文章编号: 1001-0920(2019)07-1409-08 DOI: 10.13195/j.kzyjc.2017.1640

基于多模态特征深度融合的微博流事件检测与跟踪

熊 宇[1], 张一飞[1,2], 冯 时[1,2], 王大玲[1,2†]

(1. 东北大学 计算机科学与工程学院，沈阳 110169；2. 教育部 医学影像计算重点实验室，沈阳 110169)

摘 要: 作为一种重要的社会媒体平台,分析、检测并跟踪微博内重大社会事件可以及时提供舆论焦点. 但因其碎片化、异构性和实时性,传统方法很难有效分析海量微博,为此,提出一种基于多模态特征深度融合的微博事件检测与跟踪框架. 首先基于文本处理对微博事件进行标注;然后用多模态特征深度融合实现事件的检测与表示;最后利用基于时间平滑的图变换模型完成事件流的跟踪. 在真实数据集上的实验表明,所提出的方法能有效检测和跟踪微博流事件.

关键词: 微博流；事件检测；事件跟踪；多模态；特征融合；深度学习

中图分类号: TP273 **文献标志码:** A

Event detection and tracking in microblog stream based on multimodal feature deep fusion

XIONG Yu[1], *ZHANG Yi-fei*[1,2], *FENG Shi*[1,2], *WANG Da-ling*[1,2†]

(1. College of Computer Science and Engineering，Northeastern University，Shenyang 110169，China；2. Key Laboratory of Medical Image Computing，Ministry of Education，Shenyang 110169，China)

Abstract: As an important social media platform, analyzing, detecting and tracking the important social events in microblog can provide public issues in time. However, due to the fragmentation, heterogeneity and real-time characteristics of microblog, traditional techniques can hardly analyze mass microblog efficiently. Therefore, a social event detection and tracking framework based on multimodal feature deep fusion is proposed. Firstly, in the framework, events in microblogs are labeled by text process. Then, the detection and description of events are achieved by multimodal feature deep fusion. Finally, the tracking of the event stream is accomplished by the graph variation based on time smooth. The experiments in a real dataset show that the proposed method can detect and track events in the microblog stream effectively.

Keywords: microblog stream；event detection；event tracking；multimodality；feature fusion；deep learning

0 引 言

作为最重要的社会媒体,微博已成为人们发布和获取信息的主要途径之一. 据统计,新浪微博2012年的数据量已达每天1亿,而推特2017年每天的数据量高达5亿. 通过海量信息中的大量社会事件,民众能及时获取社会焦点,而政府可以在必要时刻引导舆论. 因此,针对社会媒体的事件检测与跟踪对决策者和普通民众均有重要意义.

虽然事件检测与跟踪已研究多年并取得大量成果,但仍面临严峻挑战. 首先,海量微博的爆炸式增长使得针对静态数据的传统方法难以及时从数据海洋中提取有效信息;其次,多模态数据(如文本、视频和图片)的特征属于不同空间,无法直接用于分析模态之间的相似性,从而不能直接为社会事件的准确、全面分析提供帮助;最后,微博文本短、语法混乱、信息散乱、存在大量噪音,将进一步降低传统方法的性能. 因此,传统的事件检测跟踪技术已经难以高效地分析和处理微博数据.

本文以数量急剧增长、内容不断演变、充满噪音的微博多模态数据流作为输入,动态地检测并跟踪广泛讨论或突发性社会事件. 考虑到微博的文本短、变化快、增量多等特征,提出一种基于文本处理技术的事件快速检测方法,包括关键词提取、事件图构建、图聚类、微博文本事件标注;基于微博的多模态特征,提出一种基于多模态深度融合的事件细粒度检测表达方法,通过学习事件的多模态特征,从而准确表达完整事件;由于事件会不断演化,提出一种基于时间平滑的事件跟踪方法,通过缓和事件图的演化、事件相

收稿日期: 2017-12-04；**修回日期:** 2018-04-04.
基金项目: 国家自然科学基金项目(61772122).
责任编委: 刘民.
†通讯作者. E-mail: wangdaling@cse.neu.edu.cn.

似度实时计算,实现故事线内事件拼接. 在真实微博流上的实验验证了所提出方法的有效性,且与当前最新工作的对比实验表明,所提出的方法更适合检测与跟踪微博流的社会事件.

1 相关工作

1.1 社会事件检测与跟踪

社会事件检测与跟踪以事件为基本要素,对社会媒体数据进行分析、检测、跟踪、存储等操作. 该研究起源于传统的话题检测和跟踪模型,需要处理社会媒体数据流中的大规模噪音和歧义信息. 2010年,Sakaki等[1]以人作为社会事件的传感器,实现对日本地震位置和台风轨迹的检测跟踪. 这种理念导致当前社会事件检测与跟踪方法以语义模型[2-6]、图模型[7-10]和启发式模型[6,11-14]3种类型为主.

语义模型通常以潜在狄利克雷分布(LDA)为基础,对微博进行有监督或半监督的学习. 多模态LDA[2]、位置-时间限制话题模型[3]、多模态时空话题模型[4]、多模态事件话题模型[5]以及基于梗概的话题模型[6]等,均利用LDA对微博(或推特)的位置、时间、文本、图片、哈希标签等信息进行建模,推导话题的分布并进行动态维护. 它们仅用少量超参数便能学习模型,且对微博进行专门优化,训练效果好,但不能摆脱话题模型的固有缺陷,如数据集必需良好、训练与测试数据要同分布等.

