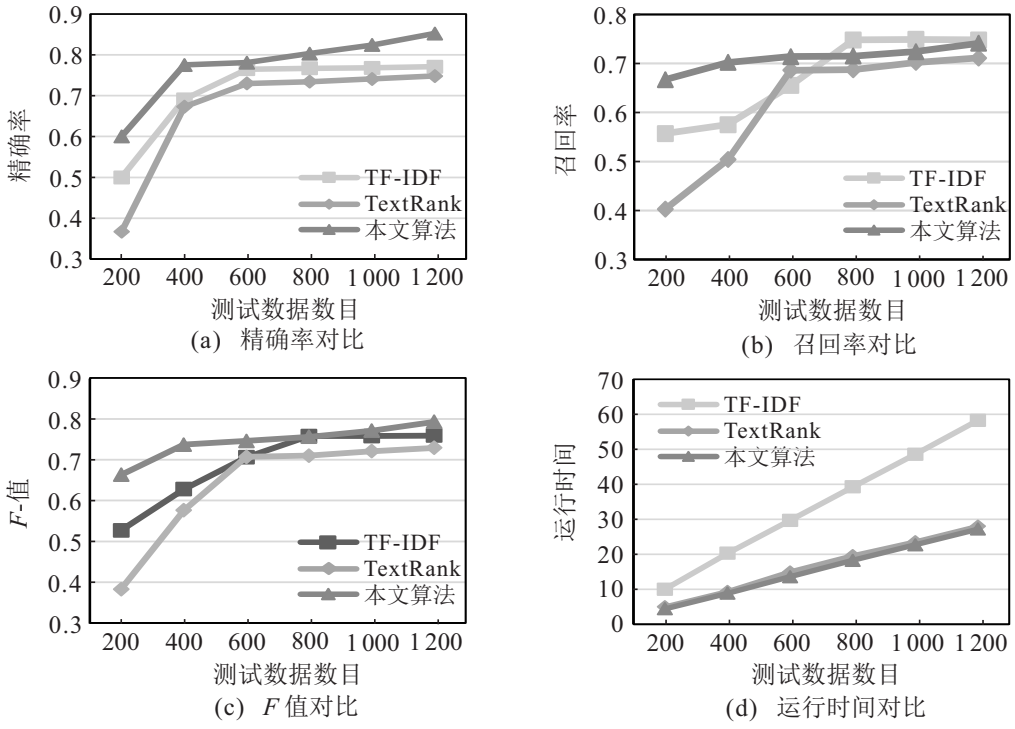

图6　实验结果对比

3　结　论

本文提出基于LDA的多模态数据主题分析算法用于舆情事件主题分析研究,对各模态数据分别进行主题分析,在其基础上实现包含多模态特征主题分析,提出了基于LDA的多模态数据主题分析算法,实现了包含多模态特征的舆情数据的主题分析. 将各模态数据融入主题分析过程中,挖掘各模态数据与文本主题的相关性,利用最优LDA主题模型以及多模态数据内容进行主题分析. 最后,通过实验验证了MD_LDA算法提取主题的语义信息更清晰、明确,且

算法效率更高.

参考文献(References)

[1] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[2] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse Processes, 1998, 25(2/3): 259-284.

[3] Hofmann T. Probabilistic latent semantic indexing[C]. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley: ACM, 1999: 50-57.

[4] Papadimitriou C H, Raghavan P, Tamaki H, et al. Latent semantic indexing: A probabilistic analysis[J]. Journal of Computer and System Sciences, 2000, 61(2): 217-235.

[5] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[6] Zhao W X, Jiang J, Weng J S, et al. Comparing twitter and traditional media using topic models[C]. Proceedings of the 33rd European Conference on Advances in Information Retrieval. New York, 2011: 338-349.

[7] Putthividhy D, Attias H T, Nagarajan S S. Topic regression multi-modal latent dirichlet allocation for image annotation[C]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 3408-3415.

[8] Barnard K, Duygulu P, Forsyth D, et al. Matching words and pictures[J]. Journal of Machine Learning Research, 2003, 3(2): 1107-1135.

[9] Wang D Z, Mao K Z. Learning semantic text features for web text-aided image classification[J]. IEEE Transactions on Multimedia, 2019, 21(12): 2985−2996.

