ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 3,May 2022,pp. 621－632
DOI:10. 16337/j. 1004-9037. 2022. 03. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于特征扩展的微博短文本流热点话题检测方法

李艳红[1,2]，谢梦娜[1,2]，王素格[1,2]，李德玉[1,2]

(1. 山西大学计算机与信息技术学院,太原 030006;2. 山西大学计算智能与中文信息处理教育部重点实验室,太原 030006)

摘　要：随着社交网络和互联网的飞速发展,产生了大量的微博短文本流数据。及时发现微博文本流中热点话题,对话题推荐和舆情监测等有重要作用。为了解决微博短文本特征稀疏问题,利用微博评论对微博进行特征扩展,提出了一种基于特征扩展的微博短文本流热点话题检测方法(Feature extension-based hot topic detection, FE-HTD)。首先利用评论用户的影响力以及评论文本的点赞数筛选评论文本,并使用词共现和词频-逆文档频率(Term frequency-inverse document frequency,TF-IDF)方法从选取的评论文本中抽取特征词完成对微博文本的特征扩展;然后计算微博文本流的词对速度、词对加速度,并根据点赞数、评论数计算微博文本强度,结合词对加速度与微博文本强度定义突发特征;最后,根据突发词对的速度确定可变长的热点话题窗口范围,通过聚类得到窗口中热点话题的主题结构。实验中,将所提算法与基于文本的话题检测(Text-based topic detection, T-TD)和基于突发词的话题检测(Burst words-based topic detection, BW-TD)进行对比实验。结果表明,本文算法FE-HTD准确率达76.4%,召回率达78.7%,与对比算法T-TD和BW-TD相比提高了10%。

关键词：微博短文本流;特征扩展;热点话题;用户影响力;增量聚类

中图分类号：TP391　　**文献标志码**:A

Hot Topic Detection Method of Microblog Short Text Stream Based on Feature Extension

LI Yanhong[1,2], XIE Mengna[1,2], WANG Suge[1,2], LI Deyu[1,2]

(1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China; 2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)

Abstract: With the rapid development of social networks and Internet, a large number of microblog short text stream data have been produced. Discovering hot topics from microblog text streams in time plays an important role in topic recommendation and public opinion monitoring. To solve the problem of sparse features of microblog, a feature extension-based hot topic detection (FE-HTD) method in microblog short text stream is proposed by using microblog comments to extend the features of microblog. To complete the feature extension of the microblog text, firstly, the comment text is selected by the influence of the comment users and the number of likes for comment text, and the feature words are extracted from the comment text by word co-occurrence and term frequency-inverse document frequency (TF-IDF) method. Then count the word pair speed, word pair acceleration and microblog text strength of the microblog short

基金项目：国家自然科学基金(62072294, 62076158, 61906112, 41871286);山西省重点研发计划(201803D421024, 201903D421041)。
收稿日期:2021-03-24;**修订日期**:2021-12-20

text stream. The burst feature is calculated by word pair acceleration and microblog text strength. Finally, the variable length window range of hot topic is determined according to the speed of the burst word pair, and the topic structure of hot topic in the window is obtained by clustering. In the experiment, the proposed algorithm is compared with the text-based topic detection (T-TD) method and the burst words-based topic detection (BW-TD) method. The results show that the accuracy of the proposed algorithm is 76.4%, and the recall rate is 78.7%, which are 10% higher than those of T-TD and BW-TD methods.

Key words: microblog short text stream; feature extension; hot topic; user influence; incremental clustering

引　言

社交网络和互联网的飞速发展使微博变为用户捕获信息的主要平台。根据新浪微博官方数据显示，2020年第一季度微博每月的活跃用户高达5.5亿，每天活跃用户高达2.41亿。微博作为一种舆情的聚焦工具，民众共享的信息或谈论主题在网络中广泛传播，使得微博的爆发力和破坏性更加强烈，从而产生巨大的社会影响。因此及时发现微博短文本流中的热点话题，有助于了解公众情绪和舆论，为话题推荐和政府决策提供依据。由于微博短文本流具有文本内容短小、特征稀疏以及话题不断变化的特点，因此如何从中实时发现热点话题是一项值得深入研究的课题。

网络上两种主要的热点话题检测方法分别以文本为中心[1]和以突发特征为中心[2]。以文本为中心检测热点话题是首先对文本聚类，然后分析各类簇的突发情况，判定有突发状态的类簇为热点话题。此类方法一般基于潜在狄利克雷分布(Latent Dirichlet allocation, LDA)主题模型[3]或者改进后的LDA主题模型[4]。例如，Kiejin[5]首先使用N-gram算法来构建多个词语，精确地捕获一个句子的含义，然后通过LDA主题模型实现话题的初步检测。Mehrotra等[6]提出优先考虑博文内容而非改进主题模型，即在文本预处理时对博文进行多方案聚合。文献[5,6]提供的各类方法前提都得保证数据是静态的，但在实际情况中，数据通常是在线文本流形式，而非静态数据。Wang等[7]首次提出了一种新的主题模型，即多属性狄利克雷分布(Multi attribute LDA，MA-LDA)，该模型把微博的时间特性和哈希标签结合到LDA主题模型中。周先琳[8]对微博文本预处理后，创建动态标签-潜在狄利克雷分布(Labeled-LDA，L-LDA)模型检测热点话题。由于利用主题模型进行参数估计比较耗时，因此很难满足实时检测热点话题。以特征为中心的方法指的是通过特征词的速度、动量等来判定突发状态，对有突发特征的微博文本进行聚类，从而确定热点话题。如Fung等[9]利用时间信息以及单词分布情况确定突发特征词；李汉才等[10]融合时序性和波动性计算话题热度；万越等[11]提出使用影响力因子，并结合热度因子修正动量模型确定突发特征；郑斐然等[12]根据突然高频出现在微博中的特征词的速度变化情况确定突发词；蔡莹等[13]结合词语权重定义突发状态，进而检测热点话题。由于以特征为中心的方法通过突发特征确定热点话题，因此省去了训练数据计算参数的时间。

