文章编号：2095-4980(2023)03-0378-06

基于改进K-均值的微博热点话题发现方法

陈阳键[1]，温秋华[2]

(1.广州开放大学(广州市广播电视大学) 数字化服务中心，广东 广州 510000；2.暨南大学 信息科学技术学院，广东 广州 510000)

摘　要：微博文本数据高维度、同义、多义特征明显，传统基于向量空间模型(VSM)联合K-均值的热点话题发现方法存在准确率低，计算复杂，聚类中心难以确定等问题。提出一种相关向量机(RVM)优化VSM的微博文本向量化方法，首先利用RVM的自适应特征选择能力对VSM特征向量进行降维，然后利用主成分分析(PCA)方法确定K-均值算法的初始聚类中心，进而采用K-均值算法得到聚类结果，最后根据微博转发、评论和高影响力用户数量定义热度指数，热度指数最大的话题即为当前热点话题。采用实际微博文本数据集开展实验，结果表明所提方法相对于2种传统方法的准确率分别提升7.3%和1.1%，实时性分别提升45%和53%。

关键词：热点话题发现；向量空间模型；话题聚类；数据降维；微博

中图分类号：TP391　　**文献标志码：**A　　**doi：**10.11805/TKYDA2020457

Micro-blog hot topic detection method based on improved K-means

CHEN Yangjian[1]，WEN Qiuhua[2]

(1.Digital Service Center，Guangzhou Radio and Television University，Guangzhou Guangdong 510000，China；
2.School of Information Science and Technology，Jinan University，Guangzhou Guangdong 510000，China)

Abstract：Micro-blog text data is high-dimensional, bearing the obvious features of synonymy and polysemy. Traditional topic detection method based on Vector Space Model(VSM) combined with K-means has some problems such as low accuracy, complex calculation, and being difficult to determine the center of clustering. A Relevance Vector Machine(RVM) optimized VSM method is proposed to realize the text vectorization. Firstly, the dimension of VSM feature vector is reduced automatically by using the adaptive feature selection ability of RVM, and then Principal Component Analysis(PCA) is applied to determine the cluster center of K-means clustering algorithm. K-means algorithm is employed to get the clustering results. Finally, according to the number of micro-blog forwarding and comments, the topic with the largest heat index is the current hot topic. The results show that compared with two traditional methods, the accuracy of the proposed method is improved by 7.3% and 1.1%, and the real-time performance is improved by 45% and 53%, respectively.

Keywords：hot topic detection；Vector Space Model；topic clustering；data dimensionality reduction；Micro-blog

微博是随着互联网技术和Web2.0时代的发展而兴起的一种新的网络媒体形式，不同于报纸、书信等传统媒体采用长文本信息交互方式，微博用户可以通过网页、移动客户端、即时通信软件等多种途径，采用140字以内的文本、图片和影音等多种媒体内容，随时随地分享身边信息，表达个人观点以及获取最新时事。由于微博具备操作简单、互动性强、传播速度快、时效性强等特点，迅速在广大网民间得到普及。以新浪微博为例，其2020年第一季度的月活跃用户高达5.5亿，日活跃用户达2.41亿。作为一种信息传播媒介，微博在为用户提供新鲜及时、丰富多样信息的同时，也带来了严重的信息过载和信息碎片化问题，在这种情况下，如何快速、有效地对海量数据进行分析整合，准确提取当前热点话题，改善用户体验，同时实现网络事件监测、网络舆情控制等功能，对网络信息安全具有重要意义[1-2]。

话题检测与跟踪技术(Topic Detection and Tracking，TDT)最早由美国国防高级研究计划署提出，旨在帮助人

收稿日期：2020-09-14；修回日期：2021-02-09

基金项目：广东省广州市高校第九批教育教学改革基金资助项目(2017F10)

