中国科学: 信息科学 2015 年 第 45 卷 第 12 期: 1547–1557
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS
社会媒体处理专题 · 论文

基于情感时间序列的微博热点主题检测

李斌阳[①], 韩旭[②*], 彭宝霖[③], 李菁[③], 王腾蛟[④], 黄锦辉[③]

① 国际关系学院信息科技学院, 北京 100091
② 首都师范大学信息工程学院, 北京 100048
③ 香港中文大学系统工程与工程管理系, 香港 999077
④ 北京大学信息工程学院, 北京 100871
* 通信作者. E-mail: hanxu@cnu.edu.cn

收稿日期: 2015–10–01; 接受日期: 2015–10–20; 网络出版日期: 2015–12–02
国家自然科学基金 (批准号: 61502115, 61370165, 61572043) 和中央高校基本科研项目 (批准号: 3262014T75, 3262015T20) 资助

摘要 随着互联网的蓬勃发展, 微博在信息传播过程中扮演着非常重要的角色, 正逐渐演变成一种新型线上交流新闻源. 人们已经习惯于通过微博平台来了解他们身边的朋友或家人在做什么, 关心这个世界正在发生什么. 然而, 由于微博平台蕴含着海量信息, 很难以人工的方式在微博上快速检测当前实时发生的重大新闻或突发事件. 因此, 面向微博的热点主题检测成为当下的一个研究热点. 然而, 现有研究主要侧重于主题识别而忽略了用户对于实时性的要求, 少数针对实时热点主题发现的方法主要基于关键词的统计分析, 实时性和准确率都有待提高. 根据我们的观察发现, 微博平台汇集了成千上万的观点与意见, 包括对社会事件的讨论、对产品的评价等, 这些观点使得微博成为一个非常有价值的观点意见数据源. 通过分析观点与情感的实时变化, 我们可以更好地了解相关主题的变化趋势, 从而辅助用户判定其是否是流行的热点主题. 本文结合微博的情感时序变化提出了一种实时的非参数化的热点主题检测方法. 该方法通过对微博情感极性分析及其强度变化来计算情感时序分布, 并利用上述特征构建一个复合模型以识别、检测微博热点主题. 实验分别在 Twitter 和新浪微博等真实数据集上进行, 结果表明我们提出的方法能够在保证检测准确率的前提下更快地识别热点话题.

关键词 情感分析 热点主题 情感时间序列 实时检测 微博

1 引言

随着互联网的蓬勃发展, 微博在信息传播中扮演着非常重要的角色, 它可以让一条很小的信息在很短的时间内传播到世界的各个角落. 可以说, 微博已经逐渐演变成一种等同于新闻源的新型线上交流信息平台 [1]. 微博的出现与流行给我们的日常生活带来了巨大影响, 它像一面镜子真实的反映身边生活与世界的变化. 人们已经习惯于通过微博平台来了解他们身边的朋友或家人在做什么, 关心这个世界正在发生什么. 例如: Justin Bieber 有新的女朋友了么? 台风海燕给菲律宾带来了怎样的灾难性损失? 葡萄牙跟瑞典, 哪个赢得了世界杯的举办权? 等等. Sakaki 等 [2] 通过案例研究发现, 相比于其

引用格式: 李斌阳, 韩旭, 彭宝霖, 等. 基于情感时间序列的微博热点主题检测. 中国科学: 信息科学, 2015, 45: 1547–1557, doi: 10.1360/N112015-00249

他传统新闻媒介, 信息在微博上的产生数量更大、传播速度更快, 因此微博更适合充当热点主题、突发事件的检测器.

以 Twitter 为例, 作为全世界最流行的微博网站之一, Twitter 拥有超过 5 亿的注册用户, 每天发表的推文 (tweets) 超过 3.4 亿条, 其信息内容几乎涵盖了世界范围内每天发生的各类事件. 然而, 面对如此大的信息量, 一名普通用户很难以人工的方式在 Twitter 上快速发现当前实时发生的重大新闻或突发事件. 因此, 面向微博的热点主题检测成为了当下的一个研究热点.