为了解决短文本特征稀疏问题，目前有两类短文本扩展方法：基于外部资源扩展[14]和基于内部资源扩展[15]。基于外部资源扩展是指使用外部语料库(如：维基百科、Probase等)对短文本扩展。比如Cheng等[16]利用维基百科语料库获取丰富的信息进行短文本扩展；Li等[17]提出从Probase中提取概念和共现术语，并对词语进行消歧，从而扩展微博短文本特征，用于微博文本的分类研究；Duan等[18]提出用特征向量代替词来训练一个巨大的外部资料库，弥补了短文本的特征稀疏问题。基于内部资源扩展是指利用短文本自身的上下文关系，构建基于文本内容的词集实现短文本的扩展。如Paulo等[19]提出利用词共现和词向量在原始文档中创建一个大的伪文档从而进行特征扩展；张萌[20]利用微博文本所带链接内容丰富短文本特征，利用主题模型发现热点话题。与利用外部资源扩展方法相比，基于自身资

源的扩展方法在相似特征选取上具有优势,并且基于外部资源扩展的边界难以确定,引入的知识太笼统,会带来一些无关词从而使得特征扩展质量不高。考虑到一般情况下一条微博会包含多条评论文本,评论者针对该条微博文本表达观点和评价,因此本文选用微博评论文本作为扩展语料,并利用突发特征来检测热点话题。对于热点话题的检测,已有研究人员做了大量相关工作,但仍存在以下问题:(1)在微博短文本利用评论文本进行特征扩展时,没有考虑评论文本的用户影响力(如用户活跃度、粉丝数等),在筛选特征词时忽略了评论文本中的特征词与微博文本中的特征词的相关性;(2)在定义突发特征时,没有考虑微博文本强度(如点赞数、评论数等)对突发程度的影响。

基于上述问题,本文提出一种基于特征扩展的微博短文本流热点话题检测方法。根据用户活跃度、粉丝数等计算评论用户影响力,结合评论文本的点赞数筛选出高质量评论文本;使用词共现和词频-逆文档频率(Term frequency-inverse document frequency,TF-IDF)方法提取评论文本中的特征词,对微博文本进行特征扩展;提出利用点赞数、评论数计算微博文本强度,并作为计算词对速度的参数,进而确定突发词对;利用突发词对速度确定突发词对窗口,合并交叉、重叠、相邻的突发词对窗口得到热点话题窗口;最后通过采用吉布斯采样狄利克雷多项式混合模型(Gibbs sampling Dirichlet multinomial mixture model,GSDMM)[21]对热点话题窗口中的微博文本聚类,分析得到的每个类簇中特征词的重要度,提取重要度高的特征词来表示热点话题的主题结构。

1 问题形式化定义及符号说明

本文将微博文本流记作 $D=\{(x_1,y_1,t_1),(x_2,y_2,t_2),\cdots,(x_i,y_i,t_i),\cdots\}$,其中 $x_i=\left(w_1^i,w_2^i,\cdots,w_b^i\right)$,$y_i=\left\{y_{i1},y_{i2},\cdots,y_{im}\right\}$ 为 $\boldsymbol{x}_i$ 对应的评论文本集合。$\boldsymbol{y}_{ij}=\left(w_1^{ij},w_2^{ij},\cdots,w_l^{ij}\right)$ 表示第 i 条微博文本对应的第 j 条评论文本的特征词。为选取高质量评论文本并从中抽取特征词扩展微博文本,定义了词共现度 $WA(w_s^i,w_t^{ij})$ 和词区分度 $d(w_t^{ij})$。将 $\boldsymbol{x}_i$ 扩展以后的微博文本记为 $\boldsymbol{cx}_i=\left(w_1^i,w_2^i,\cdots,w_b^i,w_1',w_2',\cdots,w_c'\right)$,子集 $(w_1',w_2',\cdots,w_c')$ 表示从微博文本对应的所有评论文本中抽取的 c 个特证词。对于热点话题判定问题,将 $\boldsymbol{cx}_i$ 中任意两个特征词构成词对 (w_j,w_k),定义词对速度 $v_{j,k}^{i,\Delta T}$ 和词对加速度 $a_{j,k}^i$,当词对加速度大于阈值,则判定存在热点话题。为得到热点话题主题结构,将词对加速度大于阈值的词对作为突发词对,合并多个交叉、重叠或者相邻的突发词对窗口形成热点话题窗口 W,通过聚类得到热点话题主题结构。

话题检测框架的主要符号及其说明如表1所示。

表1 话题检测框架主要符号

Table 1 Main notations of topic detection framework

符号	说明
u	用户影响力
cts	评论文本强度
λ	评论文本强度阈值
μ	词共现度阈值
ω	词区分度阈值
α	词对加速度阈值
ts	微博文本强度
$f_i(w_j,w_k)$	词对频率
β	突发词对速度阈值
BWW	突发词对窗口

2 基于评论文本的特征扩展和突发特征定义

2.1 特征扩展定义

表2给出了一条微博短文本及其对应的评论文本。从表2可发现,微博文本“有病的人找不到床位”中并没有提到任何和医保相关的内容,但是通过其评论文本可知这条微博文本是在抨击医保乱象。由此可知,可以利用微博评论对微博进行特征扩展。

