软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(10):2654−2673 [doi: 10.13328/j.cnki.jos.005146]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融合语境分析的时序推特摘要方法*

于广川[1,2], 贺瑞芳[1,2], 刘 洋[3], 党建武[1,2]

[1](天津大学 计算机科学与技术学院,天津 300350)

[2](天津市认知计算与应用重点实验室,天津 300350)

[3](北京大学 信息科学技术学院,北京 100871)

通讯作者: 贺瑞芳, E-mail: rfhe@tju.edu.cn

摘 要: 时序推特摘要是文本摘要任务中的一个重要分支,旨在从热点事件相关的海量推特流中总结出随时间演化的简要推特集,以帮助用户快速获取信息.推特作为当今最流行的社交媒体平台,其信息量爆发式的增长以及文本碎片的非结构性,使得单纯依赖文本内容的传统摘要方法不再适用.与此同时,社交媒体的新特性也为推特摘要带来了新的机遇.将推特流视作信号,剖析了其中的复杂噪声,提出融合推特流随时序变化的宏微观信号以及用户社交上下文语境信息的时序推特摘要新方法.首先,通过小波分析对推特流全局时序信息建模,实现某一关键词相关的热点子事件时间点检测;接着,融入推特流局部时序信息和用户社交信息建立推特的随机步图模型摘要框架,为每个热点子事件生成推特摘要.在算法评估过程中,对真实推特数据集进行了专家时间点和专家摘要的人工标注,实验结果表明了小波分析和融合了时序-社交上下文语境的图模型在时序推特摘要中的有效性.

关键词: 时序推特摘要;时序特性;用户社交权威性;小波去噪;上下文图模型

中图法分类号: TP391

中文引用格式: 于广川,贺瑞芳,刘洋,党建武.融合语境分析的时序推特摘要方法.软件学报,2017,28(10):2654−2673. http://www.jos.org.cn/1000-9825/5146.htm

英文引用格式: Yu GC, He RF, Liu Y, Dang JW. Context based model for temporal Twitter summarization. Ruan Jian Xue Bao/Journal of Software, 2017,28(10):2654−2673 (in Chinese). http://www.jos.org.cn/1000-9825/5146.htm

Context Based Model for Temporal Twitter Summarization

YU Guang-Chuan[1,2], HE Rui-Fang[1,2], LIU Yang[3], DANG Jian-Wu[1,2]

[1](School of Computer Science and Technology, Tianjin University, Tianjin 300350, China)

[2](Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300350, China)

[3](School of Information Science and Technology, Peking University, Beijing 100871, China)

Abstract: Temporal Twitter summarization is an important sub-task of text summarization, which aims to extract a concise tweet set with time, goes from a huge Twitter stream. It helps users quickly understand a specific event. As one of the most popular social media platforms, the explosive growth of Twitter information makes it difficult for users to find reliable and useful information. As tweets are short and highly unstructured, it makes traditional document summarization methods difficult to handle Twitter data. Meanwhile, Twitter also provides rich temporal-social context more than texts, bringing new opportunities. This paper considers Twitter stream as a kind of signal, and proposes a novel temporal Twitter summarization method by modeling macro-micro temporal context and social context through analyzing the complex noises hidden in signal. First, time points of hot sub-events are detected by modeling temporal context globally with wavelet analysis. Second, a novel random walk model is built on graph based unsupervised Twitter summarization framework, integrating both local temporal context and social user authority to generate summary for each sub-event time point. To evaluate

* 基金项目: 国家重点基础研究发展计划(973)(2013CB329301); 国家自然科学基金(61472277)

Foundation item: National Key Basic Research and Development Program of China (973) (2013CB329301); National Natural Science Foundation of China (61472277)

收稿时间: 2016-04-23; 修改时间: 2016-08-29; 采用时间: 2016-10-01

the proposed framework, a real-world Twitter dataset, including expert time point and summary, is manually labeled. Experimental results show that wavelet analysis during hot sub-event time point detection and temporal-social context in Twitter summarization are both effective.
Key words: temporal Twitter summarization; temporal context; social user authority; wavelet denoising; context based graph model

时序推特摘要旨在从热点事件相关的海量推特流中提取出随时间演化的简要推特集,以使用户通过这段文本快速地对事件有尽量全面的了解.推特是全球最著名的社交平台之一,具有用户数量庞大、信息碎片化及覆盖范围广的特点.当一个热点事件发生时,大量用户第一时间在推特发布和分享信息,使得推特平台的信息量不断扩散和增长,用户越来越难以从中发现有参考价值的资讯.因此,时序推特摘要逐渐成为自然语言处理任务中迫切需要解决的任务之一,可为社会热点事件的舆情监控以及商业竞争情报分析提供重要的技术支撑.

社交媒体平台作为大众用户发布和分享信息的媒介,其文本信息组织形式有别于传统媒体文本.社交媒体文本具有长度较短和非结构性的特点,且用户信息发布的规范性参差不齐.因此,在文本信息结构方面,社交媒体文本与传统媒体文本相比存在劣势,单纯依赖文本内容的传统文本摘要方法对时序推特摘要并不具有良好的直接迁移性.

与此同时,社交媒体平台的优势和特点也为时序推特摘要带来新的契机,如图 1 所示.其一,海量用户使得推特流时序信息对于一个事件的发生更具有时间敏感性.即当一个热点事件发生时,大量用户第一时间在推特发布和分享事件信息,信息的时效性要比传统新闻文本更加显著,且事件的起始状态与起伏过程会在推特流信号的宏观趋势中有所体现.考虑“一条文本消息出现在恰当的时刻要比出现在一个无关紧要时刻更值得参考和信赖”,可从推特流信号中捕捉与每个推特文本对应的局部瞬时时序特性来辅助推特摘要句子的选择.其二,社交媒体平台蕴含大量的用户属性信息(如关注数、粉丝数、微博数)和交互信息(如转发、评论),其对用户发布的信息质量判定有很大的参考价值.拥有较高权威性用户发布的信息会有更大的概率被认为是较高质量的信息.因此,用户属性和社交关系的强弱将对推特内容质量的判定有重要影响.

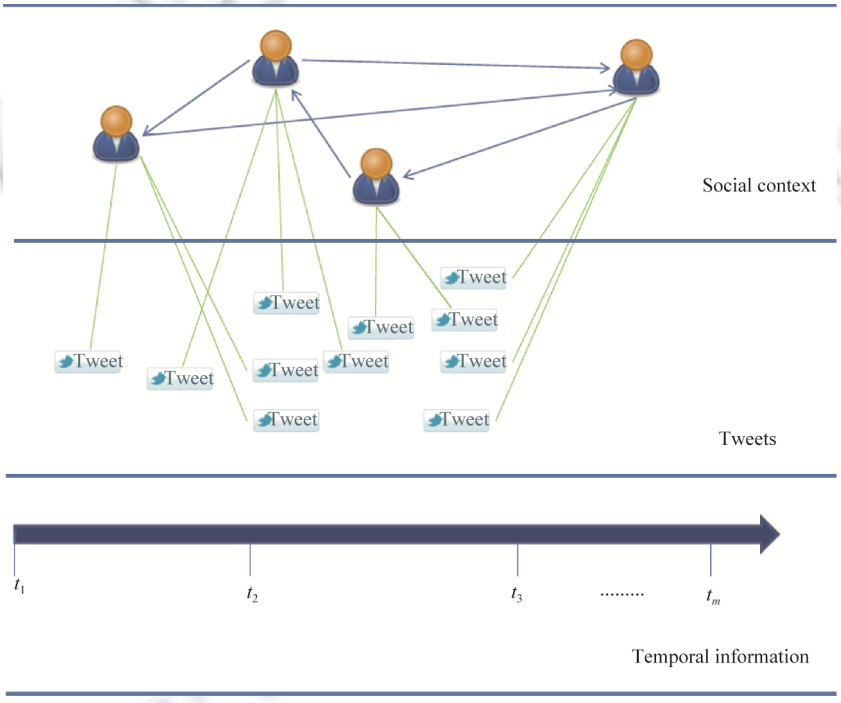


Fig.1 Twitter data with temporal-social context

图 1 包含时序和社交语境的推特数据

由此,推特流时序信息和用户社交属性关系信息在推特文本内容的基础上,可为时序推特摘要的选取提供很重要的参考价值.在目前的社交媒体文本摘要任务中,已有一些学者对推特热点事件进行了子话题检测[1−3]以及对用户属性和交互行为[1,4−6]的研究来辅助摘要的选取.在子话题检测研究中,基于推特流宏观信号的方法[1,2]对单一关键词的热点事件拥有较好的表现,但该方法忽略了推特流中复杂噪声对重要时间点选择的影响,本文利用小波分析的方法对推特流宏观信号进行降噪处理并进行重要时间点的选择;在用户权威性的研究中,已有工作分别对用户静态属性和动态交互关系进行了探索,本文综合考虑了两方面因素来衡量用户的权威性;在推特摘要研究中,一些学者利用交互关系来引导推特摘要的选取[5,6],该方法巧妙利用了社交媒体平台的交互关系对摘要选取的重要作用,但对于推特内容本身与推特上下文语境的结合有待进一步的探究.因此,本工作将推特流视作信号,剖析了推特流中的复杂噪声,提出融合时序-社交上下文语境的时序推特摘要新方法.通过采用小波分析的方法进行热点子事件时间点的检测;以此为基础,建立上下文语境的随机步图模型,实现时序推特摘要.

本文着重研究以推特流时序信息和用户社交信息辅助进行时序推特摘要的选择,包括:(1) 如何从海量的热点事件推特流信息中检测热点子事件时间点;(2) 如何分别为推特流时序信息和用户社交信息建模;(3) 如何将推特流局部时序信息和用户社交信息与推特文本内容相结合,以促进优质推特摘要的选择.

针对以上问题,本文提出通过挖掘社交媒体推特流时序信息以及社交信息,建立基于小波分析和上下文图模型的时序推特摘要方法.

本文第 1 节剖析时序推特摘要的相关工作.第 2 节对研究问题进行定义.第 3 节给出方法的描述.第 4 节介绍实验数据的准备、模型评估方法以及实验结果的讨论分析.第 5 节对本工作进行总结及展望.