图模型通过图结构(传统有向或无向图、异构图、二元图、超图等)将与社会事件相关要素(用户、位置、标签、微博等)关联起来,利用图聚类、图切割和图关联等理论,实现社会事件的检测、跟踪、预测等应用. 多模态话题与或图[9]以及推特异构图[7],采用图结构组织事件的人物、位置、标签、时间等信息,通过图切割或混合聚类等方式达到检测跟踪事件的目的. Schinas等[8]通过滑动窗口维护事件照片的多模态关联图,利用图聚类检测事件. 在图模型中,新型知识图谱能更准确高效地描述事件要素的复杂联系,从中可以挖掘更丰富的信息. 推特异构图[7]、多模态话题与或图[9]、话题敏感的意见领袖挖掘框架[10]都采用异构图或超图将推特转发、回复、提及等操作与用户、图片、社会关系等概念组织起来,通过多种图算法如信任传播、节点排序、图谱张量化等,实现更广泛的应用. 这些模型可以用大量成熟的图理论实现事件的检测与跟踪,但为了避免计算复杂度过高,需要控制其规模,因此通常不适合大规模数据的处理,且建立大规模知识图谱仍需要大量人工干预.

启发式模型通过观测事件要素的变化并结合传统的聚类分类算法,实现社会事件的检测跟踪. 常见的事件要素有词频[6,12-13]、用户关系[14]、微博内容[13]等,而事件检测跟踪算法既可以采用传统的文本聚类和分类算法[12],也可以是图关联模型[13]或话题模型[6]. 这些方法通常是词频检测与传统聚类分类方法的结合,事件类型不受限制,且拥有部分抗噪能力,可以用于大数据,但是由于它们大多属于启发式方法,一般不能直接扩展到其他应用中.

1.2 多模态数据融合

多模态数据是指两种或多种类型不兼容的数据组合,在本文特指不同的媒体形式. 多模态数据的不兼容性是指其特征无法直接用于分辨语义相似和不相似的数据[15]. 多模态数据融合的目的是为数据生成全面、准确、生动的描述. 虽然目前已有多模态文档模型[16]、异构图模型[17-18]等研究成果,但难以运用在充满噪音、数据不断变化的社会媒体上. 目前,基于社会媒体的数据融合方法主要有多模态哈希[19-21]和多模态深度学习[22-24],前者为了获取多模态编码,后者为了训练多模态融合模型.

多模态哈希是指根据语义的相似性,采用哈希映射算法对互不兼容的数据特征进行编码,保证语义相似的数据其特征编码也相似. 在多模态映射过程中,通常需要最小化损失函数或相似数据的哈希编码距离[19-20]. 此外,Xie等[21]还将跨模态哈希映射分解为共享潜在编码和动态迁移两个矩阵,通过增量式更新这两个矩阵,确保已学习到的跨模态数据哈希编码不会改变. 这些方法从大规模数据中学习表达能力强的多模态特征,一般用于数据检索,不会直接进行多模态数据聚类或分类.

多模态深度学习是指根据语义的相似性,采用深度学习算法学习不同数据的统一特征,从而生成多模态融合模型. 如多模态栈式自动编码机[22]、基于成对关系的跨模态哈希编码算法[24],都在学习多模态融合模型的过程中,最小化模态内和模态间的重构误差. 此外,Cao等[23]为了学习高质量的深度视觉语义哈希,将图片的卷积神经网络与文本的长短期记忆网络相结合,通过调节隐藏层的规模改善学习效果,能一边学习特征一边分类数据,但要求数据集良好,且学习参数多效率低下.

2 面向微博流的社会事件检测与跟踪

本文方法的整体框架包括事件检测和事件跟踪两部分. 事件检测包括基于文本处理的事件标注和基于特征融合的事件检测,其中文本处理涉及从微博文本流中提取关键词、构建事件图、图聚类,而特征融

合涉及单模态的特征学习、多模态的特征融合、基于多模态融合的事件表达. 事件跟踪包括基于时间的事件图平滑、基于平滑图的事件检测、事件相似度计算、故事线形成. 为统一起见,文中小写斜体(如x),大写斜体(如X)、小写粗体(如$\boldsymbol{x}$)、大写粗体(如$\boldsymbol{X}$)分别表示元素或标量、集合、元素特征向量、集合特征矩阵. 表1为本文所用符号.

表 1　微博流事件检测与跟踪的符号说明

符号	描述
m、M	微博、微博集
x、X	微博文本、微博文本集
y、Y	微博图片、微博图片集
t、τ、T	某天、时间窗口长度、数据集时间跨度
c、C	事件、事件集
s、S	故事线、故事线集
v、V	词语、词语集
df、idf、dfidf	词语或词组的文档频率(DF)、倒文档频率(IDF)、以及DF-IDF值
r	词语之间的相关度
G、e、E	事件图、边权重、边权重集
sim_{mc}、$\theta_{\mathrm{sim}}^{mc}$	微博与事件的相似度、微博属于某事件的相似度阈值
sim_{cc}、$\theta_{\mathrm{sim}}^{cc}$	事件之间的相似度、相同故事线中事件相似度阈值
L^0	多模态特征融合层拟合函数
L^1	多模态事件结果表达层拟合函数
$\boldsymbol{W}$、$\boldsymbol{b}$	深度模型的权重矩阵和偏置向量

由于社会事件通常按天发生,本文以天为时间单位[8]. 设一条微博为$m=\langle x,y\rangle$,x、y分别为其文本和图片. 受平台限制,每条微博有且仅有一条文本且最多只有一张图片,则第t天的微博流记为$M^t=\langle X^t,Y^t\rangle$,其中$X^t$、$Y^t$分别为当天微博的所有文本和图片. 因此,时间长度为T天的微博流可表示为$\{M^t|t\in[1,T]\}$. 在微博流中,一个社会事件c是指一天内拥有相同语义的所有微博集合,记作$c=\{m|m^c=c, t_m=t_c\}$. 其中:m^c为微博的事件标注,t_m、t_c为微博、事件的时间戳. 一条故事线s^t是指截止到第t天拥有相似语义的所有事件集合,记作$s^t=\{c|c^s=s^t, t_c\in[1,t]\}$,其中$c^s$为事件的故事线标识. 如果第$t$天的所有社会事件为$C^t=\{c_1^t,c_2^t,\cdots\}$,且$S^t=\{s_1^t,s_2^t,\cdots\}$为截止至第$t$天的所有故事线,则对于第$t$天的微博流,其输出应为当天的所有事件与故事线,即$O(M^t)=C^t\bigcup S^t$. 以下将对本文提出的社会事件检测与跟踪方法进行详细介绍.