为了解决微博短文本特征稀疏问题,使用

表2 微博文本及其评论文本

Table 2 Microblog text and its comment text

微博文本	评论文本
有病的人找不到床位	1.骗保已经是不是秘密的秘密了。小县城抬头不见低头见的,人家这么干,你不这么干,多尴尬。 2.让真正有病的人找不到床位,骗保,骗保,严查严查… 3.这都是骗保的,央视曝光过某地同样的事情。住院还给钱呢? 4.小医院真的要好好管管了,去住院,做很多没有必要的检查,也是套社保的一种方式。

微博评论文本作为扩展语料。由于微博评论文本的质量参差不齐,为了得到高质量的扩展特征,首先基于评论文本的强度来对评论文本进行筛选,然后根据词共现度和词区分度从筛选出的评论文本中抽取特征词用于微博短文本的特征扩展。

下面首先给出评论文本强度定义。

定义1 评论文本强度cts表示该条文本参考价值,可以通过发表评论用户的影响力以及评论文本对应的点赞数定义,即

$$\mathrm{cts}=\begin{cases}u+\gamma(\mathrm{suport\_n}) & \mathrm{suport\_n}<10\\ u+\lg(\mathrm{suport\_n}) & \mathrm{suport\_n}\geqslant 10\end{cases} \tag{1}$$

式中:suport_n表示点赞数,u表示用户影响力。u的计算式为

$$u=\begin{cases}\gamma\left(\dfrac{\mathrm{fan\_n}}{\mathrm{follow\_n}}+\dfrac{\mathrm{fan\_n}}{\mathrm{sum}}\right)+\mathrm{level} & \dfrac{\mathrm{fan\_n}}{\mathrm{follow\_n}}+\dfrac{\mathrm{fan\_n}}{\mathrm{sum}}<10\\ \lg\left(\dfrac{\mathrm{fan\_n}}{\mathrm{follow\_n}}+\dfrac{\mathrm{fan\_n}}{\mathrm{sum}}\right)+\mathrm{level} & \dfrac{\mathrm{fan\_n}}{\mathrm{follow\_n}}+\dfrac{\mathrm{fan\_n}}{\mathrm{sum}}\geqslant 10\end{cases} \tag{2}$$

式中:fan_n表示粉丝数,follow_n表示关注数,sum表示一条微博文本对应的所有评论用户的粉丝数,level($0\leqslant\mathrm{level}\leqslant 1$)为微博官方通过用户的活跃度、用户信誉值等综合考虑所得的一个用户信用值,γ为调节参数($0<\gamma<1$)。

为了从评论文本中提取特征词,给出词共现度和词区分度的定义。

定义2 微博文本x_i中的特征词为$\left(w_1^i,w_2^i,\cdots,w_b^i\right)$,$x_i$对应的评论文本集为$\left\{y_{i1},y_{i2},\cdots,y_{im}\right\}$,其中评论$y_{ij}$的特征词为$\left(w_1^{ij},w_2^{ij},\cdots,w_l^{ij}\right)$,将词共现度记为$WA(w_s^i,w_t^{ij})$,定义为

$$WA(w_s^i,w_t^{ij})=\frac{|\{y_{ij}|n_{w_s^i}\geqslant 1,n_t^{ij}\geqslant 1\}|}{m} \tag{3}$$

式中:$w_s^i\in x_i$,$w_t^{ij}\in y_{ij}$,$n_{w_s^i}$、n_t^{ij}分别表示特征词w_s^i、w_t^{ij}在评论文本中出现的次数。

定义3[22] 微博文本x_i对应的评论文本集为$\left\{y_{i1},y_{i2},\cdots,y_{im}\right\}$,利用TF-IDF定义$y_{ij}$中特征词的区分度$d(w_t^{ij})$为

$$d(w_t^{ij})=TF\times\mathrm{IDF}=\frac{n_t^j}{s_j}\times\lg\left(\frac{m}{n_t}\right) \tag{4}$$

式中:n_t^j表示特征词w_t^{ij}在评论文本y_{ij}中出现的次数,s_j表示y_{ij}中所有特征词的个数,n_t表示含有特征词w_t^{ij}的评论文本条数。

为了提取相似度高或代表性强的特征词,定义了词共现度阈值μ和词区分度阈值ω,当

$WA(w_s^i, w_t^{ij}) \geqslant \mu$或$d(w_t^{ij}) \geqslant \omega$时,则提取特征词$w_t^{ij}$用于特征扩展。为了更直观地说明扩展结果,通过例子来表示,如表3所示。

表3 微博文本扩展示例

Table 3 Microblog text extension examples

微博文本特征	扩展后的微博文本特征
众志成城 万众一心	众志成城 万众一心 居家 隔离 疫情 快速 褪去 武汉 封城
拔地而起 火神山医院	拔地而起 火神山 医院 医疗 工程 展开 新冠肺炎 疫情 防疫 物资 口罩 护士 志愿者 前线 支援 抗疫
防疫 物资 匮乏	防疫 物资 匮乏 医院 囤积 情况 解释 后期 陆续 到达 感谢 国民 力量 钟南山 新冠肺炎 封城
医院 囤积 解释	政府防控 管理 措施 支援 居家 隔离 抗击 疫情 口罩 消毒液 测温
摩拜 单车 结束	摩拜 单车 结束 服务 权益 账户 余额 美团
小黄车 天下	小黄车 天下 摩拜 结束 权益 接入 美团 余额 存在 账户 储值