1 相关工作

本文的研究目标为以关键词为线索的话题时序摘要,即首先进行热点子事件时间点检测,然后结合推特流时序信息和用户社交信息分别对每个热点子事件进行推特摘要.按照本文的研究内容,将分别从热点子事件检测、社交媒体用户权威度建模以及推特摘要模型这 3 个方面对相关工作进行介绍,并在第 1.4 节将本文工作与 TREC 国际评测中的时序摘要任务(TREC-TS)进行了比较.

1.1 推特流子事件检测

子事件检测是时序推特摘要的首要子任务.在目前的研究工作中,子事件检测可分为:(1) 基于推特流分析的方法;(2) 基于事件聚类的方法.

(1) 基于推特流分析的方法.2012 年,Nichols 等人[2]及 Zubiaga 等人[3]针对体育运动事件相关推特数据,结合该类型事件热点子事件爆发较集中的特点,采用尖峰检测(spike detection)的方法,在宏观推特流信号中进行子事件爆发时间点发现;2013 年,Shen 等人[7]考虑推特流宏观信息同时加入事件参与者的同步热度信息来辅助子事件时间点的选择,在一定程度上解决了推特流信息存在噪声的问题;2014 年,Gao 等人[1]对一个事件的推特数据流采用线下峰值检测(offline peak area detection)的方式进行子事件检测,设定一个固定长度的时间窗口,使其在时间线上滑动并统计窗口内的推特数量,若大于设定阈值,则视为出现一个波峰,将若干个波峰所在区域视为子事件发生的起始和结束位置.

(2) 基于事件聚类的方法.2013 年,Olariu 等人[8]针对一个未知话题数的复杂推特流,先确定该推特流中的若干个被热议的话题,然后按此信息对推特数据进行聚类,从而实现子事件检测和数据集的划分;2014 年,Gao 等人[1]以推特流信息的峰值检测结果作为子事件个数的输入,采用基于潜在狄利克雷分配的动态话题模型(dynamic topic model)对一个事件相关的数据集进行子事件聚类,将各个聚类结果中推特时间戳的均值作为该子事件对应的时间点;除此之外,2014 年,Kim 等人[9]考虑推特文本不规范的特点,以单词角度切入,以单词在推特中的共现关系构建图模型,通过最大 K 子团发现的方法进行子事件检测.

事件聚类的方法对子事件间差异性比较敏感,差异性较大的子事件聚类有很好的效果,但对于话题相似度较强的子事件检测效果不够理想.由于本工作是以关键词为事件线索的时序推特摘要,因此采取推特流分析的

方法,可以更好地捕捉推特流随时序变化的敏感性.相比已有工作,大多数工作从宏观角度进行峰值检测而忽略了多因素文本噪声的干扰,本文首先使用小波分析的原理对全局推特流信号进行降噪处理,由此提出一种基于转折率的峰值点选择算法,以实现热点子事件时间点的检测,从而更好地排除复杂噪声对真实信号的干扰.

1.2 社交用户权威度建模

用户信息挖掘在推特摘要中也起着重要作用,推特内容质量的高低与用户权威性有着紧密的联系,一些学者在社交媒体文本摘要任务中也对用户权威性进行了探究.

2012 年,Duan 等人[4]考虑文本内容、用户社会影响力和文本质量 3 个因素的互增强关系建立互增强式图模型.用线性支持向量机模型通过用户的关注数、粉丝数以及总推特数等静态信息计算每个用户话题无关的静态得分,随后该用户的整体权威性得分由静态得分和其粉丝质量所决定.在随机游走图模型中,每一步的用户权威性得分由上一步的推特得分、单词得分和用户得分这三者线性加权组成.

2013 年,Chang 等人[5]提出在一个有监督的学习框架下,用推特之间的评论关系构建文本树,并借助经济学领域的因果关系模型(granger causality influence model)思想,首先利用用户间的关注关系确定若干对有因果关系的用户,认为与文本树根节点用户有因果关系的用户具有更高的权威度.此外,还通过文本树中的用户交互关系构建用户图模型,并利用 PageRank 算法进行用户权威度排序.最终通过两方面的用户权威度来共同辅助摘要的选择.

2015,年 Li 等人[6]沿用 Chang 等人[5]的思想,利用转发关系对单一热点事件的微博扩散过程进行建模.在用户权威度度量部分,使用条件随机场模型在已构建的微博转发树上进行领导者发现(leaders detection),从而把发布信息丰富且受到大量用户认可的用户识别出来,实现对用户重要性的权衡.

上述 3 项工作分别从不同角度对用户权威性进行了研究以促进推特摘要内容质量的提升,有效地利用了用户属性和交互结构方面的特征.Chang 等人[5]及 Li 等人[6]提出用户交互模型主导的摘要算法,忽略了用户静态属性信息的作用,同时也一定程度地降低了文本内容本身的质量对摘要性能的影响.Duan 等人[4]虽然利用了用户静态信息,但用户权威性的计算依赖于推特及词的重要性同步计算,增加了用户权威性计算的时间和空间复杂性.且大多数学者以用户模型为主导,忽视了推特流时序信息在摘要抽取中的作用.除此之外,Chang 等人提出的是有监督的学习模型,对语料的人工标注需要付出较高代价.本文利用用户的动静态属性信息进行用户权威性建模,以此作为推特的社交上下文语境,通过融合推特流的社交、时序上下文语境形成一种新的无监督图模型,进行推特摘要的抽取.该方法既充分挖掘了时序-社交上下文语境,又减少了人工标注训练语料的代价.

1.3 推特摘要

文本摘要按产生形式可分为抽取式和理解式,本文提出的推特摘要框架属于抽取式方法,这里仅对抽取式推特摘要的相关工作进行总结.

结合社交媒体文本的特性,很多学者将传统文档摘要技术迁移到推特摘要中.即将一条推特看作是一个句子,然后应用多文档文摘技术进行抽取式摘要[10].典型的多文档摘要方法包括:(1) SumBasic[11]方法;(2) 基于中心性原则(centroid algorithms)[12]的方法;(3) 基于图模型[13]的方法.其中,(1)认为频率是单词主题表征性之一,通过计算在整个推特集中每个单词的出现频率来决定哪些推特最有代表性;(2)则先计算出一系列推特句子的伪中心,然后通过计算每个推特与伪中心的相似度来衡量各个推特的中心性;(3)把推特看作是图中的节点,由计算两两之间的相似度来构建节点之间的转移概率,然后通过随机游走过程完成推特的排序和摘要句的选择.

近年来,一些学者以社交媒体交互行为结构为切入点,结合用户社交权威性信息进行摘要句子选择,论述见第 1.2 节.Chang 等人[5]利用推特评论关系构建文本树,通过基于用户交互关系和关注关系的随机游走图模型计算用户权威度,从而诱导摘要句子的选择.Li 等人[6]针对中文微博,首先从转发行为的角度构建热点事件转发结构树,然后用基于内容的条件随机场模型在转发结构树中进行领导者发现,从而判别出转发结构树中的每个节点是否是有领导性意义的用户节点,并以用户节点性质结合推特内容质量来引导摘要句子的抽取.

上述推特摘要工作一方面利用了传统摘要方法的可迁移性,另一方面融合了社交媒体平台交互结构的优势,但大部分研究局限于推特文本内容和用户权威性挖掘,很少有学者同时考虑推特流时序信息与用户社交信

息来进行推特摘要的抽取,特别是推特流多因素噪声的干扰.

1.4 与国际评测 TREC-TS 任务的比较

TREC-TS(Text REtrieval Conference-Temporal Summarization)[14]是由美国国家标准与技术研究院(NIST)举办的时序摘要国际评测比赛.尽管都是时序摘要,但本文在任务定义方面与之不同.以 2014 TREC-TS Track 为例,将本文时序推特摘要任务与 TREC-TS 评测任务对比如下.

相同点:本文与 TREC-TS 评测皆为时序摘要任务,任务目标皆为从某个原始语料中,对某一对象随时间发展和演化过程中的关键信息进行自动抽取,使用户快速了解该对象的主要发展历程.

不同点:

(1) 二者处理的研究对象不同.本文为特定关键词相关的时序摘要,如“Obama”“iPad”和“Microsoft”等;TREC-TS 为面向危机事件的时序摘要,事件类型包括自然灾害、事故、爆炸、枪击及游行示威活动等;

(2) 二者处理的数据媒介类型不同.本文主要研究社交媒体文本(Twitter)时序摘要;TREC-TS 使用的语料主要为传统网页新闻文档以及部分社交媒体文本.

2 问题定义

时序推特摘要处理以关键词为事件线索的推特流,并从中总结出随时间演化的简要推特集合.假设给定某个关键词 k 相关的原始推特流数据集 Z 为输入,首先,通过数据集划分得到热点子事件时间点集合 $TP=\{imp_1, imp_2,\ldots,imp_c\}$,其中,$c$ 为热点子事件时间点总数;之后,对每个时间点提取相应日期的子事件集合 $E=\{S_1,S_2,\ldots,S_c\}$.对于每个事件集 $S_i=\{tw_1,tw_2,\ldots,tw_{num}\}$,$num$ 为 S 事件集中推特的总数,进行推特抽取得到摘要集 $S'_i=\{tw_1,tw_2,\ldots, tw_{num'}\}$,且 S'_i 中推特单词总数不超过文摘长度规定的单词数 n,num' 为摘要集中推特的总数,则 $S'=\{S'_1,S'_2,\ldots,S'_c\}$ 为时序推特摘要的最终输出.

本文提出的时序推特摘要方法主要步骤包括:

(1) 推特流热点子事件时间点检测;

(2) 局部时序特性与用户权威度建模;

(3) 融合社交媒体上下文信息的推特摘要.