不止如此, Twitter 同时还扮演了一个观点表达和情感交流平台的角色. 在 Twitter 上, 用户不仅限于提及某一事件, 更习惯于在推文中表达他们对某一事件的真实情感, 诉说他们真正的思考所在. 事实上, Twitter 平台汇集了成千上万的观点与意见, 包括对社会事件的讨论、赞扬, 或者对产品的批评、评价等等 [3]. 类似地, 在其他流行的微博平台上, 如新浪微博、腾讯微博等, 也包含了大量的针对某些热点主题的观点, 这些观点使得微博成为了一个非常有价值的观点意见数据源. 通过分析观点与情感的实时变化, 我们可以更好地了解相关主题的流行趋势, 从而辅助用户判定其是否 (可能) 是热点主题. 然而, 现有相关研究主要侧重于主题识别而忽略了用户对于实时性的要求, 少数针对实时热点主题发现的方法主要基于关键词的某种统计分析, 忽略了情感、观点等内涵更丰富的信息的作用, 例如针对某一社会事件微博用户观点强度的变化、正负极性比例的变化等. 因此, 现有方法对于热度实时变化并不敏感, 很难快速、有效地识别微博平台的热点主题.

针对上述问题, 我们结合微博上的情感时间序列提出了一种实时的非参数化的热点主题检测方法 (sentiment-topic detection, STD). 该方法通过对微博情感极性分析及其强度变化来计算情感时序分布, 并利用上述特征构建一个复合模型以识别、检测微博热点主题. 为了测试 STD 方法的有效性, 我们分别在 Twitter 和新浪微博等真实数据集中进行了相关实验, 将 STD 检测出的热门主题与 Twitter、新浪微博所展示的热点话题进行对比. 实验结果表明 STD 能够在保证检测效果的前提下更快地识别出热点主题. 同时, 由于 STD 结合了情感的时序变化特征, 因而对于与主题相关的观点非常敏感, 能够快速有效地检测出具有强情感性的话题. 通过进一步地分析发现, 其中部分话题在当时并没有引起被 Twitter、新浪等公司的重视, 但在未来的一段时间引起了社会的强烈反响.

综上, 本文的主要贡献包括以下几个方面:

(1) 结合情感分析等多种特征提出了一种复合模型, 包括推文数量、情感极性、情感强度以及参与的用户量等特征, 有效地提高了热点主题检测的准确率;

(2) 融合情感值的时序变化曲线提出了一种实时热点主题检测模型, 其有效地提高了微博平台热点主题的检测效率;

(3) 分别在 Twitter 和新浪微博等真实数据集上进行实验, 实验结果显示我们所提出的方法与当前主流方法相比, 相关指标平均可提高 4 个百分点, 可更快地识别发现热点主题, 平均提高 25 分钟.

2 相关工作

在主题发现这一研究领域中, 当前的主流方法大多基于 Latent Dirichlet Allocation (LDA) 主题模型. Diao 等 [4] 在 LDA 模型基础上对其加以处理和改进, 使其适应微博的多样性和噪音. Gao 等 [5] 提出的基于 HDP 的渐进 Gibbs 采样算法, 通过获得聚类的标签, 找出主题间的关系. 虽然这些研究工作都声称与标准主题模型相比有更好的结果, 但是这些工作不能处理流数据, 因此也不能应用于实时的热点主题发现. 而在实时主题发现方面, 当前流行的有 3 种方法 [6]. 其中最流行的方法是在相应的主题下, 分析主题活跃性的偏离程度. 此外, Twitter 监测器 [7] 通过将关键词聚类来生成主题, 并

且结合社交媒体的权威度来计算每个主题的相关热度, 相当于通过对用户在主题中的交互进行分析, 在 Twitter 数据流上来发现实时主题. Becker 等 [8]、Cataldi 等 [9]、Gao 等 [10] 也提出过类似的方法.

Nikolov [6] 提出的基于时间序列分类的模型, 认为可以通过推文的时间线分析来确定主题的趋势. 他的研究工作很具吸引力且方法较先进. 但是, 他更多地考虑了相关主题下推文数量的变化而忽略了推文中情感的重要性. 而事实上, 以 Twitter 为代表的微博平台是情感表达和意见分享的平台, 情感分析和意见挖掘在 Twitter 分析中有着特殊的价值 [11,12]. 通过分析观点与情感的实时变化, 我们可以更好地了解相关主题的变化趋势, 从而辅助用户判定其是否是流行的热门主题.

在目前与情感分析的相关工作中, 情感分类得到了最为广泛的关注. 这种分类的目标是通过利用机器学习的方法, 将一篇带有作者主观观点的文本分类为正向或者负向. 传统的情感分类研究工作主要集中在以下两个方面.