2.1　基于文本处理技术的事件标注

2.1.1　关键词抽取

为检测一般性社会事件,先从微博中提取每天与事件密切相关的关键词. 传统TF-IDF方法难以区分微博短文本中的关键词与普通词语,因此本文采用文档频率-倒文档频率(DF-IDF)算法提取每天的关键词. 对于广泛探讨的事件,词语的DF会变大;对于突发事件,词语的IDF会激增. 总之,该算法能准确地从微博文本流中提取关键词,即

$$\mathrm{dfidf}_v^{t,\tau}=\mathrm{df}_v^t\times\log\left(1+\frac{1}{\overline{\mathrm{df}}_v^{t,\tau}}\right). \tag{1}$$

式(1)为第t天词语v的DF-IDF,其DF为$|X_v^t|/|X^t|$,X_v^t为当天含有词语v的所有微博文本. 不同于传统TF-IDF,本文IDF只限于近期微博,表示第$t-\tau\sim t$天内词语v的平均DF. 由于鲜有社会事件能够被持续关注两周以上,词语IDF的时间窗口长度τ设为14,并通过实验验证其效果,基于该值,DF-IDF较大的词语即为当天的关键词,记作V^t. 实验表明,该值排前4.5 %的词语足以在微博流中进行事件聚类.

2.1.2　事件图构建与图聚类

为快速获取每日事件,利用关键词构建事件图并对其聚类,通过类的关键词定位事件. 在事件图中,节点为关键词,边为词语间的相关度. 由于微博噪音多,传统词频方法不能准确度量词语关系,因此本文依据词语和词组的重要度计算词语间相关度,即当词组重要度高于词语单独出现的重要度时,词组内词语的相关度应较高,反之应较低. 具体地,类似于$\mathrm{dfidf}_v^{t,\tau}$,词组$\langle v_i,v_j\rangle$重要度为

$$\mathrm{dfidf}_{ij}^{t,\tau}=\mathrm{df}_{ij}^t\times\log\left(1+\frac{1}{\overline{\mathrm{df}}_{ij}^{t,\tau}}\right). \tag{2}$$

其中:df_{ij}^t为第t天包含词组的微博文本频率,$\overline{\mathrm{df}}_{ij}^{t,\tau}$为第$t-\tau\sim t$天包含词组的平均微博文本频率. 基于词语和词组的重要度,词组中词语v_i、v_j相关度为

$$r_{ij}^t=\frac{\mathrm{dfidf}_{ij}^{t,\tau}}{(\mathrm{dfidf}_i^{t,\tau}+\mathrm{dfidf}_j^{t,\tau})/2}. \tag{3}$$

该值为词组重要度与词语平均重要度的比值,代表两个词语应为一个词组整体($r_{ij}^t\geqslant 1$),还是应为个体独立存在($r_{ij}^t<1$). 微博文本经常包含大量与事件无关的词语,尽管其DF-IDF有时较高,但与其他词语几乎无关联,式(3)可以有效过滤这些词语.

根据词语相关度,本文构建事件图并对其聚类,每个类即为一个事件. 由于微博流的数据量大且速度快,事件检测必须高效,本文在图模型上对关键词进行聚类. 具体地,第t天的事件图记为$G^t=\langle V^t,E^t\rangle$,$v\in V^t$为当天的关键词集合,$e_{ij}^t\in E^t$为归一化后词语间相关度集合,表示为

$$e_{ij}^t=\begin{cases}\tanh r_{ij}^t,\ v_i\in\mathrm{NN}_{v_j}^t\ \text{or}\ v_j\in\mathrm{NN}_{v_i}^t;\\ 0,\ \text{otherwise.}\end{cases} \tag{4}$$

其中NN_v^t为v在V^t内的所有最近邻. 该构建方式可以降低图的复杂度,且能提高事件检测的准确度. 图聚类采用基于密度的SCAN算法[25],该算法将G^t中

至少拥有μ个邻居且与之相似度超过ε的关键词节点作为核心点. 每次把一个未处理的核心点作为新类,并拓展该类至与其节点的相似度超过ε的所有邻居上,直至无邻居符合相似度要求. 因为G^t中节点仅被扫描一次,所以非常高效. 为准确描述事件内容,事件类的核心点至少要有两个邻居($\mu \geqslant 2$),且枢纽点由相连的多个事件类共享.

2.1.3 基于事件图的事件标注

基于事件图中事件类的关键词,利用微博与事件的文本相似度,实现微博的事件标注. 具体地,微博m与事件c的相似度由其共享词组来衡量,有

$$\mathrm{sim}_{mc} = \sum_{v_i,v_j \in \{m_x \bigcap c_v\}} e_{ij}^t, \tag{5}$$

其中c_v为类c在G^t中的所有关键词. 当该值最大且超过阈值$\theta_{\mathrm{sim}}^{mc}$时,$m$属于$c$,反之不属于任何事件. 通过下式可以实现对所有微博的事件标注:

$$m^c = \begin{cases} c,\ \mathrm{sim}_{mc} > \max\{\theta_{\mathrm{sim}}^{mc}, \mathrm{sim}_{mc'} | c' \in C \backslash c'\}; \\ \varnothing,\ \text{otherwise}. \end{cases} \tag{6}$$

2.2 基于多模态特征融合的事件检测

由于微博含有大量与事件相关的图片,为提高事件检测的准确性,基于事件标注,本文提出一种多模态特征深度融合模型,学习事件的多模态特征表达. 该模型由低到高依次为单模态的特征学习、多模态的特征融合和基于多模态融合的事件表达.