3 方 法

3.1 热点子事件时间点检测

3.1.1 全局时序热度信号构建

首先构建原始推特流数据集 Z 的时序热度信号 $f(t)$.为了更好地反映推特流的时序变化特征,提出用推特发布速度表示某一时刻关于 k 的事件正在讨论的热度,定义 t_i 时刻的推特更新速度 $v(t_i)$为

$$v(t_i)=\frac{N(t_i)}{\Delta t}\times\frac{1}{N_{all}} \tag{1}$$

其中,$N(t_i)$表示$[t_i,t_i+\Delta t]$时间段内包含关键词 k 的推特数,Δt 为时间窗口长度,N_{all} 为$[t_i,t_i+\Delta t]$时间段内所有推特数,在此起到归一化的作用.由此可以得到一维热度信号集合 $f(t)=\{v(t_1),v(t_2),\ldots,v(t_q)\}$,其中,$q$ 是时间点总个数.

图 2 是以“Obama”关键词为例的原始推特集构建的时序热度信号结果.其中,横轴代表时间点序列,时间跨度为 2011 年 1 月 1 日~7 月 31 日,时间窗口长度Δt 的单位为小时;纵轴为推特流所对应的热度信号值.可以看到,热度信号随时间分布有很明显的差异,最高值出现在 2011 年 5 月 1 日,与“奥巴马宣布美国军队击毙本拉登”这一重大历史事件的真实日期相符.其他几个较高的峰值点对应的日期也相应地与“Obama”有关的较重要事件对应的日期相吻合.

经过多个数据集实验观察发现,热度信号的峰值点除了较高的几个重要时间点比较明显之外,如果我们想获取更多的有关某一关键词的系列热点事件,很难从热度信号图中发现那些并不显著的局部峰值点.

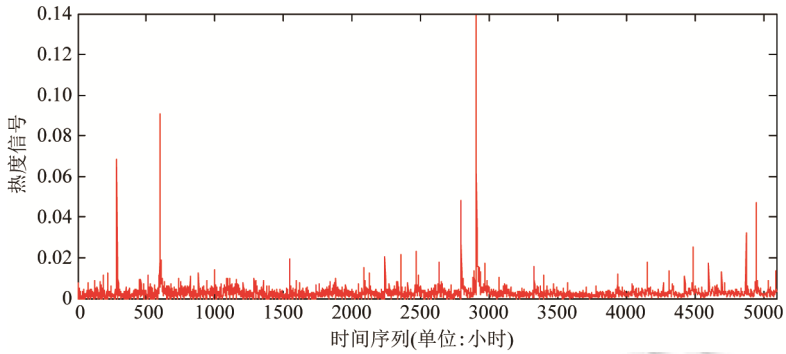


Fig.2 Global heat signal of “Obama” relevant Twitter stream from 2011/01/01 to 2011/07/31

图 2 2011 年 1 月 1 日~7 月 31 日“Obama”相关推特流的全局热度信号

3.1.2 基于离散小波变换的全局时序信号去噪方法

实际上,一个热点事件发生的同时,推特流中存在的复杂噪声信号对真实热点事件信号产生了干扰.当推特流信噪比较大时,热点事件信号易于检测;当信噪比较小时,噪声信号对真实信号影响较大,真实热点事件的信号位置很难捕捉.因此,仅从原始推特流信号中进行峰值点检测,对非极其显著的热点子事件时间点,并不能准确定位,故提出采用离散小波变换对全局时序信号进行去噪.如图 3 所示,本文分析了全局推特流中噪声信号的成因.

(1) 话题多样性:用户在谈论某一热点事件的同时也在谈论与事件主题相关性不高的其他琐碎话题内容.

(2) 时空误差:由于推特是全球性的社交媒体,当某一地方发生热点事件,该事件发生的消息扩散到世界各地用户时会有地理位置差异所导致的时空误差.

(3) 话题延续性:一个热点事件发生之后不会戛然而止,事件发生的最显著时间点之后可能还被部分网友所讨论.

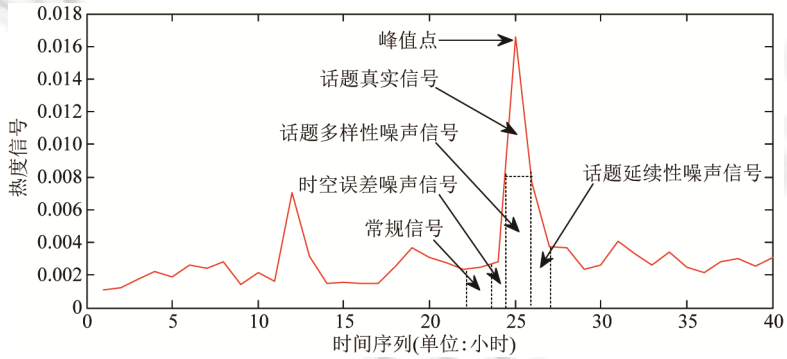


Fig.3 Real topic signal and complex noise signal in Twitter stream

图 3 推特流真实信号与多因素噪声信号

为了减少复杂噪声对重要热点子事件时间点检测的不利影响,我们采用小波分析进行热度信号去噪处理.由此提取出高频信号,而尽量削弱伪高频噪声信号,使得重构之后的信号更能反映真实高低频信号间的区别.

小波分析是通过生成一个逐渐衰减的震荡函数(母波函数$\phi(t)$)来模拟原始信号 $f(t)$,然后通过小波变换对母波函数进行缩放和变换,生成一系列线性独立的基函数,称为小波系,其定义为

$$\phi_{a,b}(t)=\frac{1}{\sqrt{a}}\phi\left(\frac{t-b}{a}\right),a>0,b\in R \tag{2}$$

其中,a 为缩放因子,b 为转换因子.本文采用离散小波变换(discrete wavelet transform)[15]进行信号重构,令缩放因子 $a=a_0^m(m\in Z,a_0\neq 1)$,转换因子 $b=nb_0a_0^m(n\in Z)$,离散母波函数可表示为

$$\phi_{m,n}(t)=|a_0|^{-m/2}\phi(a_0^{-m}t-nb_0) \tag{3}$$

原始信号 $f(t)$的离散小波变换函数可表示为

$$W_f(m,n)=\langle f(t),\phi_{m,n}\rangle=\int_R f(t)\phi_{m,n}(t)\mathrm{d}t \tag{4}$$

其中,m 为小波分辨率参数,n 为小波偏移参数.通过应用标准正交基(a_0=2,b_0=1)便可实现对原始热度信号 $f(t)$的多分辨率小波分析[16]及去噪的目的.

因此,小波降噪可分为以下几个步骤.

(1) 对原始信号 $f(t)$进行离散小波变换并计算小波系数 $W_f(m,n)$;

(2) 将每个分辨率 m 对应的小波系数与设定的阈值进行比对,由于小波系数的稀疏性,只有高频信号被保留,低频信号将被过滤掉;

(3) 用过滤后的小波系数重建热度信号,信号重建公式可表示为

$$\sum_m\sum_n W_f(m,n)\phi_{m,n}=\sum_m\sum_n\langle f,\phi_{m,n}\rangle\phi_{m,n} \tag{5}$$

在小波去噪过程中,对小波分辨率系数过滤的阈值设置尤为重要,若阈值过高可能把原本较高频的信号误当作噪声,阈值过小又起不到较好的去噪效果.因此,采用基于HeurSure阈值的启发式软阈值函数[17]来作为小波系数过滤的阈值约束,通过对一维小波分解层级的控制实现对阈值的控制,进而通过每层小波分解的分辨率 m 与阈值比较实现噪声信号的过滤.当小波去噪层级较高时,启发式阈值频率较高,此时只有高分辨率的信号被保留和用于信号重构,大量分辨率低于阈值的信号将被认为是噪声信号而被去除,重构后的信号变得更加平滑;反之,当小波去噪层级较低时,启发式阈值对应的频率较低,只有少量分辨率低于阈值的信号被认为是噪声信号而被去除,重构后的信号相对于原始信号只产生微弱的去噪和平滑效果.

图 4 为“Obama”话题热度信号在小波去噪前后的对比结果.可以看到,通过对小波系数过滤阈值的合理设置,去噪后的热度信号曲线相比去噪前更加平滑.原始信号中大量噪声信号被过滤,信号随时序的变化趋势更容易捕捉,以促进热点子事件时间点的准确检测.

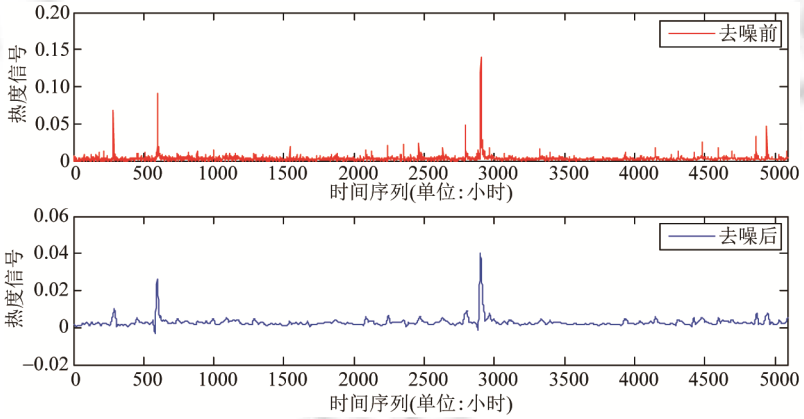


Fig.4 Comparison of “Obama” heat signal before and after wavelet denoising

图 4 “Obama”推特流全局热度信号小波去噪前后对比

3.1.3 基于去噪信号和转折率的热点子事件时间点检测

详细流程如算法 1 所示.首先设置一个固定长度滑动窗口(窗口长度设置为 3 小时),从时间序列起始位置随时序增长方向滑动:若窗口中间位置的热度信号值 $Heat(i)$大于两侧的热度信号 $Heat(i-1)$和 $Heat(i+1)$,则视为窗口中间位置对应的时间点出现了一个“伪峰值点”,并记录该时间点的位置 x;每当有“伪峰值点”出现,则窗口开

始逆向滑动,直至逆向滑动过程中遇到第 1 个低谷,此时计算该低谷的热度值 $Heat(e)$与位置 x 处热度信号 $Heat(i)$的比值 $Heat(e)/Heat(i)$,称该比值为x处伪峰值点的转折率(turning-rate);若转折率小于人工设定的转折率阈值α,则将x加入候选峰值点集合 $RealPeak$,然后窗口从x处继续沿时序增长方向滑动.重复上述过程直至窗口达到时间序列末端.最后将 $RealPeak$ 集合中的时间点按热度值从高到低排序,并依次将每个时间点对应的日期选入结果集合 TP 中,若某时间点对应的日期已入选结果集合,则略过该时间点,最终获得重要时间点对应的日期集合 $TP=\{imp_1,imp_2,\ldots,imp_c\}$,即为热点子事件时间点.