2.2 突发特征定义

微博短文本流中主要分布两类话题,即热点话题和一般话题。直观上,热点话题区别于一般话题的特征是:(1)出现热点话题时,微博文本流中涌现出大量的相关微博,而一般话题微博文本的出现较为平稳;(2)出现热点话题时,微博话题讨论量呈上升趋势,在短时间内微博文本点赞数、转发数以及评论数显著增加,而一般话题没有这一特征。热点话题微博与一般话题微博示例分别如表4和表5所示。

表4 热点话题微博博示例

Table 4 Examples of hot topics microblog

微博文本	点赞数	评论数	转发数
近日,一些国家相继采取入境管制措施。	4 105	896	1 243
火神山医院,医疗配套工程全面展开。	770	42	880
防疫物资陆续到达,感谢国民力量。	6 425	16 293	5 123
防护从我做起,勤洗手,戴口罩	623	1 203	514
快讯! 周口漯河调研检查疫情防控工作	1 101	1 616	2 183

表5 一般话题微博示例

Table 5 Examples of general topics microblog

微博文本	点赞数	评论数	转发数
买火车票还要我提供手机号。	4	0	0
双子座,不要害怕,冲鸭……	12	3	0
糟了,糟了,又要加班,文案不过关啊啊	6	2	0
每天一个小Tip,感冒远离我,哈哈哈哈	18	5	1
打包送往新加坡家庭,加油	7	2	1

基于上述现象,提出在使用词对加速度定义突发特征时,应考虑微博文本强度(点赞数、评论数和转发数)的不同。

定义4 微博文本强度ts的大小是由评论数、点赞数和转发数确定的,定义为

$$ts = u + \lg(\text{comment\_n} + \text{suport\_n} + \text{transmit\_n} + 1) \tag{5}$$

式中:u表示用户影响力,comment_n、suport_n和transmit_n分别表示评论数、点赞数和转发数。

定义5 经过特征扩展得到的微博文本cx_i中任意特征词对表示为(w_j, w_k),词对的速度$v_{j,k}^{i,\Delta T}$定义为

$$v_{j,k}^{i,\Delta T}=\begin{cases}\dfrac{1}{\Delta T}\displaystyle\sum_{t_i-\Delta T\leqslant t_q\leqslant t_i} f_i(w_j,w_k)\times\dfrac{ts}{m_{ts}}\times\exp\left(\dfrac{t_q-t_i}{\Delta T}\right) & ts>0\\ \dfrac{1}{\Delta T}\displaystyle\sum_{t_i-\Delta T\leqslant t_q\leqslant t_i} f_i(w_j,w_k)\times\exp\left(\dfrac{t_q-t_i}{\Delta T}\right) & ts=0\end{cases} \tag{6}$$

式中:ΔT表示以t_i为终止时间点的时间片;ts表示微博文本强度;m_{ts}表示数据流中微博文本强度的最大值,并且ts和$\exp\left(\dfrac{t_q-t_i}{\Delta T}\right)$均为$f_i(w_j,w_k)$的权重,后者刻画了词对距离$t_i$时刻的远近程度,$f_i(w_j,w_k)$为微博$cx_i$中词对$(w_j,w_k)$的频率[23],计算式为

$$f_i(w_j,w_k)=\begin{cases}\dfrac{(n_i^{w_j})^2-n_i^{w_k}}{n_i(n_i-1)} & w_j=w_k\\ \dfrac{n_i^{w_j}n_i^{w_k}}{n_i(n_i-1)} & w_j\neq w_k\end{cases} \tag{7}$$

式中:$n_i^{w_j}$、$n_i^{w_k}$分别表示特证词w_j、w_k在微博文本中出现的次数;n_i为cx_i中所有特征词的个数。

定义6[24] 特征词对(w_j,w_k)在t_i时刻的加速度$a_{j,k}^i$可以利用两个时间片内的特征词对速度的变化来表示,即

$$a_{j,k}^i=\frac{v_{j,k}^{i,\Delta T_2}-v_{j,k}^{i,\Delta T_1}}{\Delta T_1-\Delta T_2} \tag{8}$$

式中$\Delta T_1<\Delta T_2$。

定义7 当词对加速度$a_{j,k}\geqslant\alpha$,则称(w_j,w_k)为突发词对,α表示词对加速度阈值。

微博文本流中的突发词对速度可以反应热点话题的热度,所以可根据其确定突发词对窗口。

定义8 突发词对窗口BWW定义为

$$\text{BWW}=\left\{(cx_i,t_i)\middle|v_{j,k}^{i,\Delta T_1}\geqslant\beta,t_i\in[t_a,t_b]\right\} \tag{9}$$

式中:$v_{j,k}^{i,\Delta T_1}$表示突发词对(w_j,w_k)速度,$[t_a,t_b]$表示从t_a到t_b的时间区间,并满足条件$v_{j,k}^{a-1,\Delta T_1}<\beta, v_{j,k}^{b+1,\Delta T_1}<\beta$。

通过对微博文本流中热点话题的分析,可发现在连续时间区间内出现的微博文本一般属于同一个热点话题。此外,还发现在一个时间区间内的多个突发词对窗口往往是交叉、重叠或者相邻的,所以可通过突发词对窗口来确定热点话题窗口。

