第47卷第4期 中国科学技术大学学报 Vol.47,No.4
2017年4月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2017

文章编号:0253-2778(2017)4-0328-08

基于用户行为影响的微博突发话题检测方法

万 越,隋 杰

(中国科学院大学工程科学学院,北京 100049)

摘要:考虑到微博数据存在时序性特征以及包含用户的社交网络行为特征,提出一种动量信号增强模型算法来有效地检测微博突发话题.由于传统模型未考虑微博数据变化以及用户社交行为的影响,为此首次提出影响力因子以及热度因子,用以修正动量模型.为获取影响力因子,将计算出当前时点前给定周期内的数据对当前数据的变化差值的指数累计影响作为影响力的衡量标准,以体现词频在该区间段的重要性.影响力因子将用于修正词频序列,以获取MACD值指标.由于用户的社交行为对话题产生影响巨大,进而提出热度因子用以修正MACD值指标.当模型满足指标阈值时,特征词则列为突发特征词.最后,通过K-means聚类算法将特征词进行归类合并,以获取突发话题.实验结果表明,模型精度能达到81.82%,表现良好.

关键词:突发话题;信号增强;热度;影响力;动量;聚类

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.04.007

引用格式:万越,隋杰.基于用户行为影响的微博突发话题检测方法[J].中国科学技术大学学报,2017,47(4):328-335.

WAN Yue, SUI Jie. Bursty topic detection method for microblog based from influence of user behaviors[J]. Journal of University of Science and Technology of China, 2017,47(4):328-335.

Bursty topic detection method for microblog based on influence from user behaviors

WAN Yue, SUI Jie

(*School of Engineering Science, University of Chinese Academy of Science, Beijing* 100049, *China*)

Abstract: Social networks are becoming more and more popular where people can post anything anytime. Due to the huge user community, social network data is increasing with each passing day. Therefore how to explore the knowledge in huge data seems to be hard work. As microblog has time-related characteristics and social network behavior attributes, momentum signal enhancement model is put forward to detect bursty microblog topics effectively. Influence factor and hot energy factor are put forward to improve the momentum model. The influence factor uses the data before the current point but within a given period to calculate the difference between corresponding data and current point data, thus the difference represents the influence of a single word in one period. Therefore data series adjusted by the influence factor will be put into the momentum model to get MACD value. Then, social network behaviors are also considered in adjusting MACD value, which is represented by the hot energy factor. After the word satisfies the requirements of MACD value and MACD change value, it becomes a bursty word. Finally, K-means

收稿日期:2016-08-28;**修回日期**:2016-12-08

基金项目:国家自然科学基金(61572459)资助.

作者简介:万越,男,1992年生,硕士生.研究方向:数据挖掘及机器学习. E-mail: wanyueucas@163.com

通讯作者:隋杰,博士/副教授. E-mail: suijie@ucas.ac.cn

algorithm is used to assign bursty words to different groups. The experimental results show that the detection accuracy of the proposed algorithm is up to 80%, which indicates a good detection performance.