第一个方面是如何改进机器学习的方法并且应用于文本情感分类. Pang 等 [13] 是把这些分类算法运用于影评情感分类的先驱. 在他们的工作中, 朴素 Bayes、最大熵模型、支持向量机都被应用于确定影评情感的极性 (肯定或者否定). 1999 年, Wiebe 等 [14] 就提出了利用朴素贝叶斯的方法去确定一个文本是主观的还是客观的. 受此启发, Diao 等 [4] 使用层次分类模型, 在情感极性分类之前, 将主观观点分类作为情感分析的第一步. 无容置疑, 在 “客观句子不包含主观观点” 的假设下, 主观观点分类这一思路是合理的. 当然, 在一些特殊情况下, 它并不总是正确的.

另一个重要的方面是特征选择. 在以往研究工作所使用的特征中, 共总结出 6 种最为普遍的特征类型 [15]. 词项及其频率虽然是最简单的一种特征, 但在 Pang 等 [13] 以朴素 Bayes 和支持向量机作为情感分类器中也行之有效. 另外, 词性这一特征也很重要, 在 Riloff 等 [16] 和 Wiebe 等 [17] 的情感标注研究工作中就证明了词性特征的有效性. 而情感词以及短语可以直接表达出肯定或否定的情感, 都被认为是明确有效的情感标识符. 在特殊情况下, 形容词、副词、动词和名词都可能被认为是情感词. 此外, 否定词通常能够改变情感极性, 因而也非常重要.

3 模型与方法

本节将介绍我们所提出的基于情感时序变化的模型是如何实时检测热点主题的. 首先, 我们利用情感分类器来计算给定微博的情感值, 该值代表了微博的情感倾向极性及其强度. 进而, 通过估计情感值的分布来确定情感的时序变化曲线. 最后, 结合上述情感时序特征建立了一个复合的实时热门话题检测模型.

3.1 问题定义

在大数据时代, 从海量信息中发现热门话题是一项非常有意义的任务. 特别是在社交媒体领域, 这一观点敏锐、情感动态变化的环境, 热门情感话题识别尤为重要.

在这个背景下, 我们不妨进一步假设, 越是有很多人愿意发表观点的话题, 其更倾向于成为一个流行话题; 越是发表情感表达强烈的话题, 其更有可能成为一个流行话题 (注意不包括刚刚发生的事件尚无人评论). 实际上, 在我们的日常生活中, 尽管有些事因为其正式性被传统新闻媒体大量提及, 例如某些政府文件的制定或政策的出台等, 但是仍然很难说这是一个热点的话题, 因为很少有人会有对这件事有强烈的情感. 相反地, 一些非官方新闻事件, 却往往比较有趣且富有争议, 其中的很大一部分通常会演变成热点事件或流行话题.

针对这一现象, 我们将问题定义在 Twitter 和新浪微博平台上, 从而可以获取更广范围的主题. 同时, 微博平台提供了足够的开放接口, 也有正式的数据流 API, 方便获取、处理相关数据.

3.2 数据获取与处理

同大部分工作类似, 我们也使用 Twitter 的数据流 API 来获取数据. Twitter 数据流 API 对于一般用户而言有访问限制, 所以几乎不可能达到绝对较高的采样率. Twitter 相关文档明确规定采样的比例不得超过整体的 1%. 我们利用数据流 API 收集了超过 20 天的数据, 每天返回的数据量为 10 GB. 总共获得了 9 亿条推文, 均匀的分为训练集和测试集. 与此同时, Twitter API 也提供了每个小时的流行话题标签, 我们在收集数据流的同时也收集此类数据, 总共获取了 1400 个热点或者流行话题. 每一个标签都有时间戳, 用以表明该主题在一天的哪些时间里是流行的.

需要指出的是本文的目标是检测一个话题是不是热点话题而非一般的话题识别. 传统的方式就是利用词袋模型来表示数据, 非常的简单而且有效, 但是它仅仅是通过定义一些关键词来定义话题, 其中大部分为动词或者名词, 忽略了他们之间的次序. 此外, 传统的话题模型, 像 LDA 等, 并不适合对数据流进行实时检测.

同时, 标签被认为是一个粗粒度的话题, 一个标签通常包含一个单词或一个词语, 例如 #london-riot, #twitterparty, #nowplaying. 标签通常是 Twitter 用户定义的, 在用户之间也比较流行. Wang 等 [18] 研究了 60 万条随机挑选的推文, 发现大约 14.6% 的推文都会使用至少一个标签. 此外, 实时话题监测可以认为是对标签的流行度、热度的计算, 这样避免了话题检测通常使用的聚类引发的噪音. 换言之, 我们不关心一个话题是怎样生成的, 我们更关心如何发现一个热点主题. 因此, 我们选择标签作为一个话题, 而不是使用词袋模型表示话题, 这就避免了将不同的关键词聚类到同一个话题的过程.