单模态的特征学习以微博文本、图片为输入,利用深度模型分别学习各自的语义特征. 文本与图片语义表达的本质差异,导致其学习模型也不同. 对于图片,深层的卷积神经网络(CNN)才能有效提取其语义特征;对于短文本,浅层CNN即可学习其语义信息. 因此,本文先用不同的深度模型单独学习各模态语义特征,再进行特征融合.

多模态的特征融合通过融合微博文本和图片的语义特征,生成鲁棒性更好、表达能力更强的多模态融合特征. 在融合时,学习出来的文本和图片融合特征需要尽量减小彼此间差异,即

$$L^0 = \sum_{\langle \boldsymbol{x},\boldsymbol{y}\rangle \in \langle \boldsymbol{X},\boldsymbol{Y}\rangle} \|(\boldsymbol{W}_x^0 \boldsymbol{x} + \boldsymbol{b}_x^0) - (\boldsymbol{W}_y^0 \boldsymbol{y} + \boldsymbol{b}_y^0)\|_2^2 + \delta_w^0(\|\boldsymbol{W}_x^0\|_2^2 + \|\boldsymbol{W}_y^0\|_2^2) + \delta_b^0(\|\boldsymbol{b}_x^0\|_2^2 + \|\boldsymbol{b}_y^0\|_2^2). \tag{7}$$

其中:$\boldsymbol{X}$、$\boldsymbol{Y}$分别为所有微博的融合后文本和图片语义特征矩阵,$\boldsymbol{W}^0$、$\boldsymbol{b}^0$为多模态特征融合参数,δ_w^0、δ_b^0为调节参数. 为最小化L^0,采用梯度下降算法迭代更新L^0中的参数$\boldsymbol{W}^0$和$\boldsymbol{b}^0$,即

$$\begin{aligned}
&\boldsymbol{W}^{0^{l+1}} = \boldsymbol{W}^{0^l} - \eta \nabla_{\boldsymbol{W}} L^0,\ \boldsymbol{b}^{0^{l+1}} = \boldsymbol{b}^{0^l} - \eta \nabla_{\boldsymbol{b}} L^0, \\
&\nabla_{\boldsymbol{W}_x} L^0 = \\
&2\boldsymbol{W}_x^0(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \delta_W^0 \boldsymbol{I}) + 2(\boldsymbol{b}_x^0 - \boldsymbol{W}_y^0 \boldsymbol{Y} - \boldsymbol{b}_y^0)\boldsymbol{X}^{\mathrm{T}}, \\
&\nabla_{\boldsymbol{W}_y} L^0 = \\
&2\boldsymbol{W}_y^0(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} + \delta_W^0 \boldsymbol{I}) - 2(\boldsymbol{b}_x^0 + \boldsymbol{W}_x^0 \boldsymbol{X} - \boldsymbol{b}_y^0)\boldsymbol{Y}^{\mathrm{T}}, \\
&\nabla_{\boldsymbol{b}_x} L^0 = 2(\boldsymbol{W}_x^0 \boldsymbol{X} - \boldsymbol{W}_y^0 \boldsymbol{Y} - \boldsymbol{b}_y^0) + 2(\delta_b^0 + 1)\boldsymbol{b}_x^0, \\
&\nabla_{\boldsymbol{b}_y} L^0 = 2(\boldsymbol{W}_y^0 \boldsymbol{Y} - \boldsymbol{W}_x^0 \boldsymbol{X} - \boldsymbol{b}_x^0) + 2(\delta_b^0 + 1)\boldsymbol{b}_y^0.
\end{aligned} \tag{8}$$

其中:η为学习率,l为迭代次数.

基于多模态融合的事件表达以文本和图片的融合语义特征为输入,通过函数L^1拟合微博的事件标注,从而得到事件的多模态表示模型,记为

$$L^1 = \sum_{\langle \boldsymbol{x},\boldsymbol{y}\rangle \in \langle \boldsymbol{X},\boldsymbol{Y}\rangle} \|\boldsymbol{m}^c - (\boldsymbol{W}_x^1 \boldsymbol{x} + \boldsymbol{b}_x^1) - (\boldsymbol{W}_y^1 \boldsymbol{y} + \boldsymbol{b}_y^1)\|_2^2 + \delta_w^1(\|\boldsymbol{W}_x^1\|_2^2 + \|\boldsymbol{W}_y^1\|_2^2) + \delta_b^1(\|\boldsymbol{b}_x^1\|_2^2 + \|\boldsymbol{b}_y^1\|_2^2), \tag{9}$$

其中$\boldsymbol{m}^c$为微博的事件标注向量表示. 与L^0相同,为最小化L^1,本文同样利用梯度下降算法迭代更新L^1中的参数$\boldsymbol{W}^1$和$\boldsymbol{b}^1$,即

$$\begin{aligned}
&\boldsymbol{W}^{1^{l+1}} = \boldsymbol{W}^{1^l} - \eta \nabla_{\boldsymbol{W}} L^1, \\
&\boldsymbol{b}^{1^{l+1}} = \boldsymbol{b}^{1^l} - \eta \nabla_{\boldsymbol{b}} L^1, \\
&\nabla_{\boldsymbol{W}_x} L^1 = 2\boldsymbol{W}_x^1(\boldsymbol{X}\boldsymbol{X}^T + \delta_W^1 \boldsymbol{I}) + \\
&\qquad 2(\boldsymbol{W}_y^1 \boldsymbol{Y} + \boldsymbol{b}_x^1 + \boldsymbol{b}_y^1 - \boldsymbol{M}^c)\boldsymbol{X}^{\mathrm{T}}, \\
&\nabla_{\boldsymbol{W}_y} L^1 = 2\boldsymbol{W}_y^1(\boldsymbol{Y}\boldsymbol{Y}^T + \delta_W^1 \boldsymbol{I}) + \\
&\qquad 2(\boldsymbol{W}_x^1 \boldsymbol{X} + \boldsymbol{b}_x^1 + \boldsymbol{b}_y^1 - \boldsymbol{M}^c)\boldsymbol{Y}^{\mathrm{T}}, \\
&\nabla_{\boldsymbol{b}_x} L^1 = 2(\boldsymbol{W}_x^1 \boldsymbol{X} + \boldsymbol{W}_y^1 \boldsymbol{Y} + \boldsymbol{b}_y^1 - \boldsymbol{M}^c) + \\
&\qquad 2(\delta_b^1 + 1)\boldsymbol{b}_x^1, \\
&\nabla_{\boldsymbol{b}_y} L^1 = 2(\boldsymbol{W}_x^1 \boldsymbol{X} + \boldsymbol{W}_y^1 \boldsymbol{Y} + \boldsymbol{b}_x^1 - \boldsymbol{M}^c) + \\
&\qquad 2(\delta_b^1 + 1)\boldsymbol{b}_y^1.
\end{aligned} \tag{10}$$

