ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 3,May 2017,pp. 523—532
DOI:10. 16337/j. 1004-9037. 2017. 03. 011

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

基于滑动窗口的微博时间线摘要算法

徐 伟 赵 斌 吉根林

(南京师范大学计算机科学与技术学院,南京,210023)

摘 要: 时间线摘要是在时间维度上对文本进行内容归纳和概要生成的技术。传统的时间线摘要主要研究诸如新闻之类的长文本,而本文研究微博短文本的时间线摘要问题。由于微博短文本内容特征有限,无法仅依靠文本内容生成摘要,本文采用内容覆盖性、时间分布性和传播影响力3种指标评价时间线摘要,并提出了基于滑动窗口的微博时间线摘要算法(Microblog timeline summariaztion based on sliding window, MTSW)。该算法首先利用词项强度和熵来确定代表性词项;然后基于上述3种指标构建出评价时间线摘要的综合评价指标;最后采用滑动窗口的方法,遍历时间轴上的微博消息序列,生成微博时间线摘要。利用真实微博数据集的实验结果表明,MTSW算法生成的时间线摘要可以有效地反映热点事件发展演化的过程。
关键词: 微博摘要;时间线摘要;短文本摘要;事件演化
中图分类号: TP391 **文献标志码:** A

Microblog Timeline Summarization Algorithm Based on Sliding Window

Xu Wei, Zhao Bin, Ji Genlin

(School of Computer Science and Technology, Nanjing Normal University, Nanjing, 210023, China)

Abstract: Timeline summarization is the process of creating summaries towards topic information and development over time in natural language processing. Some algorithms are proposed to generate summaries towards long text like news, but seldom focus on timeline summaries of short text like microblog. Here, we propose a microblog timeline summarization based on sliding window (MTSW), which simultaneously incorporates content coverage, temporal distribution and influence to evaluate candidate timeline summaries. In the algorithm, representative terms are selected to represent microblog feature according to intensity of terms and entropy. We build a comprehensive indicator for evaluating the timeline summary based on the above three indicators. Then, we use sliding window to generate microblog timeline summary. Experiments on the real-world event datasets verify the effectiveness of the proposed method.
Key words: microblog summary; timeline summary; short text summary; event evolution

引 言

随着社交媒体的蓬勃发展,微博正成为人们获取信息、交流思想和分享兴趣爱好的重要平台。不同

基金项目:江苏省高校自然科学基金(13KJB520014)资助项目。
收稿日期:2014-12-05;**修订日期:**2015-06-20

于传统媒体,其独特的传播性和交互性赋予了用户更多的话语权,使得用户既是信息的接受者,也是信息的发布者和传播者。每当热点事件发生时,众多用户借助微博平台参与讨论,发表个人观点和表达自身关切。伴随事件持续发展,大量个人意见和评论在微博平台上逐渐汇聚融合形成群体观点,这样的群体性意见是社会舆论的重要组成。所以分析微博热点事件的群体性话题、构建微博话题摘要,是一个具有理论意义和应用价值的研究课题。以微博为代表的社交媒体的摘要研究是自然语言处理中的研究热点。摘要是对文本集进行内容归纳和概要生成的技术。按照摘要组织结构的不同摘要可以分为 3 种:用集合形式组织摘要[1-3]、用序列形式组织摘要[4]和用层次结构组织摘要[5]。目前大部分的摘要研究主要选择集合形式。文献[6]采用上、下文敏感的 PageRank 算法选择主题关键字,以此为基础生成候选主题短语,结合相关性和有趣性标准对主题短语进行排序生成微博摘要。文献[7]提出了一种两阶段摘要框架,先将微博分为提问、讨论、分享和聊天 4 类,然后分别对每一类别采用不同策略进行摘要处理。文献[2]利用新闻和微博的互补信息,通过构建一个无监督的联合主题模型来生成热点事件摘要。文献[8]关注个性化的微博摘要,提出一种时间感知的用户行为模型(Tweet propagation model, TPM)对微博主题和用户兴趣的动态特性进行建模,然后按照新颖性、覆盖性和多样性 3 个标准选取用户感兴趣的微博作为摘要。近年来,时间线摘要(Timeline)逐渐成为摘要研究的新方向。其基本思想是利用文本的时间特征辅助识别代表性语句用以构建基于时间维度的摘要,这样的摘要可以反映事件或者话题演化发展的完整过程。文献[4]定义了一种反映事件发展全过程的新摘要形式—序列摘要(Sequential summarization),采用基于消息流和语义两种方法识别话题摘要信息。文献[9]研究了大规模实时微博消息流的摘要问题,采用一种新的微博排序方法(Tweet cluster vector rank, TCV-Rank)生成微博的在线摘要和历史摘要,进而以此为基础自动产生微博消息流的时间线。文献[10]利用外时间(Inter-date)和内时间(Intra-date)的语句依存关系,在时间维度上对摘要相关性进行建模,提出了结合全局摘要和局部摘要的排序框架,用以生成新闻事件的时间线。文献[11]提出一种演化层次狄利克雷过程模型(Evolutionary hierarchical dirichlet process, EHDP),用来在时间线摘要中获取话题的演化模式。文献[12]总结了新闻事件在时间维度上的 3 种特征分别是时间跨度重要性、上下文时间跨度重要性和语句的时间覆盖密度,然后采用监督学习的方法提高多文本摘要的质量。文献[13]采用新闻事件中的主题层次、因果、时间和空间 4 种关系,针对特定事件生成结构化摘要,帮助用户理解新闻事件的演化过程。传统的时间线摘要研究以新闻数据为主,而本文以微博消息为研究对象研究基于时间轴的热点事件摘要方法。由于微博在数据特征上与新闻差异明显,因而新闻的时间线摘要方法无法直接应用于微博,具体理由如下:(1) 长文本的摘要方法不适用于短文本。传统的话题摘要方法主要针对新闻这样的长文本,在摘要生成过程中充分利用新闻的文本特征来度量文本的重要性。但是微博文本长度较短,文本特征不足,仅依靠文本特征很难完成摘要任务。所以微博时间线摘要还需要考虑非文本特征。(2) 新闻在话题完整性上高于微博。新闻的文本按照时间顺序发布,通常组织成序列结构。由于新闻的文本具有独立、完整和新颖的特点,因而相邻文本在内容上差异性明显,主题重复较少,话题完整性高。但是微博,尤其是转发消息是通过媒体传播的方式发布信息,按照图结构组织文本。由于任何微博文本都属于某个话题的上、下文,因而单一文本无法独立传达完整的话题信息。所以微博消息在话题完整性上的缺点妨碍了时间线摘要任务的完成。(3)新闻时间线摘要的评价方法无法应用于微博。新闻摘要普遍采用文档理解会议(Document understanding conferences, DUC)提供的数据集进行测评。该数据集以新闻数据为主,包含用于评价摘要算法的标准摘要结果。但是到目前为止社交媒体数据(包括微博)仍没有权威成熟的评测数据集的评测方法。所以无法采用权威的 DUC 数据集评价微博时间线摘要算法。