们在信息爆炸环境下更好地获取有用信息，常用的TDT方法包括隐马尔柯夫模型(Hidden Markov Model，HMM)话题预测跟踪模型，潜在狄利克雷分布(Latent Dirichlet Allocation，LDA)话题发现模型等，这类方法以广播、新闻专线等专业媒体机构产生的长文本信息流作为研究对象，采用聚类算法将语言形式的信息流分割为不同的新闻章节，能够获得较为理想的结果[3]。但是微博数据是由普通用户产生，存在格式杂乱、用词随意、内容碎片化、数据噪声较大等问题，传统TDT方法难以取得满意的检测结果[4]。

针对上述微博热点需求，国内外学者开展了大量研究，并取得了一定的成果。文献[5]以推特上与地震相关的帖子为研究对象，构建时空概率模型并基于贝叶斯理论建立预测算法，实现了对地震发生时间和地理位置的监控和提前预测；文献[6]提出一种Topic Rank热点话题检测与热度预测模型，基于时间片内关键词的分布频率和相似性定义了话题影响力指标，并利用机器学习算法实时预测话题的未来发展趋势；文献[7]提出一种基于part–of–speech和HowNet的分类聚类方法，特别适合于对短文本进行处理和分析。然而由于中文和英文在语言形式和表达方式上存在很大差异，上述基于推特的研究成果难以推广应用到中文微博。文献[8]针对微博文本内容碎片化特点，通过上下文语义分析、重复计算等方式提取有意义串，并采用Bisecting K–均值算法对其聚类，最后根据热度值指标获得当前热点话题；文献[9]将物理学中能量和加速度的概念引入热点话题发现领域，对微博文本关键字赋予能量值并采用双条件概率模型进行聚类，在新浪微博平台数据上获得了较高的热点话题发现概率；文献[10]提出一种基于隐含语义分析(Latent Semantic Analysis，LSA)和K–均值的混合模型对微博热点话题进行检测，相对于单一K–均值模型，所提方法在准确率、漏检率、错检率和F值等多个维度均可以获得更优的性能。

在上述研究的基础上，本文针对提升微博热点话题发现的准确率和实时性需求开展研究，提出一种基于相关向量机(RVM)优化VSM的微博文本向量化方法，利用RVM的自适应特征选择能力对VSM特征向量进行降维，同时针对K–均值聚类方法聚类准确率受初始聚类中心设置影响大的问题，利用主成分分析(PCA)自动获得K个正交主分量作为K–均值的初始聚类中心，提升算法的准确率。最后根据微博转发、评论和高影响力用户数量定义热度指数，选取热度指数最大的话题作为当前热点话题。

1　基于RVM优化VSM的文本向量化方法

由于聚类算法无法直接对微博文本信息进行处理，因此需要将文本信息进行数学抽象，VSM是当前应用最为广泛的一种文本表示方法，它将文本内容映射到多维向量空间，并利用欧式距离或余弦法则等空间距离描述文本内容之间的语义相似度，从而将文本内容转化为便于计算机识别处理的数学形式。对于某一特定文档T，可以用其包含的所有文档特征项t表示，即$T=\{t_1,t_2,\cdots,t_n\}$，n为文档所处向量空间的维度。VSM采用特征项权重$w_1,w_2,\cdots,w_n$组成的文档空间向量$\boldsymbol{T}=[w_1,w_2,\cdots,w_n]$实现对文档的向量化，则$M$篇文档构成的数据矩阵可以表示为：

$$\begin{cases}\boldsymbol{D}=[\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_M]^{\mathrm{T}}=\begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1n}\\ w_{21} & w_{22} & \cdots & w_{2n}\\ \vdots & \vdots & \cdots & \vdots\\ w_{M1} & w_{M2} & \cdots & w_{Mn}\end{bmatrix}\\ w_{ij}=\dfrac{tf_{ij}\times\log\left(M/n_i+0.01\right)}{\sqrt{\sum\limits_{j=1}^{M}\left(tf_{ij}\times\log\left(M/n_i+0.01\right)\right)^2}}\end{cases}\tag{1}$$