Key words: bursty topic; momentum; signal enhancement; hot energy; influence

0　引言

微博是时下较为热门的社交媒体工具，它允许用户在 140 个字内发布信息，内容包括图片、链接等.随着用户使用量的增加，微博的发布数量激增，因此如何从海量的信息中合理、高效地挖掘出具有实际价值的知识成为研究领域所关注的重点.国内外研究人员对突发话题的检测方法众多.Gagli 等[1]提出实时数据流分析方法，运用软频率模式挖掘算法，基于动态窗口选择，深入挖掘信息；Fung 等[2]提出了无参的检测方法，它通过使用超几何分布的方式来说明突发话题的特征，先以 sigmoid 函数衡量特征权重，然后在贝叶斯概率模型的基础上检测出话题；Diao 等[3]提出了 LDA 模型，说明了信息在同一时段上具有相似性质和同一用户的信息在内容上也具有相似性；于海峰等[4]根据最佳投影方向对观测样本数据进行降维，将投影特征值隐含的风险信息在风险指标论域内进行扩散，获得了突发事件不同风险等级发生的概率，利用风险熵预测突发事件发生的可能性；Kleinberg 等[5]提出了一种较为经典的检测算法，它通过构建话题状态序列及其多状态对应的概率来模拟突发话题特征的频次走势，从 2 状态模型泛化到多状态模型，以隐马尔科夫的方法来分析词频在突发状态下的分布和自动机的转移概率；Chen 等[6]采用特殊的特征词分隔方法，采用特殊的特征词处理技巧，分别计算分隔区间的加和、标准差，按照对应的公式计算结果，分析特征的相似性；He 等[7]在已有动量学模型的基础上对模型和参数进行赋权调整，设定参数步长来训练检测权值和分析参数对动量学模型的影响；贺敏等[8]提出了有意义的串比单个特征词更有说服力的论断，有意义的串所含信息更为丰富，在模型中的检测效果更佳，其实验结果也说明有意义串在数据检测上的确能够比单个特征词取得更理想的检测性能；He 等[9]在基金申请研究热点分析上使用了动量学模型，对包括特定话题的突发能量、话题频率增长、即时频率等指标的指数形式建模，构造特定词汇的突发权值，并将其运用到动量学模型中；Du 等[10]认为，用户关系在话题检测中有意义，应该将用户收听和被收听的关系纳入词频计算中；王征等[11]提出了基于信息密度的新型检测模型 MBID，该模型通过动态滑动窗口采集微博信息流，以话题树进行信息归纳，并最终通过窗口和话题树的信息密度变化进行监测，发现突发话题；申国伟等[12]提出了微博窗口的突发话题检测模型.该模型根据消息流动态地调整窗口的大小，并且在传播影响力上加入了随时间变化的突发权值来检查话题；贺敏等[13]通提出了基于时序分析的微博突发话题检测方法. 通过动量模型提取候选突发特征后，对特征的动量时间序列分别借鉴信号频域分析理论和股票趋势分析理论进行建模；郭跃秀等[14]通过凝聚式层次聚类算法对突发词集合进行聚类检测，并在簇间阈值、时间性能和指标衡量等方面进行对比；徐志明等[15]讨论微博的用户关系分析技术，将用户关系强度定义为用户之间的相似度，并给出了用户相似度计算的方法；毛佳昕等[16]主要探讨了用户的社会影响力，分析了用户行为因素之间的关系，提出了预测用户信息传播能力和社会影响力度量的方法；陈克寒等[17]提出一种基于两阶段聚类算法 GCCR，将图摘要方法和基于内容相似度计算方法结合，从而推荐主题给相应用户.

上述方法虽然在微博话题发现领域均取得了相应的成效，但在突发话题的能量预测性和当前突发性大小的定量分析方面，动量模型相对做得更好.目前研究领域关于动量模型仅涉及应用层面，并没有考虑微博数据的时间特征、用户的网络行为特点，导致动量模型的精确性有所欠缺.为此，本文从这两个方面入手，着力修正原有模型的不足，以期提高检测效果.

1　微博突发话题检测方法

1.1　突发话题定义

突发话题定义为在短时间内吸引足够关注度的话题.满足定义的话题可以认为是突发话题.这个定义可以分解为三部分：

①)时间特征.突发话题应是发生于邻接时间区

间,而不是一段时间过程;

②数量特征.在时间窗口 t-1 上的微博数不多,但在时间窗口 t 上的微博数呈迸发式增长;

③用户行为特征.这一行为和用户的社交行为特征密切相关,话题是由用户的行为促成.

1.2 动量特征定义

动量模型的主要特点是观察词频序列的发展趋势,以检测出当前状态是否存在突发性.模型的核心是计算出词频序列的一阶序列和二阶序列.其中,一阶序列就是得到 MACD 值指标的序列,体现的是突发性的大小;二阶序列是获取 MACD 变动值指标,用于表示词频序列的未来发展趋势.相类的参数含义表达如下:

质量 $M(t)$指的是在序列在时间区间 t 时的重要性.文中的热度因子体现的是质量 $M(t)$,它被用来表示 MACD 值指标在某一时间区间的重要程度.位移 $X(t)$如同股票数据序列,该序列蕴含着时间维度.文中的词频序列即为位移 $X(t)$.速率 $V(t)$是对位移 $X(t)$的一阶求导,它体现的是单位时间内的变化程度.文中的 MACD 值指标就是速率 $V(t)$,用于检测是否存在突发信号.动量 $P(t)$等于质量 $M(t)$乘以速率 $V(t)$,旨在将用户的社交行为因素考虑进动量模型的 MACD 值指标.动量方向 $H(t)$等于单位时间内的动量 $P(t)$变化.它是词频序列的二阶序列,用于预测特征词的发展方向.文中的动量方向亦称为 MACD 变化值.