定义9 将n个交叉、重叠或相邻的突发词对窗口BWW_1、BWW_2、…、BWW_n合并,定义为热点话题窗口W,可表示为

$$W=\bigcup_{i=1}^{n}\text{BWW}_i \tag{10}$$

3 基于特征扩展的热点话题检测

3.1 热点话题检测框架

本文提出的热点话题检测框架主要分为3个模块:微博短文本特征扩展、突发特征识别、热点话题窗口的确定和热点话题的主题结构,如图1所示。

(1)微博短文本特征扩展:首先基于评论文本的强度(粉丝数、关注数和点赞数)筛选评论文本;然后利用词共现与TF-IDF抽取评论文本中的特征词用来扩展微博文本。

(2)突发特征识别:通过文本强度给词对速度加权,从而计算词对加速度,并将其作为突发特征。当新到达微博时,计算微博中词对的频率、速度和加速度。当词对加速度大于阈值时,判定产生热点话题。

(3)热点话题窗口的确定和热点话题的主题结构表示:当词对加速度大于阈值时得到突发词对,通过突发词对的速度来判定突发词对窗口,合并满足条件的突发词对窗口进而确定热点话题窗口。利用GSDMM算法对热点话题窗口中的微博文本聚类,得到热点话题的主题结构。此聚类方法能够较好地处理高维、稀疏的短文本。

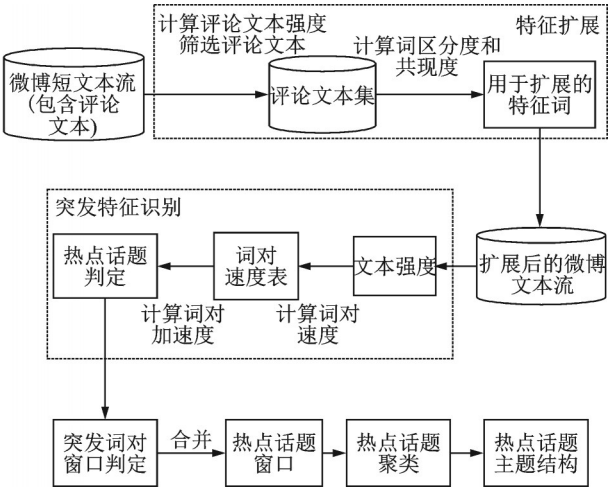

图1 基于特征扩展的热点话题检测框架

Fig.1 Framework of hot topic detection based on feature extension

3.2 基于特征扩展的热点话题检测算法

根据图1所示的特征扩展热点话题检测框架,提出了一种基于特征扩展的热点话题检测算法(Feature extension based hot topic detection,FE-HTD),如算法1所示。

算法1 基于特征扩展的热点话题检测算法

输入:微博文本流D,时间片ΔT_1和ΔT_2,文本强度阈值k,词对加速度阈值α,词共现度阈值μ,词区分度阈值ω,突发词对速度阈值β,窗口合并个数阈值n。

输出:热点话题主题结构。

①对t_i时刻出现的所有微博文本和对应的评论文本,计算文本强度,当评论文本强度cts$\geqslant\lambda$时,将该条评论文本作为扩展语料。

②计算所有微博文本中特征词w_s^i与所对应的评论文本中的特征词w_t^{ij}的词共现度$WA(w_s^i,w_t^{ij})$以及词区分度$d(w_t^{ij})$。

③若$WA(w_s^i,w_t^{ij})\geqslant\mu$或$d(w_t^{ij})\geqslant\omega$,则将特征词$w_t^{ij}$扩展到微博文本中,得到扩展微博文本$cx_i$。

④对扩展得到的微博文本cx_i,统计t_i时刻微博中所有词对出现的频率,计算词对的速度$v_{j,k}^{i,\Delta T_1}$,$v_{j,k}^{i,\Delta T_2}$和加速度$a_{j,k}^i$,最后更新词对速度表。

⑤若$a_{j,k}^i\geqslant\alpha$,则认为出现热点话题,并根据定义7得到突发词对(w_j,w_k)。

⑥根据定义8判断交叉、重叠或相邻的突发词对窗口个数,当大于n时,合并窗口从而得到热点话题窗口W。

⑦利用GSDMM算法对热点话题窗口W中的微博文本聚类,获取热点话题主题。

4 实验结果及分析

4.1 实验数据

面向微博文本流的热点话题检测无标准数据集，文中实验数据通过新浪微博(www.weibo.com)平台获取。采集了2020年1月中旬、2020年6月中旬以及2020年12月中旬3个时间段的微博短文本流作为测试数据集，共获取微博短文本33万余条，具体如表6所示。

首先对获取到的微博数据进行处理，保留微博的发布时间、内容信息、评论信息、点赞数、评论数、转发数、粉丝数以及关注数；然后对微博和评论信息进行分词、去除停用词以及删除噪声(如URL链接等)；最终得到32万条微博短文本及其相应的评论文本，经过对微博文本进行人工标注，共得到热点话题35个。

表6 测试数据集表示

Table 6 Representation of test data sets

名称	时间	微博条数/万	热点话题数
DB-F	1月10～20日	11.4	14
DB-S	6月10～20日	9.8	9
DB-T	12月10～20日	12.1	12

4.2 FE-HTD算法的准确率、召回率、F_1值

通过准确率P、召回率R、F_1值来判断算法性能，其计算公式分别为

$$P=\frac{\mathrm{RHT}}{\mathrm{ST}} \tag{11}$$

$$R=\frac{\mathrm{RHT}}{\mathrm{LT}} \tag{12}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{13}$$

式中：RHT指算法正确检测的热点话题数量；ST指算法检测出的热点话题数量；LT指人工标注的热点话题数量。

算法FE-HTD的参数取值：词区分度阈值$\omega=0.35$，评论文本强度阈值$\lambda=1.6$，词对加速度阈值$\alpha=0.15$，词共现度阈值$\mu=0.20$，突发词对速度阈值$\beta=3.0$，$\Delta T_1=15$ min，$\Delta T_2=30$ min，突发词对窗口的合并阈值$n=4$。由分析采集到的数据可发现，一些用户的粉丝数很大但是关注数很小；一些用户的粉丝数接近于关注数；粉丝数与关注数比值最小为0.47，最大为23 314。为缩小该范围，并使得用户影响力在合理区间，设定调节参数$\gamma(0<\gamma<1)$。实验中将γ取0.1。