式中：tf_{ij}为第j份文档中第i个词出现的次数，$j=1,2,\cdots,M$；n_i为文档中含有词条i的文本数量。

微博文本具有典型的多样性、碎片化和大数据量特征，经过VSM向量化后每条文本又由数十维权值特征构成，这些高维特征向量中不可避免地存在一些冗余特征，同时中文表达中特有的同义、多义特点也会引起语义模糊，上述问题都会增加后续聚类算法的复杂度和运算量，因此本文采用RVM对VSM得到的特征向量进行特征选择和降维，提升后续聚类算法的实时性和正确率。

对于VSM得到的权值特征向量$\boldsymbol{w}_j=[w_{j1},w_{j2},\cdots,w_{jn}]$，$j=1,2,\cdots,M$，利用RVM模型进行特征选择可以表示为[11-12]：

$$\begin{cases} \boldsymbol{y}=\boldsymbol{v}\cdot K\left(\boldsymbol{w},\boldsymbol{w}_j\right)+\boldsymbol{e} \\ v_n\sim N\left(0,\alpha_n^{-1}\boldsymbol{I}\right) \\ \boldsymbol{e}\sim N\left(0,\beta^{-1}\boldsymbol{I}\right) \\ \alpha_n\sim\text{Gamma}\left(a_0,b_0\right) \\ \beta\sim\text{Gamma}\left(c_0,d_0\right) \end{cases} \tag{2}$$

式中：$K\left(\boldsymbol{w},\boldsymbol{w}_j\right)$为高斯核函数，用于描述2个向量$\boldsymbol{w}$和$\boldsymbol{w}_j$在高维空间中的内积运算；$\boldsymbol{v}=\left[v_1,v_2,\cdots,v_n\right]^{\mathrm{T}}$为每个特征向量对应的权值，$v_n$服从均值为0，协方差矩阵为$\alpha_n^{-1}\boldsymbol{I}$的高斯分布，$\boldsymbol{e}$为零均值高斯白噪声。式(2)所示模型中，第一行子式与SVM模型一致，通过引入核函数将原始空间中线性不可分问题投影到高维空间转化为线性可分问题，第二行和第三行子式是RVM区别于SVM之处，RVM对模型中每个参数(v_n和$\boldsymbol{e}$)赋予后验概率分布，从而构建出全概率模型。后验概率引入的目的是实现样本在特征空间中的稀疏性，从而完成特征选择。为了构建完整的贝叶斯概率模型，第四行子式和第五行子式中，RVM进一步假设α_n和β服从伽马分布(高斯分布的共轭先验分布)，a_0,b_0,c_0和d_0为超参数。

目前常用的RVM模型参数后验概率求解算法为变分贝叶斯期望最大(Variational Bayesian Expectation Maximization，VBEM)算法，在算法迭代过程中，发现大部分α_n会逐渐趋近于无穷大，特征向量权值v_n趋近于0，RVM模型认为该权值对应的特征向量对目标分类的贡献较小，该特征向量$\boldsymbol{w}_j$被“关闭”[13]，从而自动实现对VSM权值向量的稀疏化和降维。

2 基于PCA联合K-均值的聚类方法

在实现文本内容的向量化后，需要对相似度较高的话题进行聚类，K-均值算法是当前应用最为广泛的数据挖掘算法之一，由于理论简单，算法计算效率高等优势被大量应用于文本聚类领域，K-均值算法可以总结为以下4个步骤[14-15]：

Step 1：设置聚类个数K，并从数据集$\left\{\boldsymbol{x}_l\right\}_{l=1}^{N}$中随机选取$K$个样本作为初始聚类中心$\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_K$；

Step 2：计算数据集中样本与每个聚类中心的相似度，并将其划分至相似度最高的类别中，相似度的定义如式(3)所示：

$$d\left(\boldsymbol{x}_l,\boldsymbol{u}_k\right)=\sqrt{\left(\boldsymbol{x}_l-\boldsymbol{u}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}_l-\boldsymbol{u}_k\right)} \tag{3}$$