1.3 动量信号增强模型

为获取更好的检测效果,需要对词频数据先进行数据平滑,以减少噪声影响.

$$\left.\begin{aligned}&\mathrm{EMA}(n)[X]_t=\\&\alpha x_t+(1-\alpha)\mathrm{EMA}(n-1)[x]_{t-1}=\\&\sum_{k=0}\alpha(1-\alpha)^k x_{t-1}\\&\alpha=\frac{n}{1-n}\end{aligned}\right\}\quad(1)$$

式中,α 代表对当前数据的重视程度;X_{t-k}:表示 $t-k$时点的数据大小,即位移 $X(t-k)$.$\mathrm{EMA}[X]_t$ 是第 t 时间点上的移动平均量;n 为移动平均周期.

在获取到平滑后的时间序列后,可以进行动量的计算,即通过两个不同周期的指数移动平均序列的差值来获取动量大小,即

$$\mathrm{MACD}(n_1,n_2)=\mathrm{EMA}(n_1)-\mathrm{EMA}(n_2)\quad(2)$$

式中,n 为选取的时间周期.此处的 MACD 值就是速率 $V(t)$.为了获取词频 MACD 值的变化,我们需要对动量变化的方向 $H(t)$进行计算.

$$H(t)=\mathrm{MACD}(n_1,n_2)-\mathrm{EMA}(n_3)[\mathrm{MACD}(n_1,n_2)]\quad(3)$$

通过当前的 MACD 值大小经过移动平均后的 MACD 值相减得到 $H(t)$,也称 MACD 变化值.当 MACD 变化值为正,则表示未来的词频动量将继续增强;反之,则减弱.

综上,若同时满足超过 MACD 值指标阈值和 MACD 变化值指标阈值两个条件,则表示该特征词出现了突发特征.由于传统动量模型仅在数据层面对特征词进行计算,并未考虑微博数据的时间特征和用户的网络行为特征,导致模型的检测效果不能进一步提升,因此后续部分将从这两个方面来做相应的分析探讨,以期获取更高的检测效果.

1.4 影响力因子度量

由于微博数据之间存在相关性,即前一段时间的微博发布情况会对当下时点的信息状态产生影响.比如三天前的数据呈一个稳步增长的态势,那么今天也很可能继续沿着稳步增长的态势继续增长,但时间具有衰减特性.昨天的数据对今天的数据影响力应该是基于一周前的微博对今天的影响,而这样的特征体现的就是时间衰减性.另外微博的发布数量呈指数型增长,相应地,时间影响力衰减也应该体现指数特征.

$$X_k=\sum_{t=1}^{M}(N_k-N_{(k-t)})\times \mathrm{e}^{-t}\quad(4)$$

式中,X_k 为前 N 个时段的微博发布情况对当前特征的影响;N_t 为每个时间段 t 内关于特征词的微博发布总量;M 为设定的影响周期.

影响力大小体现的是前一段周期的数据对时下数据的影响.如果前一段周期内的话题讨论量较多,说明获得用户的关注度更高,自然也会对当下的数据造成积极影响;反之,则认为当下数据的词频会随着衰减的影响力而不断减小,因此将影响力序列赋予各个特征词序列,得到以下表达式.

$$[yf(1)\times X_1^{\alpha},\ yf(2)\times X_2^{\alpha},\cdots,yf(T)\times X_T^{\alpha}]\quad(5)$$

式中,$yf(i)$是时间区间 t 时的词频序列,也即位移 $X(t)$.X_T^{α} 是影响力因子,其中 α 代表影响力的权重.

1.5　热度因子度量

现有的动量模型未考虑用户的社交行为特征，而该特征在话题的检测中又相当重要，因此定义用户社交行为的热度因子用于修正动量模型.质量 $M(t)$ 是影响力、重要性，用热度因子来体现.由于随着时间的变化，话题的讨论量应该是呈下降趋势的，旧话题被新话题逐渐代替，影响力也应该呈下降趋势.如果突发话题的词频变化呈上升趋势，那么就说明该段时间中网友的讨论量有所增加，源于该话题吸引了更多关注.若想要保持原来的序列值甚至提高序列值，那么还需要一定的热度支持，只有更多人参与到话题的讨论中，才能使序列不至于降低.MACD 体现的是数据的变化趋势，而热度同样也是预示着话题讨论的趋势，因此结合热度来探讨一阶序列的增长变化，用以下定义来衡量重要性的大小.

$$\begin{aligned} &\text{if } y_t > y_{t-1}\ F_t = \mathrm{e}^{(y_t - y_{t-1}) \times \frac{\text{hot}}{\sum \text{hot}}} \\ &\text{if } y_t < y_{t-1}\ F_t = 1 \end{aligned} \tag{6}$$