为了验证微博短文本流特征扩展对热点话题发现的作用，将本文提出的FE-HTD方法与不经过特征扩展而直接进行热点话题检测的HTD方法进行对比实验，结果如表7所示。

从表7可知，对微博文本进行特征扩展以后，可以检测到一些仅利用微博文本发现不了的热点话题。因为特征扩展可以丰富微博短文本信息，比如由微博短文本“我最爱的综艺没有之一”可以得到的特征为“最爱，综艺”，利用评论信息进行特征扩展以后成为“最爱，综艺，偶像，练习生”，可知该微博文本与话题“偶像练习生开播一周年”相关，从而提高了热点话题的识别率。

为了验证本文所提出的FE-HTD算法的性能，将其与其他3种算法进行了比较。第1种算法为上文提到的HTD；第2种为BW-TD[25]，该算法基于定长滑动窗口，将词频的变化作为突发特征，通过聚类得到突发话题；第3种算法为T-TD[26]，该算法利用词共现对文本进行扩展，并使用LDA模型检测突发话题。图2为4种算法在数据集DB-F、DB-S和DB-T上热点话题检测的P、R、F_1值的对比结果。由图2可知，在3个实验数据集上，本文所提FE-HTD算法的P、R、F_1值均高于对比算法。其原因在于BW-TD算法采用定长滑动窗口，通过实验数据可知，有的话题持续时间仅有47 min，将时间窗口设置为3 h，窗口可能会将话题切分开，导致特征词频变化率未达到阈值造成话题漏检，从

表7 部分话题检出情况

Table 7 Detection of some topics

热点话题	FE-HTD 是否检出	HTD 是否检出
2020年春运大幕拉开	是	否
深圳国企喝茅台	是	是
各大股市连续下跌	是	是
NBA湖人对雷霆	是	是
中美第一阶段经济贸易协议	是	是
偶像练习生开播一周年	是	否
斗鱼虎牙合并案被审查	是	是
暗访县医院医保乱象	是	否
摩拜单车已全面接入美团	是	是
杭州一女生感染狂犬病脑死亡	否	否
多地出现自称辛巴客服形式诈骗	否	是

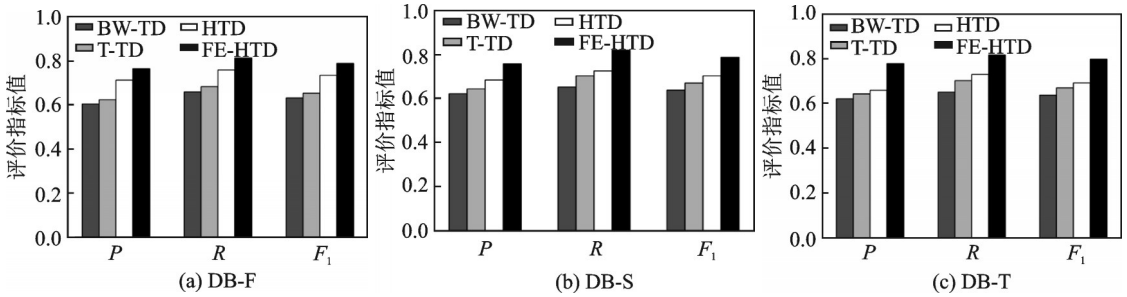

图2 4种检测算法在不同数据集上的P、R、F_1值对比

Fig.2 Comparison of P, R and F_1 values of four detection algorithms on different data sets

而热点话题检测查准率降低;算法HTD未考虑特征扩展,算法T-TD虽对文本进行特征扩展,但是在扩展的特征词质量上没有做筛选,因此导致扩展质量不高,话题检测准确度低。本文所提算法优于对比算法的主要原因在于,在特征扩展时考虑了文本强度,在检测突发特征时不但考虑了词对加速度,还考虑了微博文本点赞数、评论数、转发数以及发表微博文本用户的影响力。

4.3 FE-HTD算法时效性

为了检测本文所提算法的时效性,将FE-HTD与做了特征扩展的T-TD算法和未经特征扩展的BW-TD算法作对比实验。在表8中列举了3个热点话题的检出时间,热点话题1~3出现相关微博的时间分别为(12:37:53、13:09:03、13:41:27)。

由表8可知,本文提出的FE-HTD算法和T-TD算法相比较,FE-HTD算法提前30 min检测出热点话题。这主要是因为本文算法基于动态窗口,对微博文本实时计算特征词对加速度和词对速度,并在突发特征定义时利用文本强度为特征词对加权,当词对加速度大于阈值时则认为存在热点话题。比如实验中微博文本"万众一心,众志成城"经过特征扩展以后变成"万众一心,众志成城,新冠,疫情,隔离,医护",可以发现特征词对(新冠、疫情)出现在原始微博中,导致特征词对速度变大,从而快速检出热点话题。T-TD算法根据词共现选取评论,随后利用主题模型输出话题,最后聚类判断话题簇的突发状

表8 话题检测时效性描述

Table 8 Effectiveness description of topic detection

热点话题	检出时间		
	T-TD	BW-TD	FE-HTD
话题1:斗鱼虎牙合并案被审查	13:32:22	12:59:56	13:01:24
话题2:摩拜单车已全面接入美团	14:31:47	13:36:27	13:42:55
话题3:暗访县医院医保乱象	14:58:55	14:01:33	14:06:37