算法 1. 基于去噪信号和转折率的热点子事件时间点检测.

输入:去噪信号 $f'(t)$,起始时间点 $start$,结束时间点 end,热点子事件个数 c,转折率阈值α;

输出:重要时间点日期集合 TP.

```
RealPeak={}; TP={};
FOR i in range(start+1,end–1) DO
    IF Heat(i)>Heat(i–1) and Heat(i)>Heat(i+1): //若出现伪峰值点
        peak_flag="no";
        FOR e in range(i,start+1,–1) DO //窗口逆向回退
            IF Heat(e–1)>Heat(e): //发现第 1 个波谷
                IF Heat(e)/Heat(i)<α:
                    peak_flag="ok";
                    break;
                ENDIF
            ENDIF
        ENDFOR
        IF peak_flag=="ok": //若转折率小于设定的阈值则该伪峰值点入选
            RealPeak=RealPeak∪{i:Heat(i)};
        ENDIF
    ENDIF
ENDFOR
RealPeak.Sort() by Heat Value in descending order;
FOR i in range(0,RealPeak.length()) DO
    IF RealPeak[i].todate() not in TP: //若该时间点对应的日期在集合 TP 中不存在
        TP=TP∪RealPeak[i].date() //将该日期加入集合 TP
    ENDIF
    IF TP.length()==c:
        break;
    ENDIF
ENDFOR
```

3.2 推特流局部时序热度与用户权威度建模

3.2.1 推特摘要的随机步图模型

本文采用无监督图模型为推特摘要算法的基准框架,将每条推特视为图中的节点,推特 tw_i 和 tw_j 所对应的向量 D_i 和 D_j 之间的相似度作为节点 i 和 j 之间边的权值.在随机游走过程中,将边的值视为对应节点 i 到 j 的转移概率 M_{ij},那么一个节点的被访问或转移到的概率等于其余节点到该节点的转移概率加和,我们可以将上述过程视为该节点的一次随机游走过程.当每个节点的随机游走过程迭代多次,所有节点被访问的概率值达到稳态时,我们视一个节点此刻被访问的概率值为该节点对应推特的重要性分值,因而实现对推特的集中排序.其具体

流程如算法 2 所示.

算法 2. 基于随机步图模型的推特摘要算法 TS(Twitter summarization).

输入:推特集合 S,收敛阈值β,摘要句子总单词数 n;

输出:摘要句子集合 S_Result.

```
1:  Words=0,α=0.85;
2:  FOR sentence in S DO
3:      Compute vector D_i using TF-IDF,
4:  ENDFOR
5:  FOR i in range(0,S.length()) DO //构建转移概率矩阵 M
6:      FOR j in range(0,S.length()) DO
7:          M_ij=Sim(D_i,D_j)/∑_j'Sim(D_i,D_j'); //D_j'为与 D_i 相邻的节点对应的向量表示
8:      ENDFOR
9:  ENDFOR
10: FOR i in range(0,INT_MAX) DO
11:     R'=R;
12:     R=α×M×R+(1−α)×v; //用转移概率矩阵更新 Rank 分值向量,α为阻尼系数
13:     IF |R−R'|<β:
14:         break;
15:     ENDIF
16: ENDFOR
17: S.sort() by Rank Value in R in descending order;
18: FOR i in range(0,INT_MAX) DO
19:     S_Result=S_Result∪S[i];
20:     Words+=S[i].length();
21:     IF Words>=n:
22:         break;
23:     ENDIF
24: ENDFOR
```

对某一子事件推特集中排序后,按照最大边缘相关原则(maximal marginal relevance)[18]进行摘要推特的抽取,即选出排名较前的推特,且使后续入选的推特与已入选的推特内容冗余性较低.

3.2.2 推特流时序热度信号的瞬时特性

在获得某一热点子事件数据集后,为捕捉时序热度信号的瞬时特性对推特摘要的影响,进一步地刻画时序热度信号.我们认为,在话题热度较高时间段内发布的推特更可能与主题相关性较强,相反,在话题热度较低时(话题被热议之前和平息之后)发布的推特主题相关性较弱,且更可能包含多个主题信息,不具有高参考价值.

在基于推特内容的摘要方法中,把文本内容相似度作为节点间的转移概率,建立两两节点之间的跳跃强弱关系.在此基础上,若推特 tw_i 发布的时间点具有较高的热度,那么 tw_i 更可能是一条主题相关性较强的推特,其他推特也会有更大的几率向 tw_i 跳跃.定义推特 tw_i 的局部热度信号 $H(tw_i)$.

$$H(tw_i)=\frac{v(p(tw_i))}{\mathrm{MAX}_H} \tag{6}$$

其中,$p(tw_i)$表示推特 tw_i 的发布时间.$v(x)$代表公式(1)中定义的 x 时间点的推特更新速率.MAX_H 表示所有时间窗中推特更新率的最大值(时间窗以小时为单位),用最大值归一化的方式将局部热度信号归一到[0,1]范围.

加入时序热度信号瞬时特性信息的转移概率矩阵可表示为

$$M_{ij}=\begin{cases}\dfrac{\mathrm{sim}(D_i,D_j)\times H(tw_j)}{\sum_{j'}\mathrm{sim}(D_i,D_{j'})\times H(tw_{j'})}, & \sum_{j'}\mathrm{sim}(D_i,D_{j'})\times H(tw_{j'})\neq 0\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

若节点 j 有较高的瞬时热度值,那么其他节点跳跃到节点 j 的概率也相应地会增大.由此得到的推特摘要方法简称为 TST(Twitter summarization with temporal context).

3.2.3 用户社交权威度信号特性

推特平台拥有大量用户属性信息以及用户交互信息,本文从以下两方面假设来考虑用户权威性与其推特文本内容质量间的联系.

- 粉丝数多关注数少的用户是优质用户:粉丝数多表明广大用户对该用户网络言论和行为的信赖和关注程度高.在用户粉丝数较多的情况下,若其关注数也较多,则该用户很可能是一个靠关注行为赚取点击的广告用户;若其关注数较少,则该用户很可能是一个公众影响力较高的“明星”用户.
- 被转发数高的推特是与话题密切相关的推特:一个推特若被大量转发则说明大量用户认为该推特内容与当前热点事件有紧密联系,推特内容代表热点事件的讨论焦点因而受到大量用户转发.

基于上述假设,我们对事件集中的每个推特 tw_i 所属用户定义用户权威度属性 $AS(tw_i)$.

$$AS(tw_i)=\frac{fol(tw_i)}{fri(tw_i)}\times RT(tw_i) \tag{8}$$

其中,$fol(tw_i)$和 $fri(tw_i)$分别是推特 tw_i 所属用户的粉丝数和关注数,$RT(tw_i)$表示推特 tw_i 的被转发数.

由于 $AS(tw_i)$的值可能大于 1,类似第 3.2.2 节,采用最大值归一化方法将权威度属性分值约束到[0,1].每个推特 tw_i 的权威度属性可表示为

$$A(tw)=\begin{cases}\dfrac{AS(tw)}{\mathrm{MAX}}, & \text{if the author of } tw \text{ can be acquired}\\ AS, & \text{otherwise}\end{cases} \tag{9}$$

加入用户社交权威性属性的转移概率矩阵可表示为

$$M_{ij}=\begin{cases}\dfrac{\mathrm{sim}(D_i,D_j)\times A(tw_j)}{\sum_{j'}\mathrm{sim}(D_i,D_{j'})\times A(tw_{j'})}, & \sum_{j'}\mathrm{sim}(D_i,D_{j'})\times A(tw_{j'})\neq 0\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

若节点 j 有较高的瞬时热度值,那么其他节点跳跃到节点 j 的概率也相应地增大.由此得到的推特摘要方法简称为 TSS(Twitter summarization with social context).

3.3 融入推特流时序信息和用户社交信息的推特摘要模型(T2ST)

为了进一步改进推特摘要模型,本文将推特流局部时序热度信息和用户社交权威性信息,与推特文本内容相结合,形成改进的随机步图模型,提出融合了时序和社交上下文语境的 T2ST(Twitter summarization with social-temporal context)推特摘要方法.

由前面的建模分析可知,推特流时序信号和用户社交权威度信号分别从不同视角挖掘推特流特性,对推特摘要的提取也有不同的影响,在建构 T2ST 摘要模型的过程中,本文设计了两种推特流时序信息和用户社交信息的融合方式.

(1) 信号融合法(T2ST-M):将归一化后的推特流局部时序热度信号和社交权威度信号以相乘的方式融入转移概率矩阵 M 中,如公式(11)所示.之后,通过随机游走过程对推特进行排序,结合最大边缘相关去除冗余方法(MMR)进行推特摘要选择,直至满足一定的文摘长度.

$$M_{ij}=\begin{cases}\dfrac{\mathrm{sim}(D_i,D_j)\times H(tw_j)\times A(tw_j)}{\sum_{j'}\mathrm{sim}(D_i,D_{j'})\times H(tw_{j'})\times A(tw_{j'})}, & \sum_{j'}\mathrm{sim}(D_i,D_{j'})\times H(tw_{j'})\times A(tw_{j'})\neq 0\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

(2) 最优池化法(T2ST-Set):从集合论角度出发,考虑将分别融合局部时序热度信号和用户社交权威度信号得到的推特摘要集 TSS_{set} 和 TST_{set} 进行优化.从两个结果集中分别选取得分较高的推特进入最终的摘要集合T2ST-Set.如公式(12)所示.

$$T2ST\text{-}Set=\lambda\times TST_{set}+(1-\lambda)\times TSS_{set},\quad 0\leqslant\lambda\leqslant 1,\ \lambda|TST_{set}|+(1-\lambda)|TSS_{set}|=n \tag{12}$$

其中,$|TST_{set}|$和$|TSS_{set}|$表示摘要结果集的单词总数,λ 为调整文摘长度的比例参数,n 为人工设置的系统输出摘要长度.这一过程也同样采用最大边缘相关(MMR)方法进行冗余的去除,同时满足一定文摘长度的约束.