其中$\boldsymbol{M}^c$为所有微博的事件标注矩阵表示. 因为本文采用基于反向传播的梯度下降算法学习参数,所以要先更新基于多模态特征事件表达的参数(式(10)),再更新多模态特征融合的参数(式(8)). 通过式(10),能够得到事件的多模态表示模型,从而提高事件检测的准确率.

2.3 基于事件图平滑的事件跟踪

为跟踪不断演化的事件,在平滑事件图后,根据事件相似度,将事件连成故事线. 当事件随时间发生显著变化时,故事线内的事件无法被关联起来,因而需要减缓其演化速度. 此外,尽管微博内容常随时间

发生变化,但相同故事线的事件关键词一般不会剧变,所以本文以此计算事件间的相似度,将连续的相似事件拼接成完整的故事线.

首先,为避免相同故事线的事件随时间剧烈变化,要对事件图进行平滑,即每天的事件图都要保留之前事件图的部分信息,表示为

$$\widetilde{G}^t = (1-\lambda)G^{t-1} + \lambda G^t, \tag{11}$$

其中λ为平滑因子. 这样事件图不会剧烈变化,根据事件图检测到的事件也不会显著改变.

然后,为度量连续事件之间的关系,依据其关键词计算二者相似度. 通常,微博流中社会事件不会引起长期关注,事件之间的时差越大,其属于相同故事线的可能性越低. 为此,事件之间的相似度会随其时差的增大而减小,即

$$\mathrm{sim}_{c_1c_2} = \mathrm{e}^{-\alpha|t_1-t_2|}|c_v^1 \bigcup c_v^2|/|c_v^1|, \tag{12}$$

其中α为时差的衰减因子. 事件之间的时差越大,或者共同的关键词越少,其相似度越小. 当该值超过给定阈值$\theta_{\mathrm{sim}}^{cc}$时,认为它们属于相同故事线,即$c_1^s = c_2^s$. 当二者的时差超过$-(1/\alpha)\ln\theta_{\mathrm{sim}}^{cc}$时,无需计算其相似度. 这样,事件之间的时差上限导致事件相似度的计算量显著减少.

最后,基于事件的相似度,按时间顺序连接相似的事件以生成事件的完整故事线,从而达到对事件跟踪的目的.

3　实验分析

3.1　实验设计

本文实验数据为2012年3月25日～2012年5月25日从腾讯微博空间提取的1 000万条微博. 该数据集没有正确的事件标注,且因规模较大而不能被人工标注. 尽管微博和新闻网站的事件不同,但对于重大突发以及广泛讨论的社会事件,二者通常都会集中报道,且事件的重要信息(如关键词和关键图片)高度一致. 所以本文以新闻网站中对应时间的热点事件文章,作为微博中事件的参考标准. 表2为本文实验数据的详细信息. 由于没人会对所有事件感兴趣,与文献[3]类似,本文只研究每天前10个重要事件. 此外,为了尽可能地覆盖到微博中的事件,参考标准中的事件数量必需远大于微博流中的事件数量. 以下实验均在此数据集上进行研究.

本文以准确率和召回率双指标评价事件检测与跟踪的效果. 由于事件标注与事件跟踪是对事件的宏观掌控,实验只需判断该事件是否存在;而多模态深度融合是对事件内容的优化,实验需要判断事件内的微博是否正确. 对于事件标注或事件跟踪,以网站新闻事件作为参考标准,即

表2　实验数据的详细说明

数据集	类型	每天	总计	备注
微博流	文本	167 000	10 000 000	长度不超过140字符
	图片	57 000	340 000	每条微博最多只有一张图片
	社会事件	10	600	只研究每天前10个事件
参考标准	热点事件	37.3	2 238	平均每个事件包含2.1篇文章
	故事线	3.1	187	平均每个故事线有2.4个事件
	标注微博	71.7	4 300	来自于265个社会事件

$$\begin{aligned}\mathrm{Prec}(C) &= |C \bigcap C^*|/|C|,\\ \mathrm{Rec}(C) &= |C \bigcap C^*|/|C^*|,\end{aligned} \tag{13}$$

其中C^*为参考标准的所有事件. 式(13)衡量是否能够准确地(准确率)从微博流中挖掘出足够的事件(召回率). 对于事件检测,以人工标注的微博作为参考,有

$$\begin{aligned}\mathrm{Prec}(M) &= |\{m|m^c = m^{c*}, m \in M \bigcap M^*\}|/|M|,\\ \mathrm{Rec}(M) &= |\{m|m^c = m^{c*}, m \in M \bigcap M^*\}|/|M^*|.\end{aligned} \tag{14}$$

其中: M^*为参考标准中被事件标注的所有微博,m^{c*}为微博的正确事件标注. 式(14)分析本文方法自动标注的微博是否与人工标注的参考标准一致. 以下实验将按这两个指标分析本文参数对事件检测与跟踪的影响,并在相同环境中与当前最先进的算法进行对比. 所有实验均在一台i5-4590 CPU, 8 GB内存, 64位Windows7操作系统的电脑上运行.