目前微博时间线摘要研究的相关工作较少。文献[4]通过识别微博文本流中的子话题生成时间线摘要,利用子话题引发的"激增"区域来划定时间线上的摘要抽取区域,但是摘要区域只有当子话题在消息总量和内容变化上非常明显时才被接纳,使得一些不够明显的摘要被忽略,造成最终摘要在表现事件

发展上的不完整。研究表明,仅依靠单一的文本特征无法满足微博时间线摘要的要求,只有结合文本特征和非文本特征才能提高时间线摘要质量。本文提出了一种基于滑动窗口的微博时间线摘要算法(Microblog timeline summarization based on sliding window,MTSW),该算法综合考虑微博消息的文本内容、时间分布和传播影响力等因素,在对消息的代表性和重要程度进行评估的基础上,采用滑动窗口的方法通过遍历时间轴上的微博消息序列,解决了现有微博时间线摘要表现事件发展不完整性的问题,从而构建出完整的微博热点事件时间线摘要。

1　微博时间线摘要算法

时间线摘要是自然语言处理中自动摘要研究方向上的新问题。它对于按时间序组织的文本集进行“提炼”处理,选取代表性的文本作为摘要,反映事件发展演化的过程。目前学术界主要关注新闻时间线的研究,而微博时间线的研究还比较少。这两种文本的共同点是都具有明确的时间属性,而且消息文本都可以按照时间顺序进行组织。设微博消息序列为 $\boldsymbol{W}=\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n\}$,其中 $\boldsymbol{w}_i=(w_{c_i},w_{t_i},w_{d_i})$,$w_{c_i}$ 为微博 $\boldsymbol{w}_i$ 的文本内容,文本长度记为 $|w_{c_i}|$;w_{t_i} 为微博 w_t 的时间,时间的基本单位为“日”;w_{d_i} 为微博 $\boldsymbol{w}_i$ 的转发量。而微博时间线摘要序列为 $\boldsymbol{S}=\{\boldsymbol{s}_1,\cdots,\boldsymbol{s}_b\}$,其中 $\boldsymbol{s}_i=(s_{c_i},s_{t_i},s_{d_i})$,$\boldsymbol{s}_i\in\boldsymbol{W}$,$s_{c_i}$ 为微博的文本内容,文本长度记为 $|s_{c_i}|$;s_{t_i} 为微博 $\boldsymbol{s}_i$ 的时间;s_{d_i} 为微博 $\boldsymbol{s}_i$ 的转发量。文中所有序列和集合的规模都记为 $|\cdot|$。微博的时间线摘要问题描述为:给定热点事件的微博消息序列 $\boldsymbol{W}$,生成指定长度为 b 的时间线摘要 $\boldsymbol{S}$。时间线 $\boldsymbol{S}$ 应该反映出热点事件随时间发展演化的过程,而且要求时间线 $\boldsymbol{S}$ 中的摘要具有较低的重复性。

本文充分考虑了微博文本的特点和热点事件在微博平台上传播的特征,设计微博时间线摘要方法。每当热点事件发生,众多网络用户借助微博平台参与讨论、发表观点和表达自身关切问题。伴随热点事件持续发展,由用户交互产生的意见和评论逐渐汇聚融合形成群体观点。微博时间线摘要的主要任务就是提取代表性的微博消息作为摘要,反映热点事件在微博平台上的演化发展过程。针对微博摘要任务的上述要求,构建时间线摘要的处理框架。如图 1 所示,包含 4 个阶段:

(1)采集微博数据:根据事件的关键字和微博的时间标签收集指定微博热点事件的消息。

(2)文本预处理:微博消息的序列化处理、文本分词、停用词过滤和代表性词项选择等。

(3)时间线摘要:根据微博特征设立摘要评价指标,采用基于文本流的处理方式在时间轴上选取最优的消息组合构成时间线摘要。

(4)摘要结果展示:采用可视化技术展示时间线摘要。

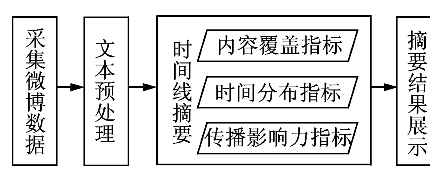

图 1　微博时间线摘要框架结构图

Fig. 1　Framework of microblog timeline summarization

1.1　微博文本预处理

微博文本预处理的主要任务是选择代表性词项来代表微博特征。微博消息是一种用户生成内容(User-generated content,UGC)。不同于新闻,微博消息不仅长度短而且文本质量低下,其中包含未登录词、错别字、谐音字和缩写等。所以,微博在分词之后会产生大量的各类词语,大致可以划分成 3 类:停用词、代表性词和非代表性词。而只有代表性词才能反映热点事件的话题信息,也才能有助于识别高价值微博消息和评价微博消息的重要性。由此可见在预处理阶段,从分词结果中识别代表性词项非常重要。本文,所有微博消息采用词项集合形式表示。微博消息 w 的词项集记为 $E(w)$,而代表性词项集记为 $R(w)$,$R(w)\subset E(w)$。所有词项的集合记为 $E=\sum\limits_{w\in\boldsymbol{W}}E(w)$,所有代表性词项的集合记为 $R=\sum\limits_{w\in\boldsymbol{W}}R(w)$。代表性词项是话题信息的基本载体,它随着热点事件话题的演化发展而交替出现。因此词项在时间轴上的分布具有明显的区域特征。例如在“马航”事件中,“越南”和“澳大利亚”这两个词项分别出现在整个事件的早期和晚期,并且它们在微博讨论中都被高频率地

提及，如图2所示。所以，根据话题演化的特点可以归纳出代表性词项的两个基本特征：高强度和爆发性。

代表性词项的高强度指词项在某些时刻被高频率地提及，并且超过了指定阈值。这样的词项具有话题的代表性，应该被保留下来。由于词项在时间轴上可能出现在多个不同时刻，因而本文在多时刻中选择最高的频率代表该词项的强度，则

$$\mathrm{Freq}(\boldsymbol{e})=\max_{wt_1\leqslant p\leqslant wt_n} f(\boldsymbol{e},p) \tag{1}$$

$$f(\boldsymbol{e},p)=\frac{|\{\boldsymbol{w}_i \mid \boldsymbol{e}\in E(\boldsymbol{w}_i)\ \wedge\ p=wt_i,\boldsymbol{w}_i\in \boldsymbol{W},i=1,\cdots,n\}|}{|\{\boldsymbol{w}_j \mid p=wt_j,\boldsymbol{w}_j\in \boldsymbol{W},j=1,\cdots,n\}|} \tag{2}$$

式中：$\mathrm{Freq}(\boldsymbol{e})$为词项$\boldsymbol{e}$的强度，$f(\boldsymbol{e},p)$为在$p$时刻词项$\boldsymbol{e}$出现的频率，通常时间单位为“日”。代表性词项的爆发性是指词项在时间轴上的分布呈现局部性，即词项只在某些时间段内被大量引用。具有爆发性的词项在时间轴上的分布往往和热点事件中话题出现的规律相符。本文采用“熵”表示词项的爆发性，其公式为

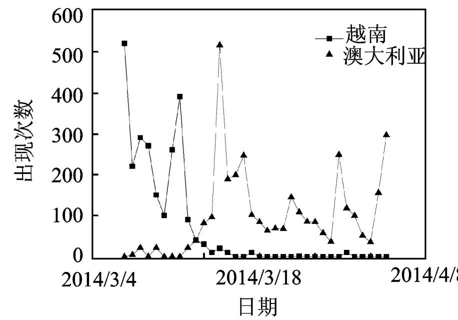

图2 在“马航”事件中词项“越南”和“澳大利亚”的分布情况

Fig. 2 Term distribution of "Vietnam" and "Australia" during the disappearance of flight MH370

$$\mathrm{Entropy}(\boldsymbol{e})=\sum_{p=wt_1}^{wt_n}\mathrm{Prob}(\boldsymbol{e},p)\log\left(\mathrm{Prob}(\boldsymbol{e},p)\right) \tag{3}$$

$$\mathrm{Prob}(\boldsymbol{e},p)=\frac{|\{w_i \mid \boldsymbol{e}\in E(\boldsymbol{w}_i)\ \wedge\ p=wt_i,\boldsymbol{w}_i\in \boldsymbol{W},i=1,\cdots,n\}|}{|\{\boldsymbol{w}_j \mid \boldsymbol{e}\in E(\boldsymbol{w}_j),\boldsymbol{w}_j\in \boldsymbol{W},j=1,\cdots,n\}|} \tag{4}$$

式中：$\mathrm{Entropy}(\boldsymbol{e})$为词项$\boldsymbol{e}$在时间轴上的熵，$\mathrm{Prob}(\boldsymbol{e},p)$为词项$\boldsymbol{e}$在$p$时刻出现的概率。综合上述两种特征，采用线性组合方法得到代表性词项的决策函数，其公式为

$$D(\boldsymbol{e})=\lambda\cdot\mathrm{Freq}(\boldsymbol{e})+(1-\lambda)\cdot\mathrm{Entropy}(\boldsymbol{e}) \tag{5}$$

给定阈值ε，如果词项$\boldsymbol{e}$的决策函数值$D(\boldsymbol{e})>\varepsilon$，则$\boldsymbol{e}$为代表性词项，否则不是。采用此方法可以得到整个微博集的代表性词项。由于无法判定词项强度和熵的权重大小，因此在本文中，λ设为0.5。