在数据获取过程中, 我们分别针对正、负样例进行处理. 针对正样例, 我们利用 Twitter API 返回的热点话题作为热点主题标签, 同时过滤掉那些存在时间过长或过短的热点主题, 例如超过 24 小时或者不到 4 小时的热点话题. 因为这些话题或者会引入一些噪音, 或者在时间上不能持续存在而成为噪音. 此外, 将热点主题存在时间限定在 24 小时, 可以有效避免重复出现的模式. 针对负样例, 我们随机从非热点主题中选取标签作为非流行话题, 处理过程基本与正样例相仿, 同样过滤掉存在时间过长或者过短的主题.

与上述过程类似, 我们同样的从新浪微博平台上收集了 3 个月的微博数据, 并按照相同处理手段对其进行标注处理.

需要指出的是, 为了加快标签抽取和分析速度, 我们采用了 MapReduce 的框架, 每一个标签都可以看作是 MapReduce 中的 Key, 合并推文的过程可以看作是 Reduce 的操作. 对于每一条推文, Mapper 产生一对标签和 ID, Reducer 将 Twitter 的 ID 合并到相关的标签列表中. 通过这种方式, 可以大大加快处理速度.

3.3 情感时序曲线

为了有效的利用情感时序曲线来检测热点主题, 首先需要计算每一条推文的情感值, 该值同时反映了情感极性和情感强度. 在本文中, 我们利用 SVM 分类器 [19], 并结合 unigram 特征来计算情感值. 根据 Pang 和 Lee [20] 的研究发现, unigram 特征简单但有效, 应用在 SVM 分类器上可取得同其他方法近似甚至更好的效果. 此外, SVM 分类器不仅仅可以判定一条推文的极性, 还给出了推文的极性强度. 推文特征向量距离 SVM 分类器超平面越远, 推文的极性程度就越高, 蕴含情感越强. 因此, 我们

使用 SVM 的得分值作为情感值. 为了训练分类器, 我们利用 Go 等 [21] 发布的 160 万条主观推文作为训练语料, 该语料根据其所包含的情感标记将推文划分为正或负极性.

如前所述, 在先前工作中, 很多人将重点放在了分类特征的选取上. 其中, 情感词常常被认为是最为重要的特征, 大量工作都是基于情感词典的方法, 如 MPQA [17]. 不幸的是, 微博用户倾向于使用非正式语言, 推文中所使用的大量情感词无法在上述词典中找到. 所以, 我们从微博训练语料中自动筛选了一部分特征词, 包括表情符号等, 来提高推文的情感分类效果.

此外, 词性 (part-of-speech) 被认为在情感分类方面非常有价值. 名词、动词、形容词和副词都是可能的情绪指标. Barboca 和 Feng [22] 总结了对于微博情感分析最有效的 4 个特性, 分别为正向词、负向词、动词、表情符号. 因此, 本文的方法所使用的特征也集中在这 4 类. 同时, 我们发现不同极性的词语词频特征也可以提高情感分类的精度, 例如快乐的、伟大的、美妙的等词会更频繁地出现在正面推文中, 因此我们也将其视为极性分类指示器. 最后, 作为特征选择的最后一步, 我们会将那些分别出现在两种极性的推文且词频近似的词去除掉, 以提高特征的有效性.

进一步, 为了方便比较情感细粒度时序变化, 我们定义了一个情感强度来指示推文的情感极性及其相应强度, 见定义 1.

定义 1 情感强度, 定义为 [−1,1] 的一个实数值, 负数表示负极性, 正数表示正极性. 其绝对值越大表明情绪强度越高.

但是, 需要指出的是, 情感得分是相对的而非绝对值. 换句话说, 很难准确的定义一条推文情感值的绝对强弱, 只有当其进行比较的时候才有意义. 例如, 不同的人对于同一条推文会有不同的评价标准, 例如, “I have done a good job”, 几乎所有人都会有正面的评价, 但是当 good 与 better 甚至 best 相比较时, 很明显后者会比前者高. 所以, 我们将所有的得分都变换到 [−1,1].

当得到每一条微博的情感值后, 就可以进一步计算情感分布与情感时序曲线. 考虑到情感值是一个相对值, 需要采用一些后处理的方式对其进行加工. 首先, 针对情感值的相对性, 我们利用一种在图像处理中非常流行的方法来对情感值进行平滑, 即对比直方图的方法 Histogram Equalization. 这种方法通常能够达到比较高的全局对比度, 尤其是当情感值比较接近的时候, 这种方法尤为奏效. 虽然这种方法一定程度上牺牲了局部的对比度, 但它能够有效地反映情感强度的分布情况, 增强了全局的对比度.