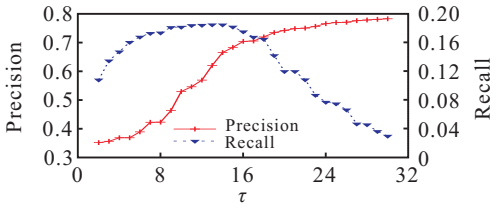
3.2　参数调节实验

为研究参数对本文方法的影响,依次对文本的事件标注、多模态深度融合的事件检测和基于事件图平滑事件跟踪的参数进行实验. 为统一起见,SCAN核心点邻居相似度ε、数量μ为0.5和5.

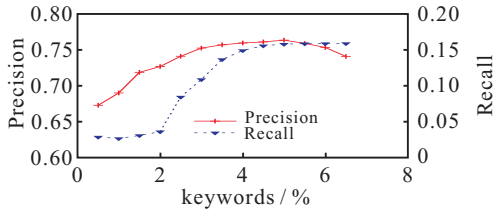
3.2.1　事件标注参数调节

首先,分析不同关键词抽取参数对事件检测的影响(如图1所示). 图1(a)显示了当关键词的IDF在不同时间窗口长度τ下进行计算时,事件检测的准确率和召回率. 当τ太小时,IDF仅在短期内被平均化,导致小干扰就会产生大量的错误关键词,进而降低准确率和召回率. 当τ从2增至14时,错误关键词逐渐被移除,检测的效果也快速提升. 当τ继续增加时,准确率缓慢增加,但召回率急剧下降. 权衡准确率和召回率后,τ最后被设为14. 图1(b)显示了对于不同关键词数量,事件检测的准确率和召回率. 显然,关键词的缺乏导致检测到事件较少(准确率和召回率都小). 随着关键词的增多,更多的事件将被发现,但太多关键

词会引发混乱,并降低准确率. 当关键词数量从0.5%增至4.5%时,事件检测的准确率和召回率均快速增长. 当数量继续增加到5%时,增长变缓. 由于少量的关键词已经能够很好地实现事件检测,权衡时间开销与检测效果,将使用4.5%的词语作为关键词用于事件分析.



(a) 时间窗口长度



(b) 关键词数量

图 1 关键词参数对事件检测性能的影响

下面分析基于图聚类的事件标注效果. 图2显示了当事件图中节点的最近邻数量变化时,事件检测的性能. 与图1(a)类似,如果关键词间的关系不足,很难在事件图中检测事件,直到最近邻的数量超过6时,事件检测的准确率才会达到最优. 图3显示了当微博-事件相似度阈值不同时,微博文本被正确标注的效果. 显然,阈值越大,事件标注的准确率也越高. 但阈值太大,大量拥有小相似度的微博将会被排除在外,严重影响召回率. 经权衡,微博-事件的相似度阈值 $\theta_{\text{sim}}^{\text{mc}}$ 最终被设为1.2.

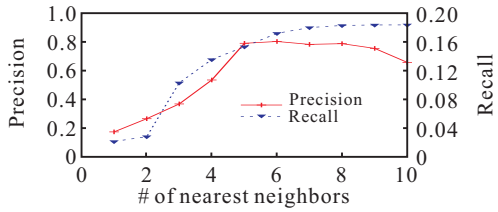


图 2 事件图中节点近邻数对事件检测性能的影响

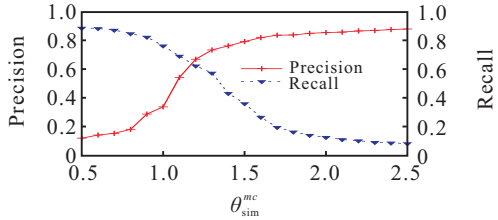


图 3 相似度阈值对事件检测性能的影响

3.2.2 事件检测参数调节

在多模态深度融合事件检测的实验中,调节参数 δ_W 和 δ_b 均为默认值 10^{-6},学习率 η 为 10^{-1},参数更新次数 l 为 10^2. 输入图片被压缩为 512×512,输入文本为单词200维word2vec向量组成的矩阵. 由于微博流中有大量与事件无关的微博,在事件表达中需要为此额外增加一个默认类. 图4和图5分别为基于深度学习检测和表达后,事件内容的准确率和召回率,其中Original、Textual、Multimodal分别表示基于事件标注、基于文本深度学习、基于多模态特征深度融合的事件检测,#2*50表示模型共2层,每层50个节点. 如图所示,即使仅对属于事件的微博文本信息进行深度学习,基于深度学习的方法都要优于原始基于共享关键词的方法. 此外,随着融合的层数以及每层节点数量的增加,事件内容的准确率和召回率在多数情形下都稳步提高,且多模态深度融合的方法要好于单模态深度学习的方法. 但当融合层的节点过多时(如Multimodal#4*200),则会过拟合,严重降低准确率和召回率. 最后,融合层被设为4层,每层100个节点.

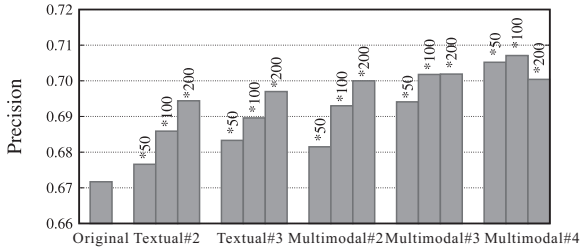


图 4 基于深度学习后事件检测的准确率

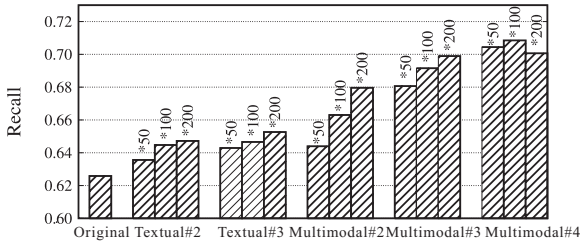


图 5 基于深度学习后事件检测的召回率

3.2.3 事件跟踪参数调节

为研究基于事件图平滑的事件跟踪效果,先分析事件图平滑因子 λ 对事件跟踪的影响. 当故事线内的事件内容随时间发生剧变时,如果在经过事件图平滑后该事件被认定为同一故事线,则说明平滑因子促进了事件跟踪;反之则没有. 如图6所示,每天的事件图只能吸收少量相邻时间事件图的信息,这样事件跟踪的准确率和召回率才会提升,吸收太多将直接失去当天事件图的内容. 为使事件跟踪效果最佳,本文将事件图平滑因子 λ 设为0.3.