1.2 摘要评价指标

根据微博消息文本的特点，可以将微博时间线摘要指标分为内容覆盖性指标、时间分布性指标和传播影响力指标。

1.2.1 内容覆盖性指标

覆盖性指标是评价候选时间线在微博消息集中重要信息的覆盖程度，是评价时间线摘要质量好坏的一个重要标准。覆盖程度越高表明时间线摘要能够包含更多更全面的事件发展信息，则它与微博消息序列的中心就越接近。因此候选时间线摘要的覆盖性是通过计算候选时间线与微博消息集中心之间的距离实现的。本文的覆盖性指标根据文本间的相似性度量计算。采用矩阵构建微博消息文档和词项间关系，通过计算文本向量间的余弦相似度实现文本相似性度量。文档-词项矩阵定义为

$$\begin{array}{cc} & \begin{array}{ccc}\boldsymbol{e}_1 & \cdots & \boldsymbol{e}_m\end{array}\\ \begin{array}{c}\boldsymbol{w}_1\\ \vdots\\ \boldsymbol{w}_n\end{array} & \begin{bmatrix}\omega_{11} & \cdots & \omega_{1m}\\ \vdots & \ddots & \vdots\\ \omega_{n1} & \cdots & \omega_{nm}\end{bmatrix}\end{array} \tag{6}$$

式中：$\boldsymbol{w}_i$对应的行向量为$\boldsymbol{w}_i=[\omega_{i1},\cdots,\omega_{im}]$，$m=|R|$，$\omega_{ij}$为在微博$\boldsymbol{w}_i$中词项$\boldsymbol{e}_j$的权重，$\boldsymbol{w}_i\in\boldsymbol{W}$，$\boldsymbol{e}_j\in\mathbf{R}$。$\omega_{ij}$权重计算基于词频-逆文档频率的方法（Term frequency-inverse document frequency，TF-IDF）。微博$\boldsymbol{w}_i$和微博$\boldsymbol{w}_j$间相似性度量函数记为$\mathrm{Sim}(\boldsymbol{w}_i,\boldsymbol{w}_j)$。微博时间线摘要$\boldsymbol{S}$的内容覆盖性指标$\mathrm{Cov}(\boldsymbol{S})$定义为

$$\mathrm{Cov}(\boldsymbol{S})=\sum_{s_i\in \boldsymbol{S}}\frac{\mathrm{Sim}(\boldsymbol{s}_i,\boldsymbol{O})}{|s_{c_i}|}\tag{7}$$

式中:$\boldsymbol{O}$为微博消息集$\boldsymbol{W}$的中心,其向量为$\boldsymbol{O}=[\omega_{o1},\cdots,\omega_{o|R|}]$。$\omega_{oi}$为$\boldsymbol{W}$中心$\boldsymbol{O}$中第$i$个词项的权重,计算公式定义为

$$\omega_{oi}=\frac{1}{|\boldsymbol{W}|}\sum_{k=1}^{|\boldsymbol{W}|}\omega_{ki}\tag{8}$$

1.2.2 时间分布性指标

在传统摘要研究中,主要依靠内容覆盖性指标较多,考虑时间分布特征较少。但是时间线摘要反映的是事件在时间维度上的变化过程,因而时间线摘要需要,考虑微博的时间分布特征。时间分布性指标是指话题摘要在时间轴上的分布情况。在热点事件的发展演化中常常包含多个子事件。它们反映了事件发展的进程,并且具有明显的阶段性和时效性。如果生成的摘要在时间轴上分布过于集中,则无法全面完整地反映事件的发展过程。微博时间线摘要$\boldsymbol{S}$的时间分布性指标TD($\boldsymbol{S}$)定义为

$$\mathrm{TD}(\boldsymbol{S})=-\sum_{i=1}^{|\boldsymbol{S}|+1}(g_i-\bar{G}(\boldsymbol{S}))^2\tag{9}$$

式中:相邻时间间隔为$G=\{g_1,\cdots,g_{b+1}\}$,热点事件的时间跨度为wt_n-wt_1,相邻微博的时间间隔$g_1=\frac{st_1-wt_1}{wt_n-wt_1}$,$g_{b+1}=\frac{wt_n-st_b}{wt_n-wt_1}$,$g_i=\frac{st_i-st_{i-1}}{wt_n-wt_1}$,$i\in[2,b]$。$\bar{G}$表示微博序列理想的时间间隔,其公式为

$$\bar{G}(\boldsymbol{S})=\frac{wt_n-wt_1}{|\boldsymbol{S}|+1}\tag{10}$$

1.2.3 传播影响力指标

微博与传统媒体最大的差别在于"社交"。在热点事件的微博讨论中,大量的微博消息由用户间的互动产生。通常在意见领袖的推波助澜下信息传播地更为广泛且更快速。因而意见领袖发布的信息比普通用户的发言更重要,被作为摘要的可能性更高。由此可见,一次微博事件中微博消息的影响力并不相同。本文采用微博转发量作为度量微博消息影响力的指标。如果一条微博消息被大量转发,则认为该消息中包含被众多用户认可的重要信息。那么,相比于那些转发少的消息,此条微博消息将更有可能被选入微博时间线摘要中。本文候选时间线摘要$\boldsymbol{S}$的传播影响力指标Inf($\boldsymbol{S}$)定义为

$$\mathrm{Inf}(\boldsymbol{S})=\sum_{i=1}^{|\boldsymbol{S}|}\frac{\boldsymbol{S}d_i}{\max\limits_{1\le j\le n}wd_j}\tag{11}$$