为了计算情感值时序曲线, 我们假设每条推文的情感值对应一个时间窗口, 并利用 Gauss 核的 Parzen 窗口估计每个时间窗口中的情感分布, 具体计算公式如下:

$$p_n(s) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{s-s_i}{n}\right), \tag{1}$$

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp^{-}\frac{x^2}{2}, \tag{2}$$

其中, n 代表推文数量, s 代表窗口中所对应推文的情感值, h 是时间窗口的宽度. 最后, 我们针对每一个时间窗口都计算其相应的一个情感强度期望值, 并将期望值的序列作为一个标签的情感时序序列.

3.4 热点主题的实时检测

如图 1(a) 所示, 原始的时间序列由于微博上内容的迅速变化, 导致了较多的震荡. 为了减少噪音, 覆盖高阶的信息, 在热点话题监测之前, 我们对时间序列进行正则化处理, 得到了如图 1(b) 所示的正则化的时间序列. 具体的正则化方法为, 假设第 n 个时刻的值为 $s[n]$, 方法包括如下几个步骤:

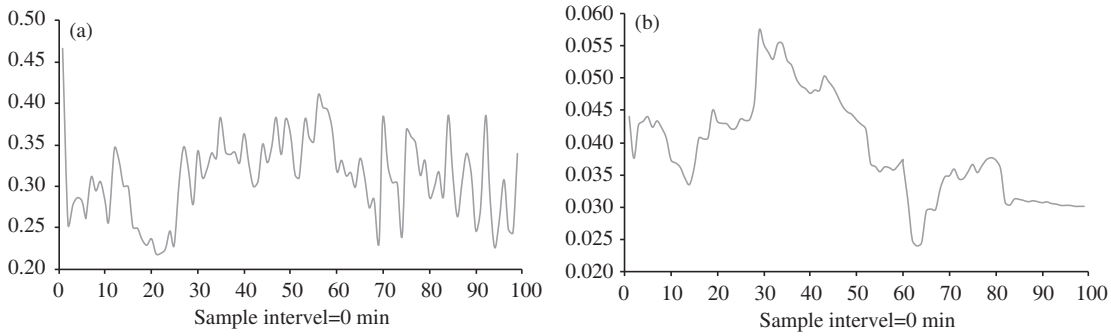


图 1　时间序列正则化前后比较图

Figure 1　Time series (a) before and (b) after normalization

(1) Baseline normalization, $s(n) = s(n)/\sum_i^N s(i)$;

(2) Spike normalization, $s(n) = (s(n) - s(n-1))^a$;

(3) Smoothing, $s(n) = \sum_{m=(n-N_{\text{smooth}}+1)^n} s(m)$;

(4) Logarithmic, $s(n) = \log s(n)$,

其中 N 是整体的序列长度.

然后, 我们采用基于监督方法, 并融合情感时序曲线这一特征来检测与之相应的热点主题.

在我们的模型中, 需要标注正负样例, 来标明监督信号. 我们利用 Twitter 系统或新浪所提供的热点主题报告中的热点标签来作为正样例. 此外, 其他的负样例, 也就是非热点标签, 来作为负样例. 进而, 我们采用如下公式来计算话题的热点度, 其中 R_+ 表示正样例, R_- 作为负样例:

$$R(s) = \frac{\sum_{c \in R_+} \exp(-r \cdot d(s,c))}{\sum_{c \in R_-} \exp(-r \cdot d(s,c))}, \tag{3}$$

r 是用来控制敏感度的参数, $d(s,c)$ 定义了两个时间节点 s 与当前时间节点的距离, 通常使用欧式距离公式. 最后, 通过式 (3) 所计算出来的 $R(s)$ 值来判断定时间序列 s 是否属于热点话题.

4　实验

实验阶段, 我们将在前面提及的 Twitter 数据集以及新浪微博数据集上对方法进行测试, 并主要从以下两个方面对结果进行评估: 主题检测模型的有效性和热点识别的效率.

4.1　有效性评估

为了验证方法的有效性, 我们采用准确率 (precision)、召回率 (recall) 以及 F-score 等指标来进行评估. 需要指出的是, 通过对微博用户行为分析, 绝大多数的用户宁可多花费时间来过滤非热点主题, 也不想错过真正的热点话题. 因此, 将评估重点放在召回率而不是准确率上. 我们期望所提出的方法 STD 能够保证较高的召回率.