态,主题模型参数估计比较耗时,并且该算法采用定长窗口,只有当一个窗口结束后才进行词频分析,因此导致检测出热点话题的时间滞后。本文算法与对比算法BW-TD相比检出时间较为接近,平均滞后5 min,主要是因为本文在话题检测前做了特征扩展。若不考虑特征扩展所需时间,对比算法BW-TD检出热点话题的时间滞后。

4.4 特征扩展的热点话题主题结构

为了得到热点话题的主题结构,本文利用GSDMM算法对热点话题窗口中的微博文本进行聚类。表9列举了3个热点话题对应不同算法检出特征词的情况。分析“司机救婴儿闯红灯家属拒绝作证”这一热点话题,FE-HTD算法检出5个突发词对,聚类分析热点话题窗口中的微博文本,得到主题结构:(1)司机闯红灯家属拒绝作证;(2)建议家属拉入征信系统。可以看出对比算法仅有少量的突发词检出,而本文算法可检出较多突发词对,所以主题结构比较丰富。

表9 热点话题主题结构

Table 9 Theme structure of hot topics

热点话题	T-TD	BW-TD	FE-HTD	
	突发词	突发词	突发词对	主题结构
暗访县医院医保乱象	免费 医院 骗保 严查	医院 医保 骗保	(免费,住院)(医院,医保)(医院,骗保)(严查,骗保)(医院,免费)	1.免费 住院 2.严查 骗保
司机救婴儿闯红灯家属拒绝作证	司机 闯红灯 拒绝 作证 系统	司机 闯红灯 拒绝 作证	(司机,闯红灯)(司机,婴儿)(家属,拒绝)(拒绝,作证)(征信,系统)	1.司机 闯红灯 婴儿 家属 拒绝作证 2.家属 征信 系统
摩拜单车已全面接入美团	摩拜 单车 停止 接入 美团	摩拜 单车 接入 美团	(摩拜,单车)(停止,服务)(摩拜,余额)(美团,权益)(接入,美团)	1.摩拜 单车 停止 运营 2.余额 权益 接入 美团

4.5 参数μ对算法P、R、F_1值的影响

为分析词共现度阈值μ对所提FE-HTD算法的P、R、F_1值的影响,分别在DB-F、DB-S、DB-T数据集上进行了实验。本文在0.15~0.25之间取不同的共现度阈值μ进行实验,结果如表10所示。

由表10可知,随着μ值的增大,算法的召回率下降。这是因为μ值的增大导致得到的特征词有限,一些可以用作特征扩展的词被过滤掉,使得算法召回率变小。通过上述实验数据可知,当μ为0.20时,算法准确率达76.4%,召回率达78.7%,实验结果最优。

表10 不同μ值对算法的P,R,F_1值的影响

Table 10 Influence of different values of μ on P, R and F_1 values of the algorithm

数据集	μ	P	R	F_1
DB-F	0.15	0.642	0.671	0.656
	0.20	**0.752**	**0.791**	**0.774**
	0.25	0.762	0.579	0.658
DB-S	0.15	0.558	0.667	0.608
	0.20	**0.776**	**0.816**	**0.795**
	0.25	0.807	0.542	0.648
DB-T	0.15	0.589	0.641	0.613
	0.20	**0.764**	**0.812**	**0.787**
	0.25	0.816	0.473	0.598

5 结束语

本文利用微博文本、评论文本和用户信息提出了一种基于特征扩展的微博短文本流热点话题检测方法。首先基于文本强度筛选评论文本,计算词共现度和词区分度提取特征词,通过文本强度和词对加速度定义突发特征,然后根据突发词对的速度确定可变长的热点话题窗口范围,最后聚类得到窗口中热点话题的主题结构。基于新浪微博平台采集的真实数据,将所提算法与对比算法进行实验,验证了所提算法的有效性和准确性。本文所提方法可以用于微博短文本流的热点话题发现,对社交网络中信息推荐和舆情监控等具有支撑作用。

参考文献:

[1] 曹中华,夏家莉,彭文忠,等.多原型词向量与文本主题联合学习模型[J].中文信息学报,2020,34(3):64-71.
CAO Zhonghua, XIA Jiali, PENG Wenzhong, et al. Multiple prototype word vectors and text topic joint learning model[J]. Journal of Chinese Information Processing, 2020, 34(3): 64-71.

[2] LI J, TAI Z, ZHANG R, et al. Online bursty event detection from Microblog[C]//Proceedings of ACM International Conference on Utility & Cloud Computing. Piscataway:ACM, 2014: 865-870.

[3] 董薇,庞峰,顾炜江.基于LDA模型的大规模文本挖掘算法研究[J].软件,2020,41(12):58-63.
DONG Wei, PANG Feng, GU Weijiang. Research on large-scale text mining algorithm based on LDA model[J]. Software, 2020, 41(12): 58-63.

[4] 伊秀娟.基于LDA主题模型的高校新闻话题发现研究[D].北京:北京交通大学,2019.
YI Xiujuan. Research on university news topic discovery based on LDA topic model[D]. Beijing: Beijing Jiaotong University, 2019.

[5] KIEJIN P.A design on informal big data topic extraction system based on spark framework[J]. KIPS Transactions on Software and Data Engineering, 2016, 5(11): 521-526.

[6] MEHROTRA R, SANNER S, BUNTINE W, et al. Improving LDA topic models for microblogs via tweet pooling and automatic labeling[C]//Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval. [S.l.]: ACM, 2013: 889-892.