式中，hot 表示某时间段 t 内所含特征词的点赞数、评论数、收藏数；$\sum$hot 表示整个时间区间所含特征词的点赞数、评论数、收藏数的总数；F_t 表示时间段 t 内的数据重要性大小；y_t 为时间区间 t 上的词频序列.在获取到热度因子后，需要对 MACD 动量序列进行修正，以反映用户社交行为对微博话题趋势的影响，因此将热度因子值分别乘以对应的 MACD 得到新的修正后值.公式表示为

$$\mathrm{MACD}(n_1, n_2)_t = \mathrm{MACD}(n_1, n_2)_t \times F_t \tag{7}$$

如果当前区间的词频数值大于上一区间的词频数值，就说明该特征词的热度在本区间内较为明显，也即网友对含该特征词的微博的关注度有所提升，那么就可以对相应的重要性赋予较高权重；反之，如果当前区间的词频数值小于上以区间的词频数值，说明很可能是微博的关注度有所下降，本文不作修改，仍旧按照原趋势计算.本文不对重要性赋予低于 1 的权值的原因在于：①算法检测的是突发话题，即短时间内从无到有或者从低水平值到高水平值，体现的是一个向上增量的关系；②如果降低权值，则降低权值后的序列值在遇到下一期原本不高的数据值看起来变化更大，会造成虚假的突发信号.

1.6　获取突发话题

话题获取部分采用的方法是使用 K-means 聚类算法来实现对突发词集的合并.由于 K-means 聚类算法是一种无监督算法，能够自动依靠词频序列之间的相似性进行合并计算，动态调整类簇的中心，分类准确率较高.相较于其他话题合并方法而言，K-means 聚类算法具有简单、易于理解和实现和具有较低时间复杂度的特点，因此本文用 K-means 聚类算法合并特征词.

满足 MACD 值指标和 MACD 变化值指标阈值后的特征词是突发特征词.随之初始化 K 个随机数据作为簇类的中心点，初始化的方式由原始序列数据位于最大值、最小值之间的随机值确定.按照以下两个公式迭代方法得到最终的 K 类的中心点以及标记后的突发特征序列数据.

$$c^{(t)} = \underset{j}{\operatorname{argmax}} \| x^{(i)} - u_j \|^2 \tag{8}$$

$$u_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \tag{9}$$

设定迭代的停机条件为中心距离的偏离值小于 α，即各类的中心点的偏离值趋于收敛.聚类所得的各簇类即为突发话题.

2　实验方案及分析

所有的实验都是在 Intel 酷睿 i5-3230MCPU，4G 内存，64 位操作系统上实现.候选特征词的获取采用的是轻量级中文分词包 IKAnalyzer，它是以开源项目 Lucene 为应用主体，结合词典分词和文法分析算法的中文分词组件.所有实验均通过 Java 程序编写.实验数据时间为 2014 年 4 月 29 日至 2014 年 5 月 12 日，按照 3 小时作为一个区间计算，共计 84 168条数据.在数据预处理上，将粉丝数低于 10 个的用户予以剔除；对于同一 ID 在一个小时内发出的三条以上的内容近似微博，仅留前两条；将微博日发布量排在前 1/4 的用户的微博重要度减少一半，排在后 1/3 的用户的微博重要度增加一倍.实验结果通过聚类算法获取，最终得到算法检测后的突发话题.实验人员通过分离出的突发话题与直接手动标记的突发话题进行对比，获取算法检测的准确度.

2.1　特征词实验结果

首先获取 2014 年 4 月 29 日至 2014 年 5 月 12 日之间的热门词汇，经过数据预处理，将经过影响力因子修正后的特征词序列作为动量信号增强模型的输入端.在阈值的设定方面，当突发阈值设的过高，能满足阈值的特征词较少，检测效果不佳；当突发阈值设的过低，满足阈值的特征词较多，易造成混乱.实验表明，当 MACD 值阈值为 0.015，MACD 变化值阈值为 0.01，检测评价指标最优.如果特征词能分

别超过该阈值,则认为在该时间段能产生突发性.图1(a)、图1(b)表示的是动量信号增强模型检测特征词的图像,而图1(A)、图1(B)是传统动量模型检测特征词的图像.横轴代表时间区间,纵轴代表衡量突发性的能量值.