测试集由中英文两种语言的数据组成. 其中英文测试集使用了 Twitter 数据 (详见 3.2 小节), 包含大约 16 万条推文, 由 400 个热点话题以及 300 个非热点话题组成. 中文测试集使用新浪微博数据, 包含大约 24374 条微博, 由 150 个热点话题以及 100 个非热点话题组成.

表 1 不同模型之间的有效性比较

Table 1 Evaluation of effectiveness between different models

Model	Twitter			Sina		
	Precision (%)	Recall (%)	F-Score (%)	Precision (%)	Recall (%)	F-Score (%)
Bursting model	60.5	48.1	53.6	58.5	50.2	54.0
V-Model	65.3	80.1	71.9	62.3	76.7	68.8
S-Model	57.5	63.7	60.4	55.4	63.1	59.0
V+U-Model	63.5	77.1	68.8	60.8	72.5	66.1
V+S-Model	65.2	82.3	73.3	67.1	83.6	74.4
V+U+S-Model	68.3	78.1	72.9	61.5	75.9	67.9

为了便于比较实验效果, 我们采用 Bursting 模型作为基准线模型, 同时比较了采用不同特征的主题检测模型以及复合模型. 其中, 将融合情感特征的话题检测模型标注为 S-Model, 将融合了微博数量特征的模型记为 V-Model, 将融合了微博用户特征的模型记为 U+Model. 除此之外, 我们还实现了采用其他特征的模型, 并对不同的特征模型进行了线性融合. 具体公式如下所示:

$$R(s) = w_1 R_v(s) + w_2 R_s(s) + w_3 R_u v(s). \tag{4}$$

实验中, 我们将各个模型检测出的热点话题与 Twitter 以及新浪微博所提供的热点话题 [14] 进行对比, 其实验结果如表 1 所示.

从实验结果可以看出, 作为实验方法的基准线, Bursting 模型的性能仅仅满足了用户的基本要求. 一个比较有趣的现象是尽管 Bursting 模型的召回率比较低, 但与其他模型相比, 其准确率还是达到了能接受的程度. 比较直观的解释是, Bursting 模型在判断一个时间序列是否会成为流行趋势的时候比较严格, 只有那些曲线变化非常明显的时间序列才能够满足判别条件.

很明显, 在各个特征之间, 微博数量是最有代表性的特征之一, 融合了微博数量特征的模型其召回率基本都达到了比较高的水准, 与召回率最高的模型相差不超过 5 个百分点. 不难理解, 微博数量通常还是热点主题形成的最主要的衡量指标. 这也是为什么一些传统的热点主题检测模型, 例如 Bursting 模型, 始终将微博数量变换曲线作为最重要的特征输入之一.

从表 1 中同样可以发现, 融合了情感特征的模型相比于基准线模型取得了更好的实验效果, 得到了较高的召回率, 这个特性也是我们期望的. 为了更好地分析实验结果, 我们比较了 S-Model 检测出的主题标签以及 V-Model 检测的主题标签, 并在表 2 和 3 中列出了其中一些典型样例.

从实验结果中可以发现 S-Model 更注重带有情感表述的话题, 尽管这些话题没有被 Twitter 公司制定为热点主题. 例如, #TellAFeministThankYou 是由 Melissa McEwan 发起的, 旨在对抗 Twitter 上女性骚扰. 这个事件震惊了整个美国社会, 很多新闻报道此事. 然而事件刚发生时, 其相关的关键词并没有呈现出显著的统计特征, 因此 Twitter 公司起初并没有将其列为热点主题. 但是, 在事件初期, Twitter 上与之相关的情感表达则非常强烈, 也正是根据这一现象 S-Model 对其进行了有效地识别. 同样的, 在新浪微博中也不难发现 # 就业季这一热点主题情况与之类似.

此外, 实验中我们同样比较了融合了多种不同特征的复合模型, 发现了一些有趣的现象. 用户数量这一特征本应具有较强的指示性, 然而当它与推文数量结合时, 性能反而下降了. 通过我们对实验结果的分析, 这是由于用户数量和推文数量之间的相关性是比较高的, 但当它们一起作为特征被使用时却产生了更多的噪声, 因而造成了性能的下降. 而我们所提出的复合模型利用了推文数量以及感情

表 2 Twitter 数据集上 V-Model 与 S-Model 检测主题典型样例比较

Table 2 Comparison of sampled hashtags detected by V-Modes and S-Model on Twitter dataset