Step 3：根据式(4)计算得到新的K个聚类中心，其中n_k为第k个子类中的样本数；

$$\overline{u}_k=\frac{1}{n_k}\sum_{l=1}^{n_k}\boldsymbol{x}_l \tag{4}$$

Step 4：按K个新聚类中心对样本集进行重新划分，若连续2次得到的划分结果一致，则算法收敛，否则重复Step 2~Step 3。

K-均值算法聚类个数和初始聚类中心的选取对最终聚类结果影响很大，如果聚类个数和初始聚类中心设置不理想，则会导致算法迭代复杂度增加，聚类准确率下降等问题。PCA[16]是一种经典的数据分析方法，通过对隐含在数据中的相关性进行分析，按相关性大小将数据划分为不同的簇，将每个簇内的信息合成为一个主分量，并保证不同簇之间的信息尽量不相关。因此PCA能够自动从数据提取K个主分量，这K个主分量相互正交并且包含了数据中的绝大部分有用信息。因此在对文本数据进行K-均值聚类之前，采用PCA对其进行分析，将得到的前K个主分量作为K-均值算法的初始聚类中心。

假设经过RVM特征选择后得到的微博文本权值特征向量矩阵为$\overline{\boldsymbol{D}}$，其协方差矩阵可以表示为$\boldsymbol{R}=\overline{\boldsymbol{D}}\,\overline{\boldsymbol{D}}^{\mathrm{T}}$，对其进行特征值分解可以得到：

$$\boldsymbol{R}=\boldsymbol{R}_{\mathrm{s}}+\boldsymbol{R}_{\mathrm{e}}=\sum_{k=1}^{K}\lambda_k\boldsymbol{s}_k\boldsymbol{s}_k^{\mathrm{T}}+\sum_{p=K+1}^{M}\lambda_p\boldsymbol{s}_p\boldsymbol{s}_p^{\mathrm{T}} \tag{5}$$

式中：$\boldsymbol{R}_{\mathrm{s}}$为主分量对应的协方差矩阵；$\boldsymbol{R}_{\mathrm{e}}$为噪声分量对应的协方差矩阵；$\boldsymbol{s}_k(k=1,2,\cdots,K)$为主分量；$\lambda_k$为对应的特征值且$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_K$；$\boldsymbol{s}_p(p=K+1,K+2,\cdots,M)$为次分量；$\lambda_p$为对应的特征值且$\lambda_{K+1}\geqslant\lambda_{K+2}\geqslant\cdots\geqslant\lambda_M$。PCA通常选择占总能量90%的特征值个数为主分量个数K，即：

$$K=\arg_{k}\left(\sum_{n=1}^{k}\lambda_k^2\Big/\sum_{n=1}^{M}\lambda_n^2=0.9\right) \tag{6}$$

3　文本预处理与热度排序

3.1 文本预处理

由于直接获取的微博文本格式杂乱，内容碎片化且包含大量与话题本身完全不相关的噪声数据，因此需要通过文本预处理方式对噪声数据进行过滤和清洗，同时由于本文不涉及图片、音视频等微博信息，所以采用如下的数据过滤和清洗准则：

1) 剔除@，#等非文本性质的交互标识符号，这些通常是代表转发、开头结尾和微博昵称等的特征符号，不含有任何语义信息；

2) 剔除受微博长度限制而显示的统一资源定位符(Uniform Resource Locators，URL)连接，以及“来自”、“转自”等与微博平台有关的信息，这些信息与话题本身内容无关；

3) 剔除同一用户短时间内发布的相同内容微博，以及文本分词数量少于3个的微博，前者通常是由于用户网络延迟问题导致用户操作失误引起，后者内容过短，不能包含一个话题的基本信息；

4) 剔除转发、点赞和评论数相加小于5的微博，一般情况下，此类微博表明该博主没有其他用户关注，因此不可能成为热点话题。

经过上述处理后，噪声数据被剔除，有用的微博文本数据大大减少，本文采用中科院ICTCLAS分词系统对上述文本进行分词，并综合4个常用的停用词库共计2 002个词汇作为本文的参考词库进行分析，剔除“但是”、“因此”等无用词汇，从而将有用微博文本处理为可用的短文本形式。