1.2.4 综合指标

为了从上述3个方面综合评价时间线,本文采用线性组合的方法定义了时间线摘要的综合指标$F(\boldsymbol{S})$。本文的时间线摘要问题就是从微博消息序列中选取使得综合指标最大化的消息子序列。本文优化综合指标,最大化为

$$F(\boldsymbol{S})=\alpha\mathrm{Cov}(\boldsymbol{S})+\beta\mathrm{TD}(\boldsymbol{S})+\gamma\mathrm{In}f(\boldsymbol{S})\tag{12}$$

约束条件为

$$|\boldsymbol{S}|=b,0\le\mathrm{Red}(\boldsymbol{S})<\theta\tag{13}$$

计算时间线$\boldsymbol{S}$重复度的函数为

$$\mathrm{Red}(\boldsymbol{S})=\sum_{i=1}^{|\boldsymbol{S}|-1}\sum_{j=i+1}^{|\boldsymbol{S}|}\mathrm{Sim}(\boldsymbol{s}_i,\boldsymbol{s}_j)\tag{14}$$

式中:参数$\alpha+\beta+\gamma=1$,θ为重复度阈值。参数α,β,γ用于平衡覆盖性、时间分布性和影响力这3个指标的权重。本文采用文献[14]的网格搜索算法确定参数α,β,γ参数值。在综合指标函数中,如果将重复度阈值θ设为0,则要求任意两条微博之间都不存在任何相同的代表性词项,这是一个非常苛刻的限制条件。从大量实验研究中发现,即使是话题信息不同的两条微博,仍然存在少量相同的代表性词项。所

以本文设立的重复度阈值 $\theta \neq 0$，但是接近于 0。

1.3　算法思想与描述

微博时间线摘要的求解是从微博消息序列 $\boldsymbol{W}$ 中选择长度为 b 的最优子序列的过程，这是一个典型的组合优化问题。由于微博消息序列 $\boldsymbol{W}$ 规模庞大，采用枚举所有组合的方法开销巨大，时间复杂度过高。因此本文采用滑动窗口的方法求解微博时间线摘要，该方法避免了选择全局最优带来的高时间复杂度，保证了时间线摘要的可行性。设滑动窗口为 $\boldsymbol{S}_w = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_{b'}\}$，$\boldsymbol{s}_i \in \boldsymbol{W}$，时间线摘要 $\boldsymbol{S} = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_b\}$，$\boldsymbol{s}_j \in \boldsymbol{W}$。MTSW 算法基本思想：采用基于滑动窗口的方法遍历时间轴上的微博消息序列 $\boldsymbol{W}$。首先依次将微博消息加入滑动窗口 $\boldsymbol{S}_w$ 中，当滑动窗口中的微博条数为 b' 时，对滑动窗口中的微博消息枚举出所有组合，即 $C_{b'}^{b}$ 个候选时间线。然后按照重复度阈值和综合指标的评分对所有组合进行过滤和排序，保留最优组合中的微博消息，丢弃其余消息。最后添加新消息加入滑动窗口，重复上述步骤直到遍历完整个序列。如图 3 所示，微博消息序列 $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \boldsymbol{w}_4, \boldsymbol{w}_5\}$，滑动窗口 $\boldsymbol{S}_w = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\}$，生成长度为 2 的时间线摘要 $\boldsymbol{S}$。首先添加消息进入 $\boldsymbol{S}_w = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3\}$。然后组合出候选时间线 $\boldsymbol{S}_1 = \boldsymbol{w}_1\boldsymbol{w}_2$，$\boldsymbol{S}_2 = \boldsymbol{w}_2\boldsymbol{w}_3$ 和 $\boldsymbol{S}_3 = \boldsymbol{w}_1\boldsymbol{w}_3$。计算所有候选时间线的重复度和综合指标评分，选择满足重复度阈值且得分最高的 $\boldsymbol{S}_3$，则 $\boldsymbol{w}_2$ 消息被丢弃。最后，新的消息 $\boldsymbol{w}_4$ 被添加入 $\boldsymbol{S}_w$，继续重复上述步骤。MTSW 算法描述如下：

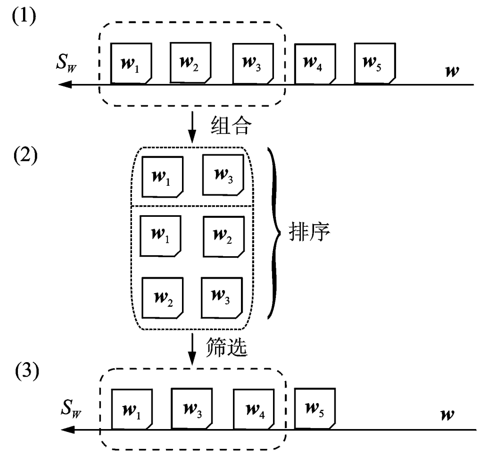

图 3　算法 MTSW 的执行过程

Fig. 3　Execution process of algorithm MTSW

输入：微博消息序列 $\boldsymbol{W} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n\}$，时间线摘要长度 b。

输出：时间线摘要 $\boldsymbol{S}$。

(1) $\boldsymbol{S} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_b\}$;　　//选取消息序列 $\boldsymbol{W}$ 的前 b 条微博作为初始时间线摘要。

(2) $\boldsymbol{W} = \boldsymbol{W} - \boldsymbol{S}$;

(3) $\boldsymbol{S}_w = \varnothing$; $\boldsymbol{C} = \varnothing$;