V-Model	S-Model
#LoQueMasDeseoEs	#HappyBirthdayHarryFromLatinas
#mbv #WaysToPissO?YourValentine	#mbv #WaysToPissO?YourValentine
#NXZEROnoEncontro #giornatadellamemoria	#NXZEROnoEncontro #giornatadellamemoria
#EresLittleMonsterSi #TuCaraMeSuena15	#EresLittleMonsterSi #TuCaraMeSuena15
#PraSempreNossoEncantoPF	#PraSempreNossoEncantoPF
#RANHariBaru #10TheBestMoviesEver	#RANHariBaru #10TheBestMoviesEver
#QueremosBandaCineNoEncontro	#QueremosBandaCineNoEncontro

表 3 新浪微博数据集上 V-Model 与 S-Model 检测主题典型样例比较

Table 3 Comparison of sampled hashtags detected by V-Modes and S-Model on sina dataset

V-Model	S-Model
# 雾霾	# 雾霾
# 也门撤侨	# 也门撤侨
# 刘翔退役	# 刘翔退役
# 隆平超级稻	# 就业季
# 油价	# 日本改换教科书
# 央行降息	# 央行降息
# 日本改换教科书	# 延迟退休

等特征, 并在所有模型中取得了最高的 F-score, 且准确度几乎与 V-Model 不相上下. 所以, 从整体性能评估而言, 可以预期 V+S-Model 能够取得最好的效果.

4.2 检测效率评估

因为待比较模型是面向流数据的实时模型, 所以热点话题的检测速度同样非常值得关注. 实验中, 我们对热点话题检测时间作了对比. 需要注意的是, 由于检测时间受不同主题的影响比较大, 所以我们采用热点主题平均检测时间这一指标进行评估, 结果如表 4 和 5 所示.

从表 4 和 5 不难看出结论, 基准线模型 Bursting Model 时效性最差, 甚至是在 Twitter 公司已经报告了热点主题之后 10 分钟才识别出相应主题, 这显然不能满足用户需求. 而其他模型能够在不影响准确率的前提下, 比 Twitter 公司更早的发现热点主题. 在上述模型中, V+S+U-Model 检测时间较慢, 同 V+S-Model 相比, 其检测效果并不优于 V+S-Model, 但速度却相差很大.

取得最快检测时间的是 S-Model. 这说明, 在某一主题的开始阶段, 其感情极性或强度越激烈且具有一定稳定性, 其越有可能成为热点主题, 这一特征不仅可以辅助判定热点主题, 同时可以提高其识别效率. 同样的, 受情感特征的影响, V+S-Model 相比于 V-Model 也具有更快的检测时间.

5 结论

在本文中, 我们提出了一种基于情感序列的非参数话题监测模型. 该方法通过微博情感极性分析

表 4 Twitter 数据集热点主题检测时间比较

Table 4 Evaluation of detecting time between different models on Twitter dataset

Model	Detecting time (min)
Bursting model	10.5
V-Model	−18
S-Model	−33
V+U Model	−15
V+S-Model	−22
V+U+S-Model	−5

表 5 新浪微博数据集热点主题检测时间比较

Table 5 Evaluation of detecting time between different models on sina microblog datasets

Model	Detecting time (min)
Bursting model	9
V-Model	−16
S-Model	−30
V+U Model	−14
V+S-Model	−20
V+U+S-Model	−3.5

及其强度变化来计算情感时序分布, 并利用上述特征构建一个复合模型以识别、检测微博热点主题. 为了测试方法的有效性, 我们分别在 Twitter 和新浪微博等真实数据集中进行了相关实验, 将检测出的热门主题与 Twitter、新浪微博所展示的热点话题进行对比. 实验结果表明我们提出的方法能够在保证检测效果的前提下更快地发现热点话题. 同时, 由于方法结合了情感时序变化特征, 因而对于与主题相关的观点非常敏感, 能够快速有效地检测出具有强情感性的话题. 通过进一步地分析发现, 其中部分话题在当时并没有引起 Twitter 公司的重视, 但在未来的一段时间引起了社会的强烈反响.