3.2 热度排序

完成聚类后，文本集被划分为K个微博话题簇，每个簇中仍然包含较多的文本关键词，如果要更加精确地获得特定时间范围内的热点话题，需要对每个簇内的关键词进行排序，提取出热度最高的话题。因此需要一个合理的指标来衡量话题的热度。根据经验可知，某个话题要想成为热点话题，首先需要引起有较高影响力的微博用户的关注，这些高影响力用户的转发和评论会迅速得到大量网民关注，从而使微博的转发数和评论数在短时间内上升，因此微博的转发数、评论数和高影响力用户的数量与话题的热度有较强的关联性，对于微博i，本文采用如下Sigmoid函数形式计算其热度值H_i：

$$H_i=\frac{1}{1+\exp\left[-\left(\sqrt{r_i}+c_i+\log\left(f_i+1\right)\right)\right]} \tag{7}$$

式中r_i,c_i和f_i分别表示微博i的转发数、评论数和高影响力用户的粉丝数。采用Sigmoid函数将热度值映射到0和1之间，H_i越接近于1，表示热度越高。

4　实验结果与分析

4.1 算法流程

对本文所提微博热点话题发现算法流程进行总结，流程图如图1所示，其具体实现思路可以描述为：

算法输入：微博文本数据集；

算法输出：热点话题(话题热度值)。

算法具体实现步骤：

Step 1：文本预处理，根据3.1小节介绍的方法将微博文本数据集中包含的无关噪声数据进行剔除；

Step 2：VSM文本向量化，采用VSM方法，根据式(1)计算得到微博文本数据集对应的数据矩阵$\boldsymbol{D}$和权值特征向量$\boldsymbol{w}$，从而将文本数据转化为计算机能够识别和处理的形式；

Step 3：RVM特征选择，利用式(2)所示RVM模型，对Step 2中得到的高维权值特征向量进行自动特征选择和降维，得到低维特征向量，进一步降低运算复杂度，提升算法实时性；

Step 4：PCA确定初始聚类中心，利用式(5)和式(6)对Step 3得到的特征向量进行分析，自动确定K–均值算法的初始聚类中心；

Step 5：K–均值聚类，利用K–均值算法对特征向量集进行聚类，得到K个微博话题簇；

Step 6：话题热度计算，根据式(7)计算得到每个话题的热度H，热度最高的话题即为当前热点话题。

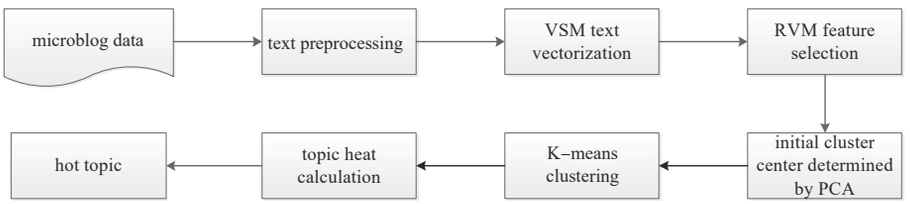

Fig.1 Flow chart of the proposed method
图1 所提热点话题算法流程

4.2 实验数据

为了验证所提方法对于微博热点话题发现的有效性，本文以新浪微博数据为研究对象，利用新浪微博提供的API随机抓取了6 378名粉丝数大于100的用户在2018年5月1日至2018年5月15日共14天发表的所有微博总共44 682条，按照3.1小节所提的文本数据预处理方法对数据进行过滤和清洗预处理后，得到18 455条有效微博作为实验数据集。

由于本文主要研究目的是提升微博热点话题发现的准确率和实时性，因此本文选取准确率、漏检率和误检率3项指标定量评估算法对热点话题发现的准确性，其中准确率指的是算法正确检测出属于某个话题的样本数和实际该话题样本总数的比值；漏检率是指本属于某话题但没有检测到的样本数与实际该话题样本总数的比值；误检率是指本不属于某话题却检测为该话题的样本数与实际该话题样本总数的比值，同时采用算法运算时间和迭代次数两项指标评估算法的实时性。