4 实验结果与分析

4.1 实验设置

4.1.1 数据准备

本文使用的原始推特数据集为 2012 年 Illinois 大学的 Li 等人[19]提供的包含约 300 万个用户的属性信息,147 909 个用户的推特共约 5 000 万条,以及用户之间的关注关系共约 2 亿条.其中,推特时间跨度大致在 2010 年末~2011 年 7 月.依据时序推特摘要的任务目标,我们分两个阶段准备实验数据.

(1) 热点子事件时间点检测数据:选取全球影响力较高的代表性人物、产品及公司关键词“Obama”“iPad”和“Microsoft”,以此为事件线索,从原始数据集中抽取包含关键词的推特流,形成时序推特摘要的输入,其详细统计信息见表 1.

Table 1 Datasets for hot sub-event time point detection

表 1 热点子事件时间点检测数据集信息

话题关键词	数据集大小(MB)	推特数量
Obama	20.7	221 364
iPad	13.6	143 887
Microsoft	15.3	172 664

(2) 推特摘要数据:在热点子事件时间点检测步骤之后,根据时间点的检测结果,抽取每个日期对应的子事件推特数据集作为摘要算法的输入.为尽量降低推特文本的非规范性对摘要性能的影响,我们对每个子事件集进行了常规的文本预处理,如字母大写转小写、句中句末标点符号剔除以及停用词剔除等.考虑到建设推特专家摘要的人工成本较高,本文以其中 4 个子事件数据集为例对推特摘要算法进行评价,详细统计信息见表 2.

Table 2 Datasets for hot sub-event summarization

表 2 热点子事件摘要数据集

话题关键词	子事件时间点	数据集大小(KB)	推特数量
Obama	2011/01/25(Obama-1)	244	1 328
Obama	2011/05/19(Obama-2)	210	1 108
iPad	2011/03/11(iPad)	231	1 308
Microsoft	2011/05/10(Microsoft)	313	1 758

4.1.2 专家时间点与专家摘要

由于在实验结果评价过程中缺乏参考标准,我们采用人工标注的方式为本实验制作评估标准语料.邀请了 9 名有自然语言处理领域研究背景且与本课题无直接联系的研究生制作专家时间点和专家摘要,其中,3 人负责专家时间点,6 人(男女各 3 人)负责专家摘要制作.

(1) 专家时间点:我们向 3 名志愿者说明热点子事件时间点检测任务的目的,并允许其通过阅读传统新闻、社交媒体等方式对每个关键词在 2011 年上半年的热点子事件进行了解,进而确定其发生时间点.每人为每个关键词为时间线索的推特流选 10 个子事件时间点,并且保证 3 人独立完成专家时间点的制作过程.

(2) 专家摘要:首先向 6 名志愿者说明推特摘要任务的目的,并引导志愿者借助相关新闻对事件有一个大致的了解,然后要求志愿者依次阅读推特数据,对每个数据集抽取 10 条推特,且要保证这 10 条推特在能反映事件

主题的前提下,推特间信息冗余度尽量低,信息丰富度尽量高.由于摘要数据集中推特数量较多,为减轻专家摘要制作的代价并提高效率,我们首先对原始数据集进行了处理,如剔除重复推特、包含较多链接的推特以及文本极短的推特等.这里,3 人负责“Obama-1”和“Obama-2”数据集,3 人负责“iPad”和“Microsoft”数据集,各自独立完成专家摘要制作工作.

4.1.3 评价指标

(1) 子事件时间点检测实验评测:我们将其看作是信息检索任务,选择基于平均准确率的 MAP(mean average precision)作为评价指标.MAP 定义为

$$MAP=\frac{1}{n}\times\sum_{i=1}^{n}\frac{Overlap(i)}{i} \tag{13}$$

其中,n 表示系统输出时间点的总数,i 表示系统输出的第 i 个时间点,$Overlap(i)$表示系统输出的前 i 个时间点中与专家时间点匹配的时间点个数.

(2) 推特摘要实验评测:以摘要国际评测任务中常用的 ROUGE[20]准则作为评价指标.ROUGE 算法的原理为:通过计算系统输出与专家摘要 N 元语法模型的匹配程度来衡量系统输出的文本质量.N 元语法模型的 ROUGE 公式定义为

$$ROUGE\text{-}N=\frac{\sum_{m\in MS}\sum_{u\in m}match(u)}{\sum_{m\in MS}\sum_{u\in m}count(u)} \tag{14}$$

其中,MS 代表专家摘要集,m 代表专家摘要每个文档单位中 N 元语法词组集合,u 为 N 元语法词组,$match(u)$代表在专家摘要和系统摘要中共同出现的 N 元语法词组个数,$count(u)$代表每个专家摘要单位中的 N 元语法词组个数.本文采用 ROUGE-1 和 ROUGE-2 作为系统性能的评价指标.

4.2 热点子事件时间点检测模型评估

4.2.1 热点子事件时间点检测总体性能

在时序推特摘要中,热点子事件时间点检测将帮助用户快速捕捉热点事件海量推特流中的关键时刻.在对推特流全局热度信号去噪之后,提出采用基于转折率的方法检测信号中宏观趋势的峰值点,以此选择热点子事件时间点.观察发现,不同类别关键词的推特流各自噪声水平和噪声模式并不一致.因此,模型中如下两个参数将对重要时间点检测的性能有直接影响:(1) HeurSure 的启发式软阈值函数下的小波去噪层级;(2) 热点子事件检测算法中的转折率α,我们将分别对这两个参数进行实验分析.考虑与相关工作的对比,大部分学者采用直接检测峰值的方法,等效于本模型中降噪前α=1 的情形.这里,小波去噪前后热点子事件时间点检测结果见表 3.

Table 3 Performance comparisons of hot sub-event time-point detection before and after wavelet denoising

表 3 小波去噪前后热点子事件时间点选择算法性能对比

	降噪前	*level*=1	*level*=2	*level*=3	*level*=4	*level*=5
α=0.5	0.526	0.532	0.533	0.586	0.562	Peak 不足
α=0.6	0.526	0.557	0.533	0.587	0.535	0.447
α=0.7	0.526	0.557	0.533	0.587	0.534	0.449
α=0.8	0.526	0.557	0.533	0.582	0.517	0.444
α=0.9	0.526	0.557	0.533	0.581	0.510	0.436
α=1.0	0.526	0.557	0.529	0.581	0.517	0.413

从表 3 中可以看到,小波去噪层级和峰值点选择算法的转折率α都对热点子事件时间点检测的结果产生重要的影响.平均来看,小波去噪层级 *level*=3、转折率α=0.7 时重要时间点选择算法获得了最优性能.我们对实验的总体性能分 3 个方面进行讨论.

(1) 与 Gao 等人[1]和 Nichols 等人[2]的工作相比:当α=1.0 时,即转折率不对任何峰值点产生过滤作用时(因为任意一个峰值点的转折率都一定<1.0),只有小波去噪层级对时间点检测模型产生影响,各个小波去噪层级时间点选择与利用未去噪的信号直接进行时间点选择相比,都对时间点选择性能起到提升作用.实际上,未去噪的

信号中存在大量伪峰值点,即由于复杂噪声的影响,未出现显著热点子事件的时间点处也可能出现峰值点,而小波去噪则消除了复杂噪声,推特流更新速度随时序变化的趋势更加明显,因此也更加有利于真正热点子事件时间点的选择.

(2) 不同小波去噪层级的影响:1~5 层小波去噪的时间点检测结果在 3 个数据集上的 MAP 均值皆高于未去噪的信号,且随层级升高呈现先升后降的趋势,在去噪层级为 3 的时候结果较好;由于原始信号存在大量噪声,使得单位时间窗口内信号波动较为剧烈,当去噪层级较低时,小波去噪对原始信号的去噪效果不显著,大多数噪声信号依然被保留;当小波去噪层级较高时,去噪力度过大使得原有的部分真实信号被误认为噪声信号而被一同去除,去噪后的信号过于平滑,失去了原有信号高信噪比的特点.所以当去噪层级为 3 时,去噪性能较好.

(3) 转折率阈值α的影响:当固定小波去噪层级,观察转折率α对时间点检测结果的影响.转折率α在小波去噪层级较高时影响比较显著,且在α=0.7 左右时,时间点检测结果更加准确;而在小波去噪层级小于等于 2 时,α对模型结果并无显著影响,这是由于,在小波去噪层级较低时,小波去噪效果并不显著,去噪后的信号波动较大,依然存在大量伪峰值点,且绝大部分伪峰值点的转折率都要高于人工设定的阈值α;而小波去噪层级较高时,大量噪声信号被过滤,推特流时序热度信号走势更加平滑,伪峰值点数量大为减少,且各个伪峰值点的转折率也有所下降,此时,转折率依然大于阈值α的伪峰值点处信号具有很高的信噪比,因此有更大概率,其为一个真正的热点子事件时间点.

4.2.2 小波去噪层级对不同话题数据集时间点检测性能的影响

为进一步观察小波去噪层级在不同数据集上的表现,固定峰值点选择算法的转折率α=0.7,时间点检测模型的实验结果见表 4.

Table 4 Performance of hot sub-event time-points detection on different topic with different denoising level

表 4 不同话题推特流的热点子事件时间点检测性能随去噪层级变化

	Obama	iPad	Microsoft	AVE
降噪前	0.682	0.372	0.523	0.526
level=1	0.669	0.479	0.523	0.557
level=2	0.710	0.384	0.506	0.533
level=3	0.683	0.444	0.635	0.587
level=4	0.531	0.422	0.649	0.534
level=5	0.169	0.535	0.643	0.449

实验结果表明:

(1) 分别地,“Obama”“iPad”和“Microsoft”数据集的最优小波去噪层级分别是 2、5、4 层,通过 3 个数据集的 MAP 性能均值观察,层级为 3 左右时比较接近各数据集的真实最优值.