[7] WANG Jing, LI Li, TAN Feng, et al. Detecting hotspot information using multi-attribute based topic model[J]. Plos One, 2017, 10(10): e0140539.

[8] 周先琳.基于动态Labeled-LDA模型的微博主题挖掘[D].合肥:合肥工业大学,2015.
ZHOU Xianlin. Microblog topic mining based on dynamic Labeled-LDA model[D]. Hefei: Hefei University of Technology, 2015.

[9] FUNG G P C, YU J X, YU P S, et al. Parameter free bursty events detection in text streams[C]//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, Norway: [s.n.], 2005: 181-192.

[10] 李汉才,徐建民,吴树芳.融合时序性和波动性的热点话题发现研究[J].河北大学学报(自然科学版),2018,38(4):416-422.
LI Hancai, XU Jianmin, WU Shufang. Research on hot topic discovery combining temporal sequence and fluctuation[J]. Journal of Hebei University (Natural Science Edition), 2018, 38(4): 416-422.

[11] 万越,隋杰.基于用户行为影响的微博突发话题检测方法[J].中国科学技术大学学报,2017, 47(4): 328-335.
WAN Yue, SUI Jie.A new method for detecting topics in microblogs based on user behavior[J].Journal of University of Science and Technology of China,2017, 47(4): 328-335.
[12] 郑斐然,苗夺谦,张志飞, 等.一种中文微博新闻话题检测的方法[J].计算机科学,2012,39(1): 138-141.
ZHENG Feiran, MIAO Duoqian, ZHANG Zhifei, et al. A method for topic detection of Chinese microblog news[J].Computer Science,2012,39(1): 138-141.
[13] 蔡莹,於跃成,谷雨.基于时间窗口包含用户行为的微博突发话题检测方法[J].计算机与数字工程,2020,48(2): 383-388.
CAI Ying, YU Yuecheng, GU Yu. A method for breaking topic detection of microblog based on time window including user behavior[J]. Computer and Digital Engineering,2020,48(2): 383-388.
[14] LI Junze, CAI Yi, CAI Zhiwei, et al. Wikipedia based short text classification method[C]//Proceedings of 22nd International Conference on Database Systems for Advanced Applications. Suzhou, Chinese:IEEE,2017: 275-286.
[15] GAO Longwen, ZHOU Shuigeng, GUAN Jihong. Effectively classifying short texts by structured sparse representation with dictionary filtering[J].Information Sciences,2015,323(6): 130-142.
[16] CHENG Xueqi, YAN Xiaohui, LAN Yanyan, et al. BTM: Topic modeling over short texts[J].IEEE Transaction on Knowledge and Data Engineering, 2014,26(12): 2928-2941.
[17] LI Peipei, HU Xuegang, ZHANG Yuhong, et al. Learning from short text streams with topic drifts[J].IEEE Transactions on Cybernetics,2018,48(9): 2697-2711.
[18] DUAN Yu, LI Chenliang, WANG Haoran, et al. Enhancing topic modeling for short texts with auxiliary word embeddings[J]. ACM Transactions on Information Systems, 2017,36(2): 1-30.
[19] PAULO B, MARCELO P, ANISIO L, et al. A general framework to expand short text for topic modeling[J]. Information Sciences, 2017,393(2): 66-81.
[20] 张萌.微博热点话题发现方法的研究和实现[D].北京:北京交通大学,2018.
ZHANG Meng.Research and implementation of microblog hot topic discovery method[D]. Beijing: Beijing Jiaotong University, 2018.
[21] LIU C Y, CHEN M S, TSENG C Y, et al.INCRESTS: Towards real-time incremental short text summarization on comment streams from social network services[J]. IEEE Transactions on Knowledge and Data Engineering, 2015,27(11): 2986-3000.
[22] 金宇杰,袁明.基于TF-IDF算法的新词发现系统原理与实现[J].信息化研究, 2020,46(5): 39-44.
JIN Yujie, YUAN Ming. Principle and implementation of new word discovery system based on TF-IDF algorithm[J].Information Research, 2020,46(5): 39-44.
[23] 李艳红,贾丽娜,王素格, 等.基于动态窗口的微博突发话题检测方法[J].计算机应用与软件, 2020,37(5): 31-37.
LI Yanhong, JIA Lina, WANG Suge, et al. A method for detecting breaking topics of microblog based on dynamic window[J]. Computer Application and Software, 2020,37(5): 31-37.
[24] HE D, PARKER D S. Topic dynamic: An alternative model of bursts in streams of topics[C]//Proceedings of the 2010 16th ACM International Conference on Knowledge Discovery and Discovery Data Mining. New York:ACM,2010: 443-452.
[25] 魏景璇.基于突发词共现的微博突发话题检测[J].滨州学院学报,2020, 39(1): 138-141.
WEI Jingxuan. Bursting topic detection in microblog based on bursting word co-occurrence[J]. Journal of Binzhou University, 2020, 39(1): 138-141.
[26] LEI Shi, GANG Cheng, XIE Shangru, et al. A word embedding topic model for topic detection and summary in social networks[J]. Measurement and Control,2019,52(10): 1289-1298.

作者简介:

李艳红(1977-),**通信作者**,女,博士,副教授,研究方向:数据挖掘、机器学习,E-mail: liyh@sxu.edu.cn。

谢梦娜(1996-),女,硕士研究生,研究方向:数据挖掘。

王素格(1964-),女,博士,教授,研究方向:自然语言处理、机器学习。

李德玉(1965-),男,博士,教授,研究方向:人工智能、数据挖掘。

(编辑:王静)