为研究事件相似度对事件跟踪的影响,接着对事件相似度时差衰减因子和相似度阈值进行分析. 图7为时差衰减因子 α 对事件跟踪的影响. 由于社会事件在微博流中通常不会持续很长时间,事件之间的时差

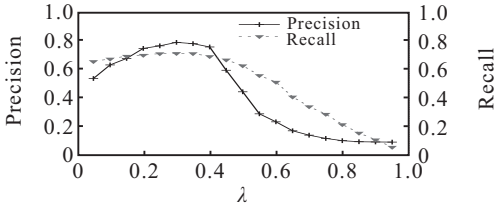


图 6　事件图平滑因子对事件跟踪性能的影响

越大,其属于相同故事线的可能性越小. 当时差衰减因子α逐渐增大时,除非两个事件的时差很小,二者属于相同故事线的可能性逐步减小. 尽管这样会产生较高的准确率,但同时带来极低的召回率. 本文希望生成一条完整的故事线,需要将相似的事件尽可能地连接起来,所以选择了一个拥有较高召回率和准确率可接受的时差衰减因子α,并将其设置为0.05. 图8为事件相似度阈值对事件跟踪的影响. 与图3类似,阈值太大导致高准确率和低召回率,太小导致低准确率和高召回率. 经权衡后,事件相似度阈值θ_{sim}^{cc}被设为0.55. 因此,当事件的时差超过$-(1/0.05)\ln 0.55 = 11.95$天,则不用计算其相似度.

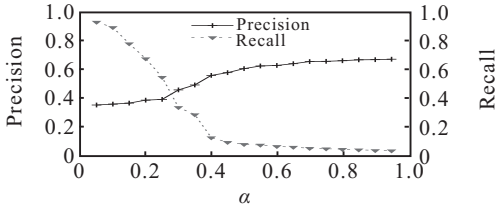


图 7　时差衰减因子对事件跟踪性能的影响

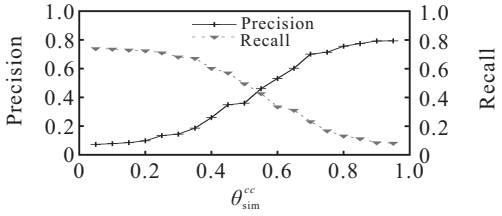


图 8　事件相似度阈值对事件跟踪性能的影响

3.3　对比实验

为表明本文方法的有效性,将所提出算法与现有的先进方法进行比较. 由于现有方法不能像本文方法一样,在充满噪音的微博流上利用多模态特征直接进行事件检测与跟踪,本文将实验拆分为事件检测与跟踪、多模态事件检测两个部分.

3.3.1　基于文本的事件检测与跟踪

对事件的检测与跟踪仅需判断某事件是否存在以及下一时刻事件的状态位置,所以将依据事件自身的准确性与各方法进行对比. 由于部分方法只能检测而无法跟踪事件,为保证公平,如果这些方法能在相邻时间段内发现正在演化的事件,则认为它们实现了事件跟踪. 图9为本文方法与STM-TwitterLDA[4]、TopicSketch[5]和CNPHGS[7]的对比实验. 可见,本文方法的准确率和召回率都远优于对比方法. 这是由于本文针对微博的特性专门设计了一种新型关键词重要度以及词语相似度的计算方法,使本文提取的事件关键词比其他方法更全面准确,根据事件图聚类得到的事件也相应地更好. 所以,本文方法在事件检测与跟踪中能够比对比方法更有效地发现微博流中隐藏的社会事件.

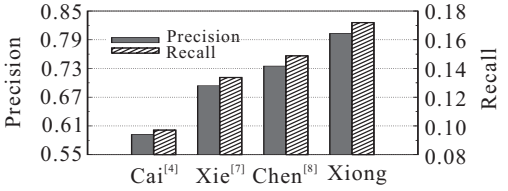


图 9　基于文本处理的事件检测与跟踪性能对比

3.3.2　基于多模态特征融合的事件检测与表达

由于目前多数方法都不能直接在含有大量噪音的微博流上运行,为保证公平,所有方法均运行在本文基于文本处理的事件标注结果上. 因为多模态事件检测可视为对基于文本事件标注的优化,所以本文依据事件内容的准确性对各方法进行评估. 图10为本文方法与Bian等的CMLDA[2]、Cai 等的STM-TwitterLDA[4]、Qian 等的mmETM[5]的对比实验. 如图10所示,由于STM-TwitterLDA不仅引入了微博的文本和图片信息,还额外利用了微博的时间与位置等特征,所以比本文方法更准确. 然而,由于本文方法是在图片与语义特征的基础上进行深度融合,可以学习出具有普通性的高层语义特征,所以召回率较高. 尽管本文方法的准确率略低于STM-TwitterLDA,但与其他方法相比均具备显著优势. 因此,在多模态事件检测任务中本文方法要优于对比方法.