(4) For all $\in \varnothing \odot$ ∶ $\prod \partial \boldsymbol{w} \in \boldsymbol{W} \in \varnothing \odot$ // $\prod \partial$ do{

(5) $\boldsymbol{S}_w = \boldsymbol{S} \cup \{\boldsymbol{w}\}$;

(6) 在 $\boldsymbol{S}_w$ 中计算所有 b 条微博的组合并插入 $\boldsymbol{C}$ 中；

(7) For all $\boldsymbol{S}' \in \boldsymbol{C}$ do

(8) If Red$(\boldsymbol{S}') \geq \theta$ then $\boldsymbol{C} = \boldsymbol{C} - \boldsymbol{S}'$;　//剔除重复度不满足阈值的微博组合。

(9) $\boldsymbol{S} = \mathrm{argmax}\{F(\boldsymbol{S}') : S' \in \boldsymbol{C}\}$;　//将使 $F(\boldsymbol{S})$ 最大的 b 条微博组合作为时间线摘要 $\boldsymbol{S}$。

(10) return

上述方法的时间复杂度为 $\boldsymbol{O}(n \times b)$，$n$ 为微博序列 $\boldsymbol{W}$ 中的微博数，b 为时间线摘要长度。

2　实验与结果分析

2.1　基准测试算法

为了验证微博时间线摘要算法 MTSW 的可行性和有效性，本文设计了 4 种不同类型的基准测试算法。分别是文本聚类摘要（Text clustering summarization, TCS）算法[15]、最大边缘相关（Maximal marginal relevance, MMR）算法[16,17]、有监督的机器学习摘要算法（Probabilistic support vector machine, PSVM）[18] 和基于微博文本流的时间线摘要方法（Stream-based subtopic detection and ordering, SSDO）[4]。

TCS算法采用带噪声的基于密度的聚类算法(Density-based spatial clustering of applications with noise, DBSCAN)聚类实现的抽取式摘要算法。该算法在不考虑文本间链接关系的情况下只关注消息文本的内容,按照文本相似度进行聚类,从规模最大的 k 个聚类簇中选择微博消息作为微博摘要,并且满足重复度阈值的约束要求。其中文本相似度计算采用 TF-IDF 向量和余弦相似性方法进行度量。MMR 算法将文本摘要定义成一个排序问题。采用贪心策略对文本进行重排序,然后选择前 k 条文本作为最终的摘要。PSVM 算法将文本摘要定义成一个分类问题。首先对消息文本的摘要特性进行特征选取,采用支持向量机预测是否可以作为摘要,然后对符合摘要特性的句子进行排序,最后选择前 k 条语句作为最终的摘要。SSDO 算法通过识别微博文本流中的子话题生成时间线摘要。子话题的出现往往伴随文本内容和文本量的双重变化。SSDO 算法通过识别子话题引发的“激增”区域可以准确划定时间线上摘要的抽取区域。为了保证最终摘要结果中不存在重复的内容和信息,本文所有算法采用相同的重复度阈值作为约束条件,依序选择不重复的前 k 条语句作为摘要结果。

2.2　时间线摘要评价方法

摘要的评估一直是一个比较困难的问题。2005 年美国国家标准与技术研究院(National institute of standards and technology, NIST)组织的文档理解会议(Document understanding conferences,DUC)评测中确定了新闻摘要的评价指标和测试数据集。但是针对微博时间线摘要问题学术界至今仍然没有权威的测试数据集。在评价指标方面,本文采用信息检索中的准确率、召回率和 F_1 值 3 种指标[19]评价算法生成时间线的正确情况和全面程度。需要说明的是,本文使用词项和语句不同的粒度进行指标评价。词项粒度的指标主要反映时间线摘要中话题信息的准确和全面程度,而语句粒度的指标旨在评估时间线摘要中子事件发展演化的准确和完整程度。在词项粒度评价方面,本文将算法的摘要表示成代表性词项集并且合并。基于此词项集合的计算准确率、召回率和 F_1 值。此方法可以从整体上评估算法生成摘要的子话题质量。在语句粒度评价方面,如果直接比较人工摘要语句和算法摘要语句是否相同,可能过于严苛。本文使用摘要语句包含的代表性词项集代表语句本身。如果两个摘要语句的代表性词项集相似度超过 70%,则视为相同。

2.3　测试数据集和实验结果分析

为了研究微博热点事件的时间线摘要问题,本文采用腾讯微博 API 收集了“马航 MH370 失联”事件(MH370)的微博消息。根据整个数据集的时间分布情况,采用随机抽样的方法将数据集分为训练集和测试集。训练集主要用于求解综合指标中的参数 α,β 和 γ,而测试集用于验证算法的正确性和有效性。表 1 介绍了该事件微博数据集的基本情况。

表 1　MH370 事件数据集

Tab. 1　Dataset description of MH370 event

数据集	时间跨度	微博数/条	大小/KB
MH370 训练集	2014/03/15～2014/04/18	10 850	4 083
MH370 测试集	2014/03/15～2014/04/18	3 129	1 172

本文采用人工抽取的方式从 MH370 测试集中识别出人工时间线摘要(简称人工摘要)。但是文献[20]发现,在没有任何指导信息的情况下,不同人员对于同样的文本集做出的摘要结果平均只有 8%的内容相同。所以为了尽可能降低人的主观因素对人工摘要结果的影响,本文将 MH370 事件的新闻时间线[21]作为人工摘要的指导信息,然后由 5 名不同研究人员分别对 MH370 测试集进行人工摘要,得到 5 份结果。如果在特定时间段内某条微博消息出现在 4 份以上的摘要结果,则此微博应该被选入最终的人工时间线摘要。按照该方法,最终的人工时间线包含 15 条微博。参与微博时间线摘要实验的各种