(2) 进一步地,“iPad”和“Microsoft”数据集的 1~5 层小波去噪结果的 MAP 性能相比原始信号均有所提高,最优去噪层级比原始信号分别提高了 43%和 24%,证明了小波去噪方法在热点子事件时间点检测中的显著作用.相对而言,“Obama”数据集的表现并不理想,只有第 2、3 层小波去噪的 MAP 性能超过了原始信号,最优去噪层级的性能只比原始信号提高了 4%.观察发现,“Obama”数据集降噪前推特流信号的时间点选择 MAP 性能大幅度高于“iPad”和“Microsoft”数据集,且已达到 0.682 这样一个较高性能水平.这可能是由于“Obama”数据集的推特流整体信噪比相对较高,噪声信号对真实信号的影响较弱,因此小波去噪和热点子事件时间点检测算法在“Obama”数据集的优势并不明显.

(3) 时间点检测性能在每个数据集上随小波去噪层级的变化趋势也不尽相同,如图 5 所示,虽然小波去噪的方法在 3 个数据集上都对热点子事件时间点检测模型的结果有提升作用,但时间点选择的结果随去噪层级的变化并不稳定,而且在个别数据集(如“Obama”数据集)起伏较大,这可能是由于不同类别话题数据集的推特流信息的噪声水平和噪声模式存在差异,因而,需要结合不同话题类别数据的噪声水平和噪声模式对小波去噪参数进行适当调节,以实现最优时间点的选择.

以“Obama”数据集为例,从原始全局热度信号中检测出的前十个峰值点所对应的时间点,以及从经过小波去噪的信号中实施峰值点选择算法之后的时间点选择结果和对应事件见表 5,其中,小波去噪的 *level* 为 2 层,峰值点选择算法的阈值为 0.7.可以看到,原始热度信号的结果中有两个时间点是无显著性事件的时间点,分别是

2011/07/02 和 2011/04/08.其中,2011/07/02 属于 2011/07/01 时间点“奥巴马宣布击毙本拉登”事件的延续,而 2011/04/08 时间点并无某显著事件与“Obama”相关,但由于网友对“Obama”话题相关琐碎杂乱的讨论稍多,使得该时间点热度信号比常规热度信号有一个小的增幅波动,因而也被误当作一个关于“Obama”的热点子事件.

相比较而言,采用小波去噪和峰值点选择算法之后的时间点选择结果则更具有话题代表性:(1) 上述两个时间点被有效过滤,算法可以过滤掉有话题延续性、冗余性以及无显著事件的时间点;(2) 新的有代表性的时间节点被检测出来,如 2011/03/28 和 2011/05/19 皆为奥巴马关于“美国在中东和北非部分区域政治政策的演讲”等活动,而 2011/06/13 所对应的“奥巴马与美国就业和竞争委员会座谈”事件虽然不是奥巴马政治活动中最典型的事件,但也在当时激发了网友较为集中的讨论,该时间点也理应为“Obama”关键词的话题时间轴中较重要的一个时间点,因此,该时间点也被有效地检测出来.

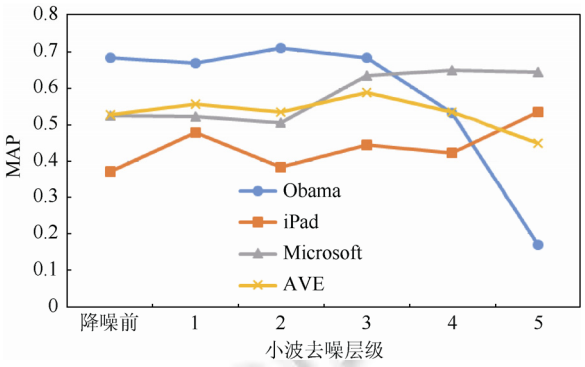


Fig.5 The performance trend of hot sub-event time point detection on different denoising levels

图 5 热点子事件时间点检测性能随去小波噪层级变化趋势

Table 5 Comparison of time point detection results and relevant sub-events before and after wavelet denoising

表 5 小波去噪前后的时间点检测结果和对应子事件对比

小波去噪和热点子事件时间点选择算法之前	时间点对应事件	小波去噪和热点子事件时间点选择算法之后(α=0.7,*level*=2)	时间点对应事件
2011/01/12	关于亚利桑那州枪击事件的演讲	2011/01/12	关于亚利桑那州枪击事件的演讲
2011/01/25	发表国情咨文演说	2011/01/25	发表国情咨文演说
2011/04/08	无显著事件	2011/03/28	在美国国防大学发表关于利比亚政策的演讲
2011/04/13	发表关于美国财政政策的讲话	2011/04/20	标普调整美债评级后,奥巴马强调债务问题的紧迫性
2011/04/27	白宫出示奥巴马完整版出生证明文件	2011/04/27	白宫出示奥巴马完整版出生证明文件
2011/05/01	宣布击毙本拉登	2011/05/01	宣布击毙本拉登
2011/05/02	无显著事件,依旧与宣布击毙本拉登事件相关	2011/05/19	发表关于中东和北非政策的演讲
2011/07/06	通过 Twitter 市政厅和网友进行交流	2011/06/13	奥巴马在科锐公司总部与美国就业和竞争委员会座谈
2011/07/22	奥巴马宣布结束军中同性恋歧视政策	2011/06/22	宣布阿富汗撤军计划
2011/07/25	奥巴马和众议院议长约翰·博纳进行关于债务问题演讲	2011/07/25	奥巴马和众议院议长约翰·博纳进行关于债务问题演讲

综上所述,本文提出的热点子事件时间点检测算法是有效的,但由于不同类别关键词的推特流各自噪声水平和噪声模式不一致,因而对不同类别关键词并不能设置统一的参数.在未来工作中,我们将进一步扩大推特语料和专家语料规模,分别在不同类别话题下补充多个关键词数据集,并在每个类别下分别训练,形成类别自适应的热点子事件时间点检测模型,以更好地捕捉小波去噪参数和时间点选择阈值的规律.

4.3 子事件推特摘要评估

对推特流中每个重要的时间点,本文提出融合社交-时序上下文情景的 T2ST 推特摘要方法.通过将推特流局部时序热度信息和用户社交权威性信息与推特文本内容结合,形成上下文随机步图模型,并结合最大边缘相关去冗余方法(MMR)进行推特摘要抽取.在实验中,首先根据每个重要时间点对应子事件数据集的 3 份专家摘

要,计算专家摘要的平均单词数,并以此作为 T2ST 系统输出推特摘要长度的约束.

4.3.1 推特摘要的对比实验算法

为了验证该方法的有效性,本文设计了相关的对比实验,见表 6.

Table 6 Comparision algorithms for T2ST

表 6 T2ST 推特摘要的对比算法

算法名称	算法原理简介
Random	从给定子事件数据集中随机选取推特作为摘要
Sumbasic[21]	计算整个文档的单词词频分布,每个推特的重要性得分为其包含的单词词频分布之和,以此得分对文档中的推特由高到低排序,最后通过最大边缘相关原则进行摘要句筛选
TS	采用图模型的基本摘要算法,如算法 2 所示
TSS	加入用户社交权威性信息的 TS 模型,如公式(10)所示
TST	加入局部时序热度信息的 TS 模型,如公式(7)所示
T2ST-M	采用信号融合法直接融合用户社交信息与局部时序热度信息的摘要模型,如公式(11)所示
T2ST-Set	采用最优池化法融合用户社交信息与局部时序热度信息的摘要模型,如公式(12)所示
SumBasic+S	在 SumBasic 模型中加入用户权威性信息,即在每个推特分值基础上乘以归一化的用户权威性分值,具体分值如公式(9)所示
SumBasic+T	在 SumBasic 模型中加入局部时序热度信息,即在每个推特分值基础上乘以归一化的局部时序热度值,具体分值如公式(6)所示
SumBasic+ST-M	采用信号融合的方式在 SumBasic 模型中同时加入用户权威性信息和局部时序热度信息,具体模式如公式(11)所示
SumBasic+ST-Set	采用池化的方式在 SumBasic 模型中加入用户权威性信息和局部时序热度信息,具体模式如公式(12)所示

4.3.2 T2ST 及对比实验算法性能评估

在 4 个子事件数据集上,我们分别用上述方法进行了实验,并计算了各个系统输出的 ROUGE-1 和 ROUGE-2 值,见表 7 和表 8.

Table 7 ROUGE-1 performance of T2ST and comparision algorithms

表 7 T2ST 及对比算法的 ROUGE-1 性能

ROUGE-1	Obama-1	Obama-2	iPad	Microsoft	AVE	IPR(%)
Random	0.278 24	0.278 46	0.323 92	0.392 31	0.318 23	–
TS	0.324 27	0.363 82	0.331 45	0.384 62	0.351 04	Baseline
TSS	0.349 37	0.384 15	0.363 47	0.417 95	0.378 74	+7.779
TST	0.332 64	0.382 11	0.361 58	0.397 44	0.368 44	+4.850
T2ST-M	0.366 11	0.410 57	0.354 05	0.397 44	0.382 04	+8.720
T2ST-Set	0.366 11	0.390 24	0.380 41	0.428 21	0.391 24	+11.338
Sumbasic	0.315 90	0.333 33	0.335 22	0.346 15	0.332 65	Baseline
SumBasic+S	0.380 75	0.310 98	0.337 10	0.423 08	0.362 98	+9.117
SumBasic+T	0.297 07	0.339 43	0.352 17	0.348 72	0.334 35	+0.510
SumBasic+ST-M	0.366 11	0.308 94	0.337 10	0.428 21	0.360 09	+8.249
SumBasic+ST-Set	0.359 83	0.282 52	0.329 57	0.407 69	0.344 90	+3.683

Table 8 ROUGE-2 performance of T2ST and comparision algorithms

表 8 T2ST 及对比算法的 ROUGE-2 性能

ROUGE-2	Obama-1	Obama-2	iPad	Microsoft	AVE	IPR(%)
Random	0.048 42	0.040 90	0.049 24	0.077 52	0.054 02	–
TS	0.067 37	0.098 16	0.107 95	0.124 03	0.099 38	Baseline
TSS	0.094 74	0.118 61	0.121 21	0.149 87	0.121 11	+21.866
TST	0.069 47	0.104 29	0.138 26	0.134 37	0.111 60	+12.296
T2ST-M	0.094 74	0.114 52	0.142 05	0.142 12	0.123 36	+24.130
T2ST-Set	0.103 16	0.126 79	0.134 47	0.149 87	0.128 57	+29.377
Sumbasic	0.071 58	0.055 21	0.115 53	0.103 36	0.086 42	Baseline
SumBasic+S	0.117 89	0.042 94	0.123 11	0.160 21	0.111 04	+28.486
SumBasic+T	0.046 32	0.055 21	0.113 64	0.103 36	0.079 63	–7.854
SumBasic+ST-M	0.107 37	0.042 94	0.123 11	0.155 04	0.107 12	+23.947
SumBasic+ST-Set	0.098 95	0.034 76	0.100 38	0.129 20	0.090 82	+5.094

主要实验观察如下.