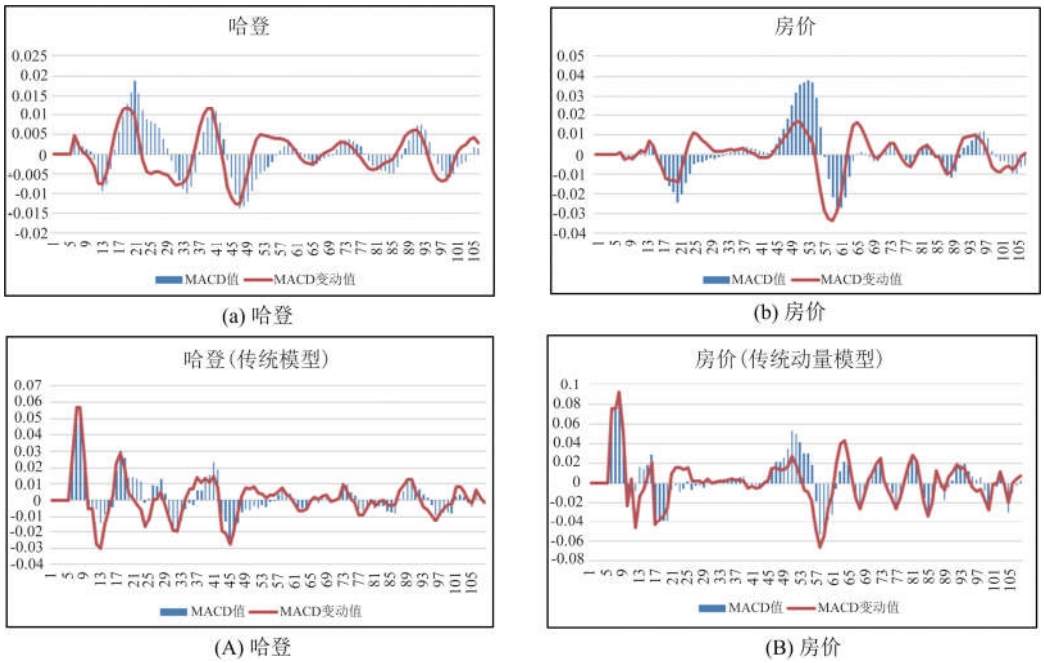

图1 特征词能量图

Fig.1 Energy of feature word

图1(a)显示的特征词是哈登,表示的是2013-2014赛季NBA季后赛的赛事情况.季后赛的比赛时间为4月21日至5月3日,对应的微博发布数量最高点的时间在5月3日火箭队对开拓者最后一场比赛,火箭队被绝杀淘汰.从图1(a)中可见,在当时时点上,MACD的能量值达到高峰,MACD值接近0.02,突破了设定的突发阈值0.015.由于火箭队的中国粉丝基础较好,并且关于被看好的火箭队再次止步季后赛首轮的结果,因此在随后的几天时间充满了关于哈登以及主帅、林书豪等的赛场表现的分析讨论,不过由于没有具体事件的影响因素,因此关于哈登的关注度也较为分散,随后的MACD值大多位于0.005以下,没有达到突发性的阈值,符合实际情况.从MACD变动值上看,当该词体现突发特征时,变动值指标同样超出阈值0.01,虽然后续部分也偶有超过阈值,但由于不满足MACD值指标,故不构成突发性.图1(a)是动量模型下的特征词检测图像,最开始部分显示了很强的突发信号,但从文本数据看并未体现突发性,原因在于初始的文本数据量小和词频量小,导致词频序列略微增大后,能够得到较大的MACD值,缘于基数的变化.此时用户行为特征也较小,故若能考虑热度因子,就能降低MACD值的过快增长,过滤不必要的虚假信号.

图1(b)是关于房价大跳水后的讨论.5月6日左右,温州、长沙、武汉等地房价争相降价,月最大降幅比例在4.61%,引发了网友关于全国房价大跳水的广泛争论.因为住房是民生不可或缺的一环,所以这次降价吸引了足够的关注.从图1(b)中的MACD绝对值的基数大小来看,房价的MACD值在一定程度上要远超过前一个特征词MACD值.房价的MACD值大小最高时能够达到接近0.04,远超突发阈值.相较图1(B),图1(b)在中间部分的MACD值一度拔高,远超其他区间段,而图1(b)部分则相对不明显.文本数据显示,该段的微博发布量并不是最多的,但是评论数却屡创新高,用户之间的互动相对频繁,热度因子刻画该状态得到了图1(b)的走势.从这个特征来说,刻画了用户的社交行为后的特征的MACD值应该更准确地把握了整个过程的趋势走势.