算法参数设置为：TCS算法的DBSCAN半径长度和最小密度阈值分别设为0.7和2；MMR算法的相关性和重复度的权重均设为0.5；PSVM算法选取话题代表性词项作为分类特征。SSDO算法中的步长时间单位为“天”。MTSW算法参数$\alpha=0.4$，$\beta=0.4$，$\gamma=0.2$，$b=15$。为了保证5种算法比较的公平性，本文设定4种算法的摘要结果重复度指标都必须满足相同的阈值，即所有算法在相同的重复度约束下进行比较。文中该阈值设置为0.05。本文实验在CPU为Intel Core i5 3.2GHz、内存大小为4GB的PC机上进行。为了验证MTSW算法在微博时间线摘要上的有效性和可行性。在准确率、召回率和F_1值指标方面采用词项和语句两种不同粒度对5种算法进行性能比较，实验结果如表2所示。表3为所有算法的运行时间。从实验结果可以发现，TCS算法的摘要效果一般，而且运行时间较长。此方法采用聚类算法将内容相似的消息聚集成簇，然后从簇中选取摘要。该方法仅依靠文本内容信息生成摘要，仅选取文本强度大的消息作为摘要，因而可能忽略“马航”事件后期子事件（事件发展平缓）的摘要，无法完整表现事件发展的全部过程。此外，聚集算法中的相似性度量非常耗时，导致该算法无法在短时间内完成时间线摘要任务。除了MTSW算法以外，MMR算法优于其他方法。主要原因是MMR算法根据文本内容和重复度指标进行文本的排序，排在前列的文本在内容上具有较好的代表性词项，并且文本间的重复度较低。这样可以避免摘要集中出现在事件爆发阶段。因此，摘要效果要好于TCS算法。此外MMR算法运行时间要比TCS算法缩短许多，但是仍然比MTSW算法长许多。PSVM算法在词项和语句两种粒度评价上都表现最差。原因是PSVM在摘要排序中只考虑了语句的位置和长度特征，没有利用文本内容信息生成摘要。所以最终的时间线摘要效果不理想。SSDO算法的处理方式接近于MTSW算法，都采用面向文本流的处理方法生成摘要。SSDO算法虽然考虑了话题内容和话题强度在时间维度上的变化，但是在语句粒度指标上表现较差。原因是SSDO发现的摘要区域只有当子话题在消息总量和内容变化上非常明显时才被接纳。这导致一些不够明显的摘要被忽略。如热点事件发展到后期往往出现话题变化不明显、消息量变化趋缓。为了保证事件发展状态的完整性，依然需要识别相应的摘要。在所有指标上，MTSW算法的时间线摘要效果明显优于其余算法。MTSW算法综合考虑文本内容、时间分布和社交影响力3种特征，识别具有代表性的微博作为摘要。由此可见，综合考虑微博多种特征的算法要优于采用单一文本特征的算法。本文的实验效果证明了该方法的有效性。

表2 MH370测试集上5种时间线摘要算法的准确率、召回率和F_1值比较

Tab. 2 Performances of five algorithms for timeline on MH370 test dataset %

算法	词项粒度			语句粒度		
	准确率	召回率	F_1	准确率	召回率	F_1
TCS	57	58	57	40	40	40
MMR	59	67	63	63	63	63
PSVM	51	55	53	20	20	20
SSDO	62	55	58	33	33	33
MTSW	75	78	77	67	67	67

表3 MH370测试集上时间线摘要算法运行时间比较

Tab. 3 Runtimes of five algorithms for timeline on MH370 test dataset

算法	运行时间
TCS	≥4 h
MMR	166.03 s
PSVM	0.59 s
SSDO	0.14 s
MTSW	33.42 s

为了展现MTSW算法生成摘要的具体效果，表4列出了此算法生成的MH370事件的时间线。由于篇幅原因，本文只用简短的语句替代微博消息原文。不难发现，在事件发展的时间轴上MTSW算法生成的摘要和人工摘要有很多的相同子话题。

表 4　MTSW 算法生成的“MH370 事件”时间线摘要

Tab. 4　MH370 timeline summary generated by algorithm MTSW

时间	人工时间线摘要	MTSW 算法的时间线摘要
2014.3.16	马航称正在刑事调查是否劫机。	马航称正在刑事调查是否劫机。 马军方人士披露 MH370 失联前完整航线。
2014.3.17	4 大疑问指向机长。	4 大疑问指向机长。 航空专家徐勇凌解读几大疑点。
2014.3.18	中方排除中国乘客涉恐嫌疑。 马航事件谣言满天飞。	中方排除中国乘客涉恐嫌疑。
2014.3.20	澳大利亚发现疑似飞机残骸。 疑似残骸不在常规航道。	澳大利亚发现疑似飞机残骸。 疑似残骸不在常规航道。 两大碎片可能来自失联航班。
2014.3.21	无	正在确认疑似残骸。 澳大利亚搜寻疑似失联客机物件海域无突破性发现。
2014.3.22	中国“高分一号”卫星发现疑似物。	无
2014.3.24	马方确认 MH370 在南印度洋坠毁。	马方确认 MH370 在南印度洋坠毁。
2014.3.25	家属前往马驻华使馆抗议。 习近平指示即派我国政府特使赴马处理马航客机失联事件。 “雪龙号”前往搜索地点。	家属前往马驻华使馆抗议。 习近平指示即派我国政府特使赴马处理马航客机失联事件。
2014.3.31	马航 MH370 最后通话内容公布。	无
2014.4.6	疑似黑匣子海域水深 4 000～4 500 m。	无
2014.4.8	搜寻一个月,决不放弃。	搜寻一个月,决不放弃。
2014.4.10	“海盾号”检测到黑匣子脉冲信号。	“海盾号”检测到黑匣子脉冲信号。