参考文献

1 Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010. 591–600

2 Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010. 851–860

3 Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media. Stroudsburg: Association for Computational Linguistics, 2011. 30–38

4 Diao Q, Jiang J, Zhu F, et al. Finding bursty topics from microblogs. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. Stroudsburg: Association for Computational Linguistics, 2012, 1: 536–544

5 Gao Z J, Song Y, Liu S, et al. Tracking and connecting topics via incremental hierarchical dirichlet processes. In: IEEE 11th International Conference on Data Mining (ICDM), Vancouver, 2011. 1056–1061

6 Nikolov S. Trend or no trend: a novel nonparametric method for classifying time series. Dissertation for Ph.D. Degree. Boston: Massachusetts Institute of Technology, 2012

7 Mathioudakis M, Koudas N. Twittermonitor: trend detection over the twitter stream. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. New York: ACM, 2010. 1155–1158

8 Becker H, Naaman M, Gravano L. Beyond trending topics: real-world event identification on twitter. In: International

AAAI Conference on Web and Social Media, Barcelona, 2011, 11: 438–441
9 Cataldi M, Di Caro L, Schifanella C. Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the 10th International Workshop on Multimedia Data Mining. New York: ACM, 2010, 4
10 Gao W, Li P, Darwish K. Joint topic modeling for event summarization across news and social media streams. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012. 1173–1182
11 Hayashi K, Maehara T, Toyoda M, et al. Real-time top-R topic detection on Twitter with topic hijack filtering. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015. 417–426
12 Pang B, Lee L. Using very simple statistics for review search: an exploration. In: International Conference on Computational Linguistics (Posters), Manchester, 2008. 75–78
13 Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002. 79–86
14 Wiebe J M, Bruce R F, O'Hara, et al. Development and use of a gold-standard data set for subjectivity classifications. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1999. 246–253
15 Liu B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies. Morgan Claypool Publishers, 2012, 5: 1–167
16 Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping. In: Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg: Association for Computational Linguistics, 2003. 25–32
17 Wiebe J, Riloff E. Creating subjective and objective sentence classifiers from unannotated texts. In: Gelbukh A, ed. Computational Linguistics and Intelligent Text Processing. Berlin: Springer, 2005, 3406: 486–497
18 Wang X, Wei F, Liu X, et al. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM, 2011. 1031–1040
19 Joachims T. SVM-Light Support Vector Machine. Dortmund: University of Dortmund, 1999
20 Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2004. 271
21 Go A, Bhayani R, Huang L. Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report, Stanford, 2009. 1–12
22 Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Stroudsburg: Association for Computational Linguistics, 2010. 36–44

Incorporating sentiment series into trending topic detection on microblog

LI BinYang[1], HAN Xu[2*], PENG BaoLin[3], LI Jing[3], WANG TengJiao[4] & WONG Kam-Fai[3]

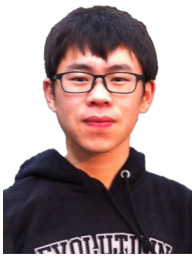
1 *College of Information Science and Technology, University of International Relations, Beijing* 100091*, China*;
2 *College of Information Engineering, Capital Normal University, Beijing* 100048*, China*;
3 *Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong* 999077*, China*;
4 *School of Electronic and Computer Engineering, Peking University, Beijing* 100871*, China*
*E-mail: hanxu@cnu.edu.cn

Abstract Twitter plays a significant role in information diffusion, and it has evolved to become an important information resource as well as news feed. There is a widespread interest in what is happening on Twitter, and the instantaneous news information that is passed on. However, with the large amount of data, it is impossible to manually determine what topic is trending, which makes real-time topic detection attractive and significant. Furthermore, Twitter provides a platform for the sharing of opinions and providing feedback for events, news, and products, etc. Because users tend to express their real thoughts on Twitter, it is recognized as a valuable source of opinions. Nevertheless, most works about trending topic detection fail to consider sentiments. In this work, we develop a non-parametric supervised real-time-trending topic-detection model with a sentimental feature. By performing experiments, we show that our model successfully detects trending sentimental topic in a short time. After applying a combination of multiple features, e.g., tweet volume and user volume, the proposed model demonstrates impressive effectiveness with an 82.3% recall rate, surpassing all of the competitors.

Keywords sentiment analysis, trending topic detection, sentiment series, real-time tracking, microblog

LI BinYang was born in 1982. He received the Ph.D. degree in 2012. Currently, he is an associate professor at the University of International Relations, and his research interests are natural language processing, sentiment analysis, and social computing.

HAN Xu was born in 1984. She received the Ph.D. degree in 2011. Currently, she is an assistant professor at the Capital Normal University, and her research interests are artificial intelligence and cloud computing.



PENG BaoLin is a Ph.D. student at the Chinese University of Hong Kong. He obtained his Master's degree from Beihang University and his Bachelor's degree from Yantai University. His research focus is on deep learning for natural language processing, especially end-to-end and data-driven dialog modeling and document summarization.

LI Jing is now a Ph.D. candidate in the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong. Before that, she received her B.S. degree from the Department of Machine Intelligence, Peking University, in July 2013. Her research interests are in automatic text summarization and NLP for social computing.