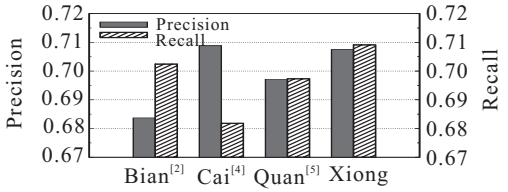


图 10　基于多模态特征融合的事件检测性能对比

4　结　论

本文将数量迅速增加、内容不断变化的微博数据流作为输入,动态地检测并跟踪社会事件. 通过基于文本处理的事件标注、基于多模态特征融合的事件检测与表达、基于时间平滑的事件跟踪机制,实现在微博流中社会事件的发现及随时间演化的完整故事线的多模态形式表达. 在真实数据集上的实验验证了本文方法的有效性,且在多数情况下比当前最新方法更适合在微博流中检测并跟踪社会事件.

在未来工作中,会将本文工作应用至其他研究领

域,例如事件摘要生成、内容推荐、事件可视化等. 此外,由于本文的事件检测与跟踪算法包含大量参数,未来将研究如何自动设置参数.

参考文献(References)

[1] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: Real-time event detection by social sensors[C]. Proc of 19th World Wide Web. New York: ACM, 2010: 851-860.

[2] Bian J, Yang Y, Chua T S. Multimedia summarization for trending topics in microblogs[C]. Proc of 22nd Conf on Information and Knowledge Management. New York: ACM, 2013: 1807-1812.

[3] Zhou X, Chen L. Event detection over twitter social media streams[J]. J on Very Large Data Bases, 2014, 23(3): 381-400.

[4] Cai H, Yang Y, Li X, et al. What are popular: Exploring twitter features for event detection, tracking and visualization[C]. Proc of 23rd Multimedia. New York: ACM, 2015: 89-98.

[5] Qian S, Zhang T, Xu C, et al. Multi-modal event topic model for social event analysis[J]. IEEE Trans on Multimedia, 2016, 18(2): 233-246.

[6] Xie W, Zhu F, Jiang J, et al. TopicSketch: Real-time bursty topic detection from Twitter[J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(8): 2216-2229.

[7] Chen F, Neill D B. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs[C]. Proc of 20th Knowledge Discovery and Data Mining. New York: ACM, 2014: 1166-1175.

[8] Schinas M, Papadopoulos S, Petkos G, et al. Multimodal graph-based event detection and summarization in social media streams[C]. Proc of 23rd Multimedia. New York: ACM, 2015: 189-192.

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

[10] Fang Q, Sang J, Xu C, et al. Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning[J]. IEEE Trans on Multimedia, 2017, 16(3): 796-812.

[11] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: A survey of approaches and applications[J]. IEEE Trans on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.

[12] Cameron M A, Power R, Robinson B, et al. Emergency situation awareness from twitter for crisis management[C]. Proc of the 21st World Wide Web. New York: ACM, 2012: 695-698.

[13] Lin C, Lin C, Li J, et al. Generating event storylines from microblogs[C]. Proc of the 21st Conf on Information and Knowledge Management. New York: ACM, 2012: 175-184.

[14] Yin H, Cui B, Lu H, et al. A unified model for stable and temporal topic detection from social media data[C]. Proc of the 29th International Conf on Data Engineering. Los Alamitos: IEEE Computer Society, 2013: 661-672.

[15] Xiong Y, Zhang Y, Wang D, et al. Picture or it didn't happen: Catch the truth for events[J]. Multimedia Tools and Applications, 2017, 76(14): 15681-15706.

[16] Yang Y, Wu F, Xu D, et al. Cross-media retrieval using query dependent search methods[J]. Pattern Recognition, 2010, 43(8): 2927-2936.

[17] Sun Y, Aggarwal C C, Han J. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes[J]. Proc of VLDB Endowment, 2012, 5(5): 394-405.

[18] Xiong Y, Wang D, Zhang Y, et al. Multimodal data fusion in text-image heterogeneous graph for social media recommendation[C]. Proc of the 15th Web-Age Information Management. Cham: Springer, 2014: 96-99.

[19] Wu F, Yu Z, Yang Y, et al. Sparse multi-modal hashing[J]. IEEE Trans on Multimedia, 2014, 16(2): 427-439.

[20] Wu B, Yang Q, Zheng W S, et al. Quantized correlation hashing for fast cross-modal search[C]. Proc of the 24th Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI Press, 2015: 3946-3952.

[21] Xie L, Shen J, Zhu L. Online cross-modal hashing for web image retrieval[C]. Proc of the 30th Association for the Advancement of Artificial Intelligence. Menlo Park: AAAI Press, 2016: 294-300.

[22] Wang W, Ooi B C, Yang X, et al. Effective multi-modal retrieval based on stacked auto-encoders[J]. Proc of VLDB Endowment, 2014, 7(8): 649-660.

[23] Cao Y, Long M, Wang J, et al. Deep visual-semantic hashing for cross-modal retrieval[C]. Proc of the 22nd Knowledge Discovery and Data mining. New York: ACM, 2016: 1445-1454.

[24] Yang E, Deng C, Liu W, et al. Pairwise relationship guided deep hashing for cross-modal retrieval[C]. Proc of the 31st Association for the Advancement of Artificial Intelligence. Menlo Park: AAAI Press, 2017: 1618-1625.

[25] Xu X, Yuruk N, Feng Z, et al. SCAN: A structural clustering algorithm for networks[C]. Proc of the 13th Knowledge Discovery and Data mining. New York: ACM, 2007: 824-833.

作者简介

熊宇 (1985−), 男, 博士生, 从事多模态数据融合及其应用的研究, E-mail: xiongyu@stumail.neu.edu.cn;

张一飞 (1977−), 女, 讲师, 博士, 从事图像处理及机器学习的研究, E-mail: zhangyifei@cse.neu.edu.cn;

冯时 (1981−), 男, 副教授, 博士, 从事情感分析及文本挖掘等研究, E-mail: fengshi@cse.neu.edu.cn;

王大玲 (1962−), 女, 教授, 博士生导师, 从事社会媒体分析、数据挖掘等研究, E-mail: wangdaling@cse.neu.edu.cn.

(责任编辑: 郑晓蕾)