(1) Random 摘要算法:由于 Random 摘要方法完全采取随机抽取的方式,推特内容信息以及推特流时序特性和用户权威度特性皆未被利用,所以抽取出的推特虽是包含主题关键词的推特,但在话题相关度及内容丰富度方面均有欠缺,因此,ROUGE-1 和 ROUGE-2 指标性能相对于基准推特摘要方法 TS 和 Sumbasic 均较低.

(2) TS 系列对比实验:在 4 个数据集的 ROUGE-1 和 ROUGE-2 上,加入推特流局部时序信号信息和用户权威度信息都对基准推特摘要算法 TS 性能有提升作用,验证了用户属性信息和社交信息以及推特流局部热度信息的确对推特文本摘要的抽取有积极的参考价值.将用户权威度信息和推特流局部时序热度信号结合起来对推特摘要性能也有提升,且提升度要高于单一特征的摘要方法.

在两种特征融合方法中,采用最优池化法在大部分子事件集中比信号直接融合的方式的效果更好.更进一步地,T2ST-M 特征融合法的摘要性能虽然相比 TS 方法有所提高,但其相对于单纯利用用户权威度信息的 TSS 方法并未有进一步的提升,这说明,虽然用户权威度特征和推特流时序信息特征都对推特摘要的抽取起到积极作用,但二者在同一摘要框架下的共存互助方式还有待进一步探究,而 T2ST-Set 最优池化的方法不仅对基准推特摘要框架(TS)有显著改进,且相对于融入单一用户权威度特征(TSS)和融入单一推特流时序特征(TST)的方式也有显著提升.原因在于,T2ST-Set 将两种上下文信息分别融合到图模型的随机游走过程,并采用池化的方式从两个排序后的推特集中择优抽取摘要,从而避免了 T2ST-M 方法相对粗暴的特征融合方式.

(3) SumBasic 系列对比实验:除了“ipad”数据集外,单纯的 SumBasic 方法在大部分数据集中的性能皆低于 TS 方法.这是由两个原因造成的:其一,SumBasic 方法是一种基于词频分布的摘要抽取方法,社交媒体文本具有口语化的特点,用户对同一事件信息描述的多样性会产生“一义多词”现象,这使得整个推特集的词表长度拉长,真正代表事件主题的核心词词频有所降低,导致整个词表稀疏性增大;其二,推特文本普遍具有较短的特点,使得单词词频特征的表达力度降低,仅依据每条推特所包含单词的词频分布来进行推特的重要性排序,使得 SumBasic 方法相对于 TS 方法对社交媒体文本的鲁棒性有所降低.

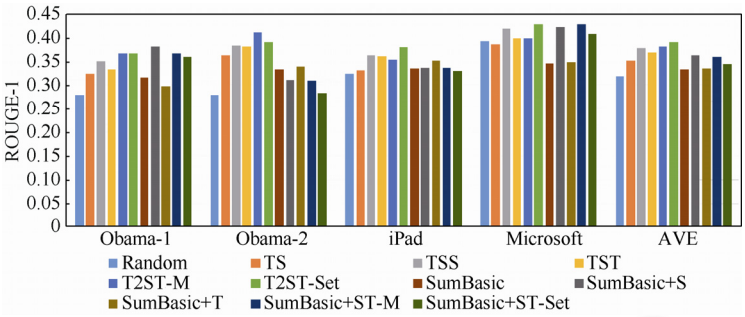
另外,可以看到,在“obama-1”和“Microsoft”数据集中,加入社交上下文的 SumBasic+S 方法以及加入时序社交上下文的 SumBasic+ST-M 和 SumBasic+ST-Set 方法的性能均有所提升.这是因为,时序社交上下文信息独立于词频特征而存在,弥补了词表的稀疏和推特文本较短的缺陷;在“obama-2”和“ipad”数据集中,加入时序社交上下文信息未能使 SumBasic 方法性能有明显提升,甚至略低于单纯的 SumBasic 方法.这可能是由于时序-社交上下文语境蕴含在大量的动态交互信息中,将词频分布和动静态上下文信息直接耦合,未能捕捉到推特间的内在联系,使得该方法不能获得词的自然属性,因此整个模型只在部分数据集有较好表现,普适性较低.

为了更清晰地观察实验效果,我们给出了 T2ST 及各推特摘要对比算法在 4 个不同推特数据集的性能可视化结果,如图 6 所示.可以看到,在 4 个数据集中,T2ST-M 和 T2ST-Set 方法的 ROUGE-1 和 ROUGE-2 性能均比 TS 方法有所提升,其中,ROUGE-1 性能的提升率为 10%左右,ROUGE-2 性能的提升率为 25%左右.另外,SumBasic+S 和 SumBasic+ST-M 算法在“Obama-1”和“Microsoft”数据集中相对于 Sumbasic 方法也有提升,这表明,社交和时序上下文信息在部分数据集中也对 SumBasic 摘要模型起积极作用.但将时序上下文信息单独融入 SumBasic 模型的方法表现较差,甚至性能低于不加任何上下文信息的 SumBasic 方法,这可能由于用户社交权威度信号相对于推特流时序信号更适用于 SumBasic 方法,而如何将社交和时序上下文信息更合理地与 SumBasic 方法进行结合还有待进一步的探究.

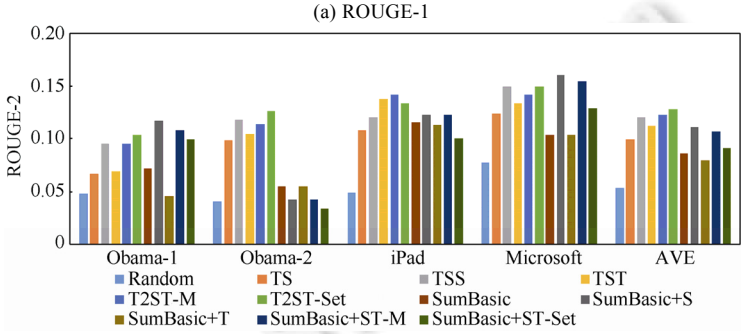
4.3.3 推特摘要示例

为了更直观地观察 T2ST 算法的优势,这里以“Obama-1”数据集为例,将 TS 算法和 T2ST-Set 算法的系统输出摘要以及专家摘要示例进行展示,见表 9.可以看到,

(1) 单纯基于内容的推特摘要算法 TS 抽取出的推特虽然可以向读者反映“奥巴马发表国情咨文演说”这一事件的大致情况,但从这部分推特中,读者只能了解到“国情咨文演说”事件的开始时间和网络直播地址等周边信息,很难对“国情咨文演说”的具体内容有所了解.且 TS 算法选出的推特中,存在用户个人情感的抒发,这些包含个人意见和情感的推特并不利于读者对事件的客观了解.



(a) ROUGE-1



(b) ROUGE-2

Fig.6 Performance comparison of algorithms in different Twitter datasets

图 6 算法在不同推特数据集上的性能对比

Table 9 Summary of "Obama-1" topic in TS, T2ST-Set and Expert

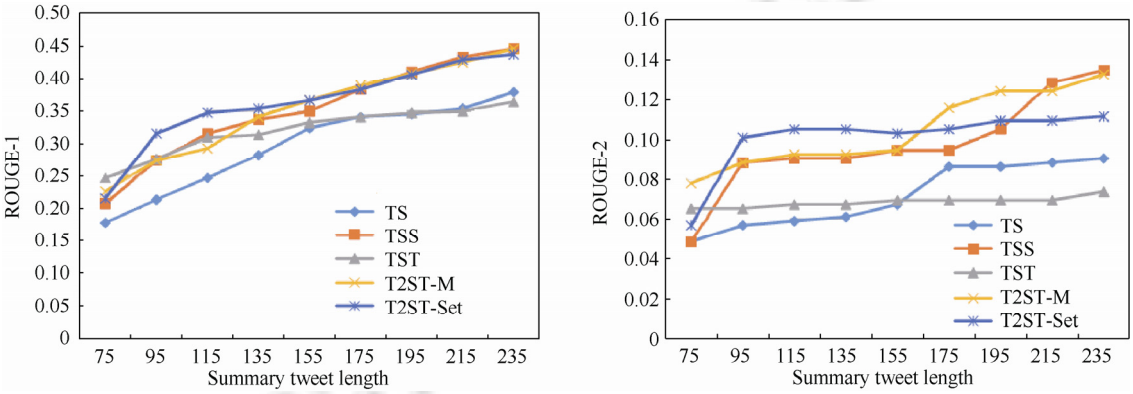
表 9 "Obama-1"话题在 TS 和 T2ST-Set 算法的摘要结果及专家摘要示例