2.2 实验评价

在传统的实验结果检验中,准确率(Precision)、召回率(Recall)、F值(F-measure)是检验实验结果的重要指标.由于不存在可供参考的标准事件库,导致无法知晓给定区间段的突发话题内容,因此传统

的检验指标难以衡量实验结果的准确性.为更有效地验证模型的准确性,在此通过大量的实验观察以及手工提取突发话题的方法来抓取突发话题,近似地作为实际发生的突发话题.此处采用以下方式计量对应的指标:

$$\text{Precision}=\frac{\text{right\_classify}}{\text{total right\_classify}} \tag{10}$$

$$\text{Recall}=\frac{\text{right\_classify}}{\text{total\_classify}} \tag{11}$$

$$F\text{-measure}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{12}$$

式中,right_classify 是被模型正确分类的特征词数目;total right_classify 是标准库中分类的特征词数目;total_classify 是模型分类出的所有特征词数目.

此处 Precision 指标表示特征词被正确分类的准确率.Recall 指标表示模型在本身检测出的特征词中,被正确分类的概率.F-measure 是综合指标,表示模型的整体检测水平.表 1 是实验准确率与话题分类数的关系,图 2 是话题分类准确率的结果.

表 1　实验精度比较(%)

Tab.1　Comparision of accuracy of cluster numbers

话题数	7	8	9	10	11	12
Precision	58.9	64.64	76.36	81.82	69.09	61.82
Recall	50.01	62.50	55.56	70.21	72.73	63.64
F-measure	54.09	63.53	64.32	75.57	70.86	62.73

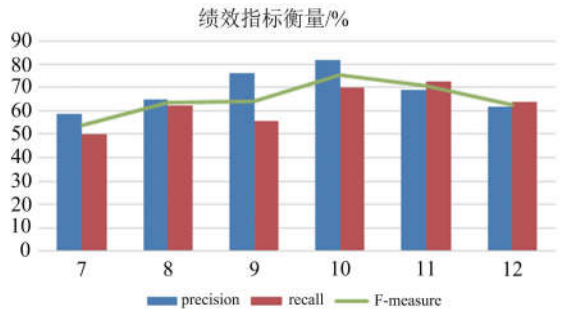

图 2　话题分类准确率

Fig.2　Accuracy of topic classification

从整体的话题分类准确率上看,其准确率大致都在 60%以上.随着话题类别数量的不断增加,准确率呈上升的趋势.当话题数为 7 时,准确率为 58.9%,表示将特征的词频序列划分到 7 类话题中的准确率只有 60%左右,而当所给定的话题分类数目增加时,特征词的选择范围更广泛,准确率从 64.64%增加到 76.36%,相比于给定 7 类话题的准确率要高很多.当分类为 10 类时,准确率达到 81.82%,相比于 7 类分类的准确率高出将近 20%,而比 9 类话题要高出 5%.当给定的话题分类更多时,那么原有的最优话题分类将被打破,使得特征词不得不重新归于新的类别.从表 1 可以看出,准确率迅速降到 70%以下,降低幅度超过 10%.当话题分类数目进一步增加时,分类的准确率降低到 61.82%,相比于 11 类话题要减少 7%左右.整体趋势表明,随着话题分类的不断增加,准确率提升显著,说明特征词的归类更加合理;当话题分类数目超过最优值,即图 2 中的 10 类话题,准确率降低,因此从图表中可以看出,最优的话题分类为 10 类.从 F-measure 值的水平看,随着话题数不断增加,F-measure 值也相应增长,达到 10 类话题时的 75.57%,F-measure 值此时处于最优状态.随着话题数的进一步增加,使得 F-measure 值也相应减少,F-measure 值由最高的 75%下降到 63%左右,表明当话题数设定为 10 类时,特征词能够被正确分类.最终的突发话题及其特征词检测结果如表 2 所示.

表 2　检测出的十类话题所属特征词

Tab.2　Detected feature words of topics

话题分类	特征词
NBA 季后赛	林书豪、哈登、利拉德、开拓者、防守、霍华德、火箭、绝杀、赛季、助攻
雾霾天气	雾霾
转基因食品	转基因、大豆、安全、农业部、食品
小米手机发布会	新浪、小米、发布会、魅族、浏览器、客户端、手机
贪官	贪官、高层、中国、美国、政府
房价	北京、超速、房价、成交、深圳、财政部、开发商、城市
《同桌的你》	电影、同桌的你、青春、大学
聘任制	聘任制、公务员
广州恒大	恒大、广州、大阪、中超、足球
韩剧	韩国、韩剧

从表 2 的结果来看,单个话题至少包含 1 个特征词,最多包含 10 个特征词。前者为雾霾天气话题,后者为 NBA 季后赛。从特征词的语意来看,特征词的关联度较大,通常出现在同一微博语句中,因而被检测、划定为同一话题的概率较大。

2.3 动量信号增强模型与其他动量模型的比较

贺敏等[8]提出的有意义的串动量模型是话题检测领域较为经典的算法，它能在较小复杂度的条件下完成对特征词的突发性检测.有意义的串的动量模型能够更好地挖掘出话题信息的原因在于有意义的串包含更多的信息，稳定性更强.本文将动量信号增强模型与传统动量模型和有意义串动量模型做比较，结果如表3所示.