4.3 话题发现结果与分析

对于预处理后的18 455条试验数据，根据图1所示算法流程，首先利用VSM对其进行文本向量化操作，获得383 962个有效的权值特征，此时如果直接对权值特征进行聚类需要消耗大量的运算资源，并且权值特征中同义词汇大量重复的存在会降低聚类正确率，因此本文采用RVM对权值特征进行分析和特征选择，得到201 166个特征向量，数据量减少了1倍。然后采用PCA对特征向量进行分析，将获得的主分量作为K-均值聚类的初始聚类中心，完成聚类，最后根据式(7)计算得到每个聚类中热度最高的前5个文本作为热点话题。同时为了进行横向对比，采用传统VSM联合K-均值的热点话题发现方法以及文献[14]所提方法在相同条件下开展试验，其中K-均值算法采集经验试错法确定初始聚类中心，取正确率最高的一次实验。

表1给出了所提方法、VSM联合K-均值方法、文献[14]所提方法得到的检测结果和实时性指标，可以看出由于PCA能够自动获取稳定的初始聚类中心，有效克服K-均值对初始聚类中心随机选择的敏感性，因此所提方法相对于VSM联合K-均值方法可以获得更高的准确率、更低的漏检率、误检率。同时所提方法在采用RVM进行特征选择后，大大降低了特征向量的维度，因此所提方法在算法运算时间和迭代次数2项指标上相对于VSM联合K-均值方法也表现出较大优势。文献[14]方法将多条微博之间的相互转发关系量化为转发图矩阵，并将其作为K-均值方法的输入变量，在一定程度上增加了聚类先验信息，能够提升话题发现性能，其准确率、漏检率、误检率与所提方法接近，均明显优于VSM联合K-均值方法，但是转发图矩阵的引入大大增加了算法的运算复杂度。

表1 不同方法性能比较

Table1 Performance comparison of different methods

detection algorithm	accuracy	probability of missing	probability of error detection	operation time/s	number of iterations
proposed method	0.821 3	0.164 3	0.014 4	2 946	7 817
VSM combined K-means	0.748 3	0.192 8	0.058 9	5 337	17 359
reference[14]	0.810 5	0.177 4	0.030 6	6 215	19 031

表2给出了采用本文算法得到的每个热点话题中热度最高的5个文本特征词与新浪微博官方提供的同时段热点话题，可以看出当话题热度较高时，所提方法可以准确有效地发现话题，同时提取的特征词能够很好地反映话题所包含的内容和主题，随着热度的降低，所提方法的准确性出现了一定程度的下降，例如对于“朝韩首脑会晤”话题，所提方法提取了“口误”这个与话题无关的特征词，对于“随手拍景点”话题，所提方法提取了“校长”和“传销”2个与话题不相关的特征词。对数据进行复核发现，当天除了表2所列5个话题外，同时还存在“北大校长口误事件”、“特大传销组织‘云联慧’被摧毁”2项热度稍低的话题，其中“北大校长口误事件”热度值为0.311 6，“特大传销组织‘云联慧’被摧毁”热度值为0.302 4，与“朝韩首脑会晤”和“随手拍景点”的热度接近，给算法带来了一定的混淆。

表2 同时段新浪微博热门话题与所提方法结果对比

Table2 Comparison of hot topics on Sina Weibo and the results of the proposed method

rank	hot topics on Sina Weibo	proposed method	calorific value
1	Labor Day	Labor Day, holiday, travel, congestion, train	0.751 4
2	flight attendant was killed in a didi taxi late at night	flight attendant, online car hailing, female safety, Liu mouhua, didi	0.706 6
3	Tianjin talent introduction	civil servant, Tianjin, Beijing, household registration, house price	0.612 9
4	inter Korean summit	Kim Jong un, post-80s generation, CCTV news, slip of the tongue, photo	0.425 1
5	take photos of scenic spots	Sanya, Xiamen, principal, MLM, Taishan	0.337 0