算法	推特摘要
TS	1) Reacts to president obama's state of the union address. video: 2) Watch president obama's state of the union live [video] - facebook, google and more in this year's speech. 3) If anyone wld lk 2 watch president obama's state of the union address tuesday at the nbc action news station in call (816) 932-4141 4) What do you want obama to say tonight? - the state of the union is an annual occasion in which a president receives... 5) Tune in tonight at 8:00 pm cst as president obama delivers his state of the union address. watch it live here: 6) The text of president obama's state of the union address 7) Michelle obama phat sheeeeesh lol 8) Great speech president obama. i enjoyed it and it was so special to me. i love president obama 9) What president obama needs to do tonight - glenn thrush - politico.com 10) What do you want to hear from president obama at the state of the union?
T2ST-Set	1) President obama's state of the union address streams live on hulu tonight at 6p, with coverage now in progress 2) Interesting. no mention of the poor or even middle class in president obama's speech 3) We want to reward good teachers and stop making excuses for bad ones. obama: yes please!!! 4) There's a lot of discussion about the author of the new "o" book on pres. obama. i want to end any speculation now: it wasn't me. 5) Obama: "I'm asking congress to eliminate the billions in taxpayer dollars we currently give to oil companies." 6) Obama: "we need to out-innovate, out-educate, and out-build the rest of the world." 7) The full text of president barack obama's state of the union address: -cc 8) Obama:..innovation that created..millions of new jobs. this is our generation's sputnik moment." 9) Michelle obama phat sheeeeesh lol 10) Full text of president obama's state of the union speech 11) Senator stabenow's reaction to president obama's state of the union address
Expert	1) Paul Ryan hit a home run and decimated just about every argument Obama made. That's how you give a speech. 2) So far, Obama has praise Muslims, illegal immigrants, and unions--the three biggest problems facing the United States. 3) We commend and the RSC for De-Funding ObamaCare in their Spending Reduction Act! 4) RT President Obama co-opted much of the GOP agenda in his speech. Now let's see if he'll walk the talk. 5) RT Obama: "We Are a Nation of Google and Facebook" - 6) RT Obama in asks Congress to make permanent 4-year college tax credit 7) Obama gives a major shoutout to repeal and calls for college campuses to welcome ROTC and military recruiters. 8) RT Obama in says he's "willing to look at" medical malpractice reform 9) Obama on clean energy, biofuels, cutting oil subsidies, complicated salmon management, and high speed rail 10) Members of Congress less keen on Obama's proposal to freeze domestic spending for next five years

(2) 加入推特流局部时序信号和用户权威性信息后的推特摘要算法 T2ST-Set 输出的摘要推特所表达的内容信息更全面和丰富,包含事件周边信息的同时,也体现了奥巴马在国情咨文演说中的重要演讲内容,且对事件各方位信息的描述更倾向于事件主题内容的客观陈述.此类推特被选作摘要也更有助于事件主题和内容信息的展示.因此相比于 TS 算法,一个事件的陌生者通过 T2ST-Set 算法的输出,可以更好地对该事件有快速、客观和全面的了解.

4.3.4 推特摘要长度对性能的影响

为观察系统摘要长度对性能的影响,对系统摘要长度参数设置进行了实验.由于 ROUGE 是基于召回率的评价指标,在专家摘要总单词数恒定的前提下,系统摘要输出的总单词数 n 的取值将对 ROUGE 指标的结果有很大影响.我们使系统摘要总单词数在 75~235 范围内,以 20 单词为间隔分别在不同摘要算法下进行了 9 组实验,系统性能随摘要长度的变化趋势如图 7 所示.



(a) ROUGE-1 (b) ROUGE-2

Fig.7 The influence trend of summary length in different summarization method

图 7 系统摘要长度约束对推特摘要性能的影响趋势

可以观察到:

(1) ROUGE-1 指标中,随着系统摘要长度的增长,摘要性能存在先快速升高再缓慢升高的趋势,这个转折点位置恰为专家摘要平均长度的位置.

(2) ROUGE-2 指标中,在专家摘要平均长度的位置之后,TSS 和 T2ST-M 方法的摘要性能依然有上升的趋势,也就是说,融入了用户社交信息和推特流时序信息的随机步图模型将更多、更高质量的推特推送到了摘要候选序列顶端,当系统摘要长度继续扩大时,会继续有高质量的推特入选结果集,这也体现了用户社交信息和推特流时序信息在摘要算法中的积极作用.

因此,我们在实验中将系统摘要长度控制在 3 份专家摘要平均长度左右,尽可能保证系统摘要的最佳性能.

5 总结与展望

时序推特摘要是自动文摘领域的新问题.本文借助社交媒体信息的优势,融合了推特流时序信息和用户属性及交互上下文语境信息,提出基于小波分析和上下文随机步图模型的时序推特摘要方法.主要贡献包括:(1) 我们将推特流看作是一种信号,通过对全局热度信号进行小波去噪,进一步提出基于去噪信号和转折率的热点子事件时间点检测模型,有效地降低了推特流中复杂噪声对时序推特摘要关键时刻捕捉的干扰;(2) 融合了推特流热度信号中局部瞬时特性和社交用户属性及交互信息,提出基于上下文随机步图模型的推特摘要框架,一定程度上弥补了短文本由于内容稀疏性和非结构化的劣势对推特内容质量判断造成的不利影响;(3) 在推特流真实数据集上建设了专家时间点和专家摘要,以进行模型的实验和评估.时序推特摘要不同阶段的实验

结果表明了本文方法的有效性.

另外,研究发现:(1) 不同类别关键词的推特流各自噪声水平和噪声模式并不一致,在小波去噪时不适宜采用统一的参数和阈值,未来可探索建立关键词话题类别自适应的小波去噪参数和时间点选择阈值,设计更健壮的热点子事件时间点检测算法;(2) 可挖掘更丰富的用户属性信息和交互结构信息进行权威度的建模;(3) 如何将社交媒体文本新特征更合理地融合到时序推特摘要模型中,使其与推特文本内容相辅相成,生成更高质量的推特摘要,还有待进一步探索和实践.

致 谢 衷心感谢审稿专家的悉心指导及本刊编辑的辛勤工作.

References:

[1] Gao DH, Li WJ, Cai XY, Zhang RX, Ouyang Y. Sequential summarization: A full view of Twitter trending topics. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2014,22(2):293–302. [doi: 10.1109/TASL.2013.2282191]

[2] Nichols J, Mahmud J, Drews C. Summarizing sporting events using Twitter. In: Proc. of the 2012 ACM Int'l Conf. on Intelligent User Interfaces. ACM, 2012. 189–198. [doi: 10.1145/2166966.2166999]

[3] Zubiaga A, Spina D, Amigó E, Gonzalo J. Towards real-time summarization of scheduled events from twitter streams. In: Proc. of the 23rd ACM Conf. on Hypertext and Social Media. ACM, 2012. 319–320. [doi: 10.1145/2309996.2310053]

[4] Duan YJ, Chen ZM, Wei FR, Zhou M, Shum HY. Twitter topic summarization by ranking tweets using social influence and content quality. In: Proc. of the 24th Int'l Conf. on Computational Linguistics. 2012. 763–780. http://www.aclweb.org/anthology/C12-1047

[5] Chang Y, Wang XH, Mei QZ, Liu Y. Towards Twitter context summarization with user influence models. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. ACM, 2013. 527–536. [doi: 10.1145/2433396.2433464]

[6] Li J, Gao W, Wei ZY, Peng BL, Wong KF. Using content-level structures for summarizing microblog repost trees. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 2168–2178.

[7] Shen C, Liu F, Weng FL, Li T. A participant-based approach for event summarization using Twitter streams. In: Proc. of the NAACL-HLT. Atlanta: Association for Computational Linguistics, 2013. 1152–1162.

[8] Olariu A. Hierarchical clustering in improving microblog stream summarization. In: Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer-Verlag, 2013,7817:424–435. [doi: 10.1007/978-3-642-37256-8_35]

[9] Kim TY, Kim J, Lee J, Lee JH. A tweet summarization method based on a keyword graph. In: Proc. of the 8th Int'l Conf. on Ubiquitous Information Management and Communication. ACM, 2014. 1–8. [doi: 10.1145/2557977.2558045]

[10] Inouye D, Kalita JK. Comparing Twitter summarization algorithms for multiple post summaries. In: Proc. of the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE the 3rd Int'l Conf. on Social Computing (SocialCom). 2011. 298–306. [doi: 10.1109/PASSAT/SocialCom.2011.31]

[11] Nenkova A, Vanderwende L. The impact of frequency on summarization. Technical Report, MSR-TR-2005-101, Washington: Microsoft Research, 2005.

[12] Radev DR, Jing HY, Styś M, Tam D. Centroid-Based summarization of multiple documents. Information Processing & Management, 2004,40(6):919–938. [doi: 10.1016/j.ipm.2003.10.006]

[13] Erkan G, Radev DR. LexRank: Graph-Based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 2004,457–479. [doi: 10.1613/jair.1523]

[14] Aslam J, Diaz F, Ekstrand-Abueg M, McCreadie R, Pavlu V, Sakai T. TREC 2014 temporal summarization track overview. 2015. http://trec.nist.gov/pubs/trec23/papers/overview-tempsumm.pdf

[15] Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. on Information Theory, 1990,36(5):961–1005. [doi: 10.1109/18.57199]

[16] Mallat SG. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Trans on Pattern Analysis and Machine Intelligence, 1989,11(7):674–693. [doi: 10.1109/34.192463]

[17] Misiti M, Misiti Y, Oppenheim G, Poggi J-M. Wavelet Toolbox. MA: The MathWorks Inc., 1996.

[18] Carbonell J G, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 1998. 335–336. [doi: 10.1145/290941.291025]

[19] Li R, Wang SJ, Deng HB, Wang R, Chang K. Towards social user profiling: Unified and discriminative influence model for inferring home locations. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2012. 1023–1031. [doi: 10.1145/2339530.2339692]

[20] Lin CY. Rouge: A package for automatic evaluation of summaries. In: Proc. of the ACL-04 Workshop. Barcelona: Association for Computational Linguistics, 2004. 8.

[21] Vanderwende L, Suzuki H, Brockett C, Nenkova A. Beyond SumBasic: Task-Focused summarization with sentence simplification and lexical expansion. Information Processing & Management, 2007,43(6):1606–1618. [doi: 10.1016/j.ipm.2007.01.023]

于广川(1992－),男,辽宁东港人,硕士生,主要研究领域为时序文本摘要,社会媒体计算.

刘洋(1991－),男,博士生,CCF 学生会员,主要研究领域为篇章分析,语义关系分类.

贺瑞芳(1979－),女,博士,副教授,CCF 专业会员,主要研究领域为自然语言处理,社会媒体挖掘,机器学习.

党建武(1956－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为语音科学,言语生成与感知,语音识别,神经生理建模.