表3 动量信号增强模型与其他动量模型的比较(%)

Tab.3 Comparision of momentum signal enhancement model and other models

	Precision	Recall	F-Measure
动量信号增强模型	81.82	70.21	75.57
有意义串动量模型	72.67	73.25	72.96
传统动量模型	64.69	68.21	66.4

从表3的对比结果看，动量信号增强模型在精确率指标上占优，达到81.82%，远超过有意义串动量模型的72.67%和传统动量模型的66.4%.在召回率指标上，动量信号增强模型的比率为70.21%，而传统动量模型也较为近似，达到68.21%的水平，而有意义串动量模型的表现则最优，达到73.25%.总体来看，动量信号增强模型的F-measure指标能够达到75.57%，超过有意义串动量模型大约2.6%左右，超过传统动量模型的指标值9%左右.

从数据的结果来看，动量信号增强模型能够达到较高的准确率，准确地把突发词挖掘出来，这得益于动量信号增强模型对突发信号的敏感性，能放大突发信号，有利于检测，这一特征上与传统动量模型有所近似，而前者检测效果好于后者的原因在于前者考虑了微博的时间特征以及用户的社交行为特征，对数据的处理更占优势，因此动量信号增强模型效率更高.有意义串动量模型是借助有意义的串来进行话题的挖掘，稳健性更高，这一点在召回率上有所体现.总体来看，动量信号增强模型在精确率上要高于后两者10%左右，而在召回率上则落后3%，但总体的75.6%要优于后两者，因此动量信号增强模型在检测上有一定的优势.

3 结论

本文通过包含影响力因子和热度因子的动量信号增加模型检测出突发特征词，使用K-Means聚类算法对特征词进行聚类，获取突发话题.实验表明，在实验条件下，当选择的话题数达到10个时，准确率最高达到81.82%，说明动量信号增强模型在突发话题获取上具有一定的优势，具有一定的参考价值.考虑到用户的关联关系以及用户影响力的传导可能对话题产生深刻影响，因此未来计划进一步研究用户关系在突发话题检测上的重要性.

参考文献(References)

[1] GAGLIO S, RE G L, MORANA M. A framework for real-time Twitter data analysis [J]. Computer Communications, 2016, 73: 236-242.

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

[3] DIAO Q M, JIANG J, ZHU F D, et al. Finding bursty topics from microblogs[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea: Association for Computational Linguistics, 2012: 536-544.

[4] 于海峰，王延章，卢小丽，等. 基于知识元的突发事件风险熵预测模型研究[J]. 系统工程学报，2016，31(1)：117-126.
YU Haifeng, WANG Yanzhang, LU Xiaoli, et al. Emergency risk entropy forecasting model based on knowledge element [J]. Journal of Systems Engineering, 2016, 31(1): 117-126.

[5] KLEINBERG J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373-397.

[6] CHEN Y, YANG S, CHENG X Q. Bursty topics extraction for web forums[C]// Proceedings of the 11th International Workshop on Web Information and Data Management. Hong Kong, China: ACM Press, 2009: 55-58.

[7] HE D, PARKER D S. Topic momentums: An alternative model of bursts in streams of topics[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D C, USA: ACM Press, 2010: 443-452.

[8] 贺敏，杜攀，张瑾，等. 基于动量模型的微博突发话题检测方法[J]. 计算机研究与发展，2015，52(5)：1022-1028.
HE Min, DU Pan, ZHANG Jin, et al. Microblog bursty topic detection method based on momentum model [J]. Journal of Computer Research and Development, 2015, 52(5): 1022-1028.

[9] HE D, PARKER D S. Learning the funding

momentum of research projects [J]. Knowledge Discovery and Data Mining, 2011, 6635(2):532-543.

[10] DU Y, HE Y, TIAN Y, et al. Microblog bursty topic detection based on user relationship[C]// Proceedings of the 6th IEEE Joint International Information Technology and Artificial Intelligence Conference. Chongqing, China: IEEE Press, 2011, 1: 260-263.

[11] 王征, 王林森, 赵磊. 基于信息密度的微博突发话题检测模型研究[J]. 情报理论与实践, 2016, 39(3): 125-129.