[10] Takahashi H, Lakhani U, Raza A. Video abstract extraction using semantic categorization[C]. The 6th Global Conference on Consumer Electronics. Nagoya: IEEE, 2017: 1-3.

[11] Li W X, Joo J, Qi H, et al. Joint image-text news topic detection and tracking by multimodal topic and-or graph[J]. IEEE Transactions on Multimedia, 2017, 19(2): 367-381.

[12] 冯霞, 胡志毅, 刘才华. 跨模态检索研究进展综述[J]. 计算机科学, 2021, 48(8): 13-23.
(Feng X, Hu Z Y, Liu C H. Survey of research progress on cross-modal retrieval[J]. Computer Science, 2021, 48(8): 13-23.)

[13] 李荣杰, 蒋兴浩, 孙锬锋. 一种基于音频词袋的暴力视频分类方法[J]. 上海交通大学学报, 2011, 45(2): 214-218.
(Li R J, Jiang X H, Sun T F. Violent videos classification algorithm based on bag of audio words[J]. Journal of Shanghai Jiao Tong University, 2011, 45(2): 214-218.)

[14] 吴悦, 雒江涛, 刘锐, 等. 基于感知哈希和切块的视频相似度检测方法[J]. 计算机应用, 2021, 41(7): 2070-2075.
(Wu Y, Luo J T, Liu R, et al. Video similarity detection method based on perceptual hashing and dicing[J]. Journal of Computer Applications, 2021, 41(7): 2070-2075.)

[15] 龚志, 邵曦. 基于多模态的音乐推荐系统[J]. 南京信息工程大学学报: 自然科学版, 2019, 11(1): 68-76.
(Gong Z, Shao X. A music recommendation system based on multi-modal fusion[J]. Journal of Nanjing University of Information Science & Technology: Natural Science Edition, 2019, 11(1): 68-76.)

[16] 蒋雨肖, 丁晟春, 吴鹏. 基于BiLSTM-VGG16的多模态信息特征分类研究[J]. 情报理论与实践, 2021, 44(11): 180-186.
(Jiang Y X, Ding S C, Wu P. A study on the classification of features of multi-modal information based on BiLSTM-VGG16[J]. Information Studies: Theory & Application, 2021, 44(11): 180-186.)

[17] 谢坷珍. 融合人脸表情和语音的双模态情感识别研究[D]. 青岛: 中国海洋大学, 2015.
(Xie K Z. Research on bimodal emotion recognition based on facial expression and speech signal[D]. Qingdao: Ocean University of China, 2015.)

[18] 倪宁宁. 跨媒体话题检测与观点分析研究[D]. 北京: 北京邮电大学, 2019.
(Ni N N. Research on cross-media topic detection and opinion analysis[D]. Beijing: Beijing University of Posts and Telecommunications, 2019.)

[19] Cao J, Xia T, Li J T, et al. A density-based method for adaptive LDA model selection[J]. Neurocomputing, 2009, 72(7/8/9): 1775-1781.

[20] 王涛, 李明. 改进的关键词提取算法研究[J]. 重庆师范大学学报: 自然科学版, 2019, 36(3): 98-104.
(Wang T, Li M. Study on an improved keyword extraction algorithm[J]. Journal of Chongqing Normal University: Natural Science, 2019, 36(3): 98-104.)

作者简介

赵越(1966−), 女, 教授, 博士, 从事医疗信息管理、医学人工智能等研究, E-mail: zhaoyue@bmie.neu.edu.cn;

郝琨(1989−), 男, 博士后, 从事大数据、区块链技术等研究, E-mail: haokun@bmie.neu.edu.cn;

时彩云(1998−), 女, 硕士生, 从事主题分析的研究, E-mail: yun17051026@163.com;

解胜震(1997−), 女, 硕士, 从事人工智能的研究, E-mail: xieshengzhen_neu@163.com;

王之琼(1980−), 女, 教授, 博士, 从事医疗大数据、人工智能等研究, E-mail: wangzq@bmie.neu.edu.cn;

信俊昌(1977−), 男, 教授, 博士, 从事大数据管理、数据库技术等研究, E-mail: xinjunchang@mail.neu.edu.cn.