3　结束语

本文以微博消息为对象,研究微博热点事件的时间线摘要问题。按照处理流程的不同,微博时间线摘要处理框架分为 4 个阶段,分别是采集微博数据、文本预处理、时间线摘要和摘要结果展示。由于微博消息的特点,无法仅依靠文本内容信息度量微博消息的重要程度,因而本文综合利用内容覆盖性、时间分布性和传播影响力 3 种指标,构建出评价时间线摘要的综合评价指标,提出了基于滑动窗口的微博时间线摘要算法 MTSW。该算法采用滑动窗口的方法求解微博时间线摘要,避免了选择全局最优带来的高时间复杂度,保证了时间线摘要的可行性。利用真实微博数据进行实验,结果表明 MTSW 算法可以有效地抽取出代表性的微博消息,反映热点事件发展演化的过程。

参考文献:

[1] Hu M, Sun A, Lim E P. Comments-oriented blog summarization by sentence extraction [C]//16th ACM Conference on Information and Knowledge Management(CIKM/07). Lisbon, Portugal: ACM, 2007: 901-904.

[2] Gao W, Li P, Darwish K. Joint topic modeling for event summarization across news and social media streams [C]//21st ACM International Conference on Information and Knowledge Management (CIKM′12). Maui, HI, USA: ACM, 2012: 1173-1182.

[3] Ma Z Y, Sun A, Yuan Q, et al. Topic-driven reader comments summarization [C]// 21st ACM International Conference on Information and Knowledge Management (CIKM′12). Maui, HI, USA: ACM, 2012:265-274.

[4] Gao D H, Li W J, Zhang R X. Sequential summarization: A new application for timely updated twitter trending topics [C]// Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL′13). Sofia, Bulgaria: ACL, 2013:567-571.

[5] Christensen J, Soderland S, Bansal G, et al. Hierarchical summarization: Scaling up multi-document summarization [C]//

Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL'14). Baltimore, Maryland, USA: ACL, 2014:902-912.

[6] Zhao X W, Jiang J, He J, et al. Topical keyphrase extraction from twitter [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL'11). Portland, Oregon: ACL, 2011:379-388.

[7] Weng J W, Yang C L, Chen B N, et al. IMASS: An intelligent microblog analysis and summarization system [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL'11). Portland, Oregon: ACL, 2011: 133-138.

[8] Ren Z H, Liang S S, Meij E, et al. Personalized time-aware tweets summarization [C]// Proceeding of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13). Dublin, Ireland: ACM, 2013:513-522.

[9] Shou L D, Wang Z H, Chen K, et al. Sumblr: Continuous summarization of evolving tweet streams [C]// Proceeding of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13). Dublin, Ireland: ACM, 2013:533-542.

[10] Yan R, Kong L, Huang C R, et al. Timeline generation through evolutionary trans-temporal summarization [C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing(EMNLP'11). Edinburgh, UK: ACL, 2011:433-443.

[11] Li J W, Li S J. Evolutionary hierarchical dirichlet process for timeline summarization [C]// Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics(ACL'13). Bulgaria: ACL, 2013:556-560.

[12] Ng J P, Chen Y, Kan M Y, et al. Exploiting timelines to enhance multi-document summarization [C]// Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics(ACL'14). Baltimore, Maryland, USA: ACL, 2014: 923-933.

[13] Tran G B. Structured summarization for news events [C]// Proceedings of the 22th International Conference on World Wide Web (WWW'13). Republic and Canton of Geneva, Switzerland: ACM, 2013:343-348.

[14] Christensen J M, Soderland S, et al. Towards coherent multi-document summarization [C]// Proceedings of NAACL-HLT. Atlanta, Georgia: ACL, 2013:1163-1173.

[15] Nenkova A, McKeown K. Automatic summarization [J]. Foundations and Trends in Information Retrieval, 2011, 5(2):103-233.

[16] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries [C]// Proceeding of the 21th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98). Melbourne, Australia: ACM, 1998:335-336.

[17] Murray G, Renals S, Carletta J. Extractive summarization of meeting recordings [C]// Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal: ISCA, 2005:593-596.

[18] Wong K F, Wu M L, Li W J. Extractive summarization using supervised and semi-supervised learning [C]// Proceeding of the 22nd International Conference on Computational Linguistics (Coling 2008). Stroudsburg, PA, USA: ACL, 2008:985-992.

[19] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. New York: Cambridge University Press. 2008:142-143.

[20] Rath G J, Resnick A, Savage T R. The formation of abstracts by the selection of sentences. Part I. Sentence selection by men and machines [J]. American Documentation, 1961, 12(2):139-141.

[21] 百度百科. MH370 [EB/OL]. http://baike.baidu.com/view/12368712.htm,2014-10-30.

作者简介：

徐伟(1990-),男,硕士研究生,研究方向:数据挖掘技术与应用,E-mail:xwnjnu@163.com。

赵斌(1978-),通信作者,男,博士,讲师,研究方向:Web数据挖掘。

吉根林(1964-),男,教授,博士生导师,研究方向:数据挖掘技术及应用。