[12] 申国伟, 杨武, 王巍, 等. 面向大规模微博消息流的突发话题检测[J]. 计算机研究与发展, 2015, 52(2): 512-521.
SHEN Guowei, YANG Wu, WANG Wei, et al. Burst topic detection oriented large-scale microblogs streams [J]. Journal of Computer Research and Development, 2015, 52(2): 512-521.

[13] 贺敏, 徐杰, 杜攀, 等. 基于时间序列分析的微博突发话题检测方法[J]. 通信学报, 2016, 37(3): 48-54.
HE Min, XU Jie, DU Pan, et al. Bursty topic detection method for microblog based on time series analysis[J]. Journal on Communications, 2016, 37(3): 48-54.

[14] 郭跑秀, 吕学强, 李卓. 基于突发词聚类的微博突发事件检测方法[J]. 计算机应用, 2014, 34(2): 486-490, 505.
GUO Yixiu, LYN Xueqiang, LI Zhuo. Bursty topics detection approach on Chinese microblog based on burst words clustering [J]. Journal of Computer Applications, 2014, 34(2): 486-490, 505.

[15] 徐志明, 李栋, 刘挺, 等. 微博用户的相似性度量及其应用[J]. 计算机学报, 2014, 37(1): 207-218.
XU Zhiming, LI Dong, LIU Ting, et al. Measuring similarity between microblog users and its application [J]. Chinese Journal of Computers, 2014, 37(1): 207-218.

[16] 毛佳昕, 刘奕群, 张敏, 等. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报, 2014, 37(4): 791-800.
MAO Jiaxin, LIU Yiqun ZHANG Min, et al. Social influence analysis for micro-blog user based on user behavior[J]. Chinese Journal of Computers, 2014, 37(4): 791-800.

[17] 陈克寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报, 2013, 36(2): 349-359.
CHEN Kehan, HAN Panpan, WU Jian. User clustering based social network recommendation [J]. Chinese Journal of Computers, 2013, 36(2): 349-359.

(上接第 303 页)

[8] 余伟, 李石君, 杨莎, 等. Web 大数据环境下的不一致跨源数据发现[J]. 计算机研究与发展, 2015, 52(2): 295-308.
YU Wei, LI Shijun, YANG Sha, et al. Automatically discovering of inconsistency among cross-source Data based on Web big Data [J]. Journal of Computer Research and Development, 2015, 52(2): 295-308.

[9] 张安珍, 门雪莹, 王宏志, 等. 大数据上基于 Hadoop 的不一致数据检测与修复算法[J]. 计算机科学与探索, 2015, 9(9): 1044-1055.

[10] 金连, 王宏志, 黄沈滨, 等. 基于 Map-Reduce 的大数据缺失值填充算法[J]. 计算机研究与发展, 2013, 50: 312-321.

[11] 罗元剑, 姜建国, 王思叶, 等. 基于有限状态机的 RFID 流数据过滤与清理技术[J]. 软件学报, 2014, 25(8): 1713-1728.
LUO Yuanjian, JIANG Jianguo, WANG Siye, et al. Filtering and cleaning for RFID streaming *Data* technology based on finite state machine[J]. Journal of Software, 2014, 25(8): 1713-1728.

[12] 陈振国, 田立勤. 信任模型在雾霾感知源评价中的应用[J]. 计算机应用, 2016, 36(2): 472-477.
CHEN Zhenguo, TIAN Liqin. Application of trust model in evaluation of haze perception source [J]. Journal of Computer Applications, 2016, 36(2): 472-477.

[13] 罗涛, 李俊涛, 刘瑞娜, 等. VANET 中安全信息的快速可靠广播路由算法[J]. 计算机学报, 2015, 38(3): 663-672.
LUO Tao, LI Juntao, LIU Ruina, et al. A fast and reliable broadcast routing algorithm for safety related information in VANET [J]. Chinese Journal of Computers, 2015, 38(3): 663-672.

[14] 田立勤, 林闯, 张琪, 等. 物联网监测拓扑可靠性设计与优化分析[J]. 软件学报, 2014, 25(8): 1625-1639.
TIAN Liqin, LIN Chuang, ZHANG Qi, et al. Topology reliability design and optimization analysis of IoT-based monitoring[J]. Journal of Software, 2014, 25(8): 1625-1639.

[15] 钟晓睿, 马春光. 基于动态累加器的异构传感网认证组密钥管理方案[J]. 通信学报, 2014, 35(3): 124-134.
ZHONG Xiaorui, MA Chunguang. Dynamic accumulators-based authenticated group key management scheme for heterogeneous wireless sensor network [J]. Journal on Communications, 2014, 35(3): 124-134.