5　结论

本文利用RVM对VSM提取的高维微博文本特征向量进行分析和降维，有效减小数据量并提升后续聚类效率，同时针对K-均值聚类算法对初始聚类中心敏感的问题，利用PCA自动提取主分量，并将其作为K-均值初始聚类中心从而提升K-均值聚类的准确率。结合转发数、评论数和高影响力用户数定义热度值指标，热度值最大的话题即为当前热点话题。采用新浪微博数据开展实验，结果表明，所提方法相对于传统VSM联合K-均值方法准确率提升7.3%，漏检率和误检率分别提升2.9%和4.5%，实时性提升45%，与文献[14]所提方法相比，本文方法准确率提升1.1%，漏检率提升1.3%，误检率提升1.6%，算法实时性提升53%。

另外也应注意到，本文算法在面对热度不高但是热度值接近的话题时，提取的关键词出现了混淆的现象，这在下一步工作中需要重点研究。同时作为一种实时信息交互方式，除了文本外，微博数据中还包含大量的图片、音视频等多媒体文件，这些文件中同样包含大量的有用信息，因此后续研究需要考虑将文本与上述多媒体文件相结合，进一步提升话题发现的准确率。

参考文献：

[1] 丁兆云,贾焰,周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展, 2014,51(4):691-706. (DING Zhaoyun,JIA Yan,ZHOU Bin. Survey of data mining for microblogs[J]. Journal of Computer Research and Development, 2014,51(4):691-706.)

[2] YANG Y,CARBONELL J,BROWN R,et al. Multi-strategy learning for topic detection and tracking[M]. New York:Springer US, 2002.

[3] 李心妍,刘俐俐. 浅析微博中的“微舆情”[J]. 新闻世界, 2011(7):111-112. (LI Xinyan,LIU Lili. Analysis of "micro public opinion" in micro blog[J]. News World, 2011(7):111-112.)

[4] 薛峰,周亚东,高峰,等. 一种突发性热点话题在线发现与跟踪方法[J]. 西安交通大学学报, 2011,45(12):64-69. (XUE Feng, ZHOU Yadong, GAO Feng, et al. An online detection and tracking method for bursty topics[J]. Journal of Xi'an Jiaotong University, 2011,45(12):64-69.)

[5] SAKAKI T, OKAZAKI M, MATSUO Y. Earthquake shakes Twitter users: real-time event detection by social sensors[C]// Proceedings of the 19th International Conference on World Wide Web. ACM:[s.n.], 2010:851—860.

[6] 杨冠超. 微博客:热点话题发现策略研究[D]. 杭州,浙江:浙江大学, 2011. (YANG Guanchao. Research of hot topic discovery strategy on microblogging plotforms[D]. Hangzhou,Zhejiang,China:Zhejiang University, 2011.)

[7] LIU Z, YU W, CHEN W. Short text feature selection and classification for microblog mining[C]// Proceedings of International Conference on Computational Intelligence and Software Engineering. Wuhan,China:IEEE, 2010:1-4.

[8] 贺敏,王丽宏,杜攀,等. 基于有意义串聚类的微博热度话题发现方法[J]. 通信学报, 2013(S1):256-262. (HE Min,WANG Lihong,DU Pan,et al. Microblog hot topic detection method based on meaningful string clustering[J]. Journal on Communications, 2013(S1):256-262.)

[9] 林思娟,林柏钢,许为,等. 一种基于词语能量值变化的微博热点话题发现方法研究[J]. 信息网络安全, 2015(10):46-52. (LIN Sijuan,LIN Bogang,XU Wei,et al. Research on microblog hot topic detection method based on term energy change[J]. Netinfo Security, 2015(10):46-52.)

[10] 马雯雯,魏文晗,邓一贵. 基于隐含语义分析的微博话题发现方法[J]. 计算机工程与应用, 2014,50(1):96-100. (MA Wenwen, WEI Wenhan,DENG Yigui. Micro-blog topic detection method based on latent semantic analysis[J]. Computer Engineering and Applications, 2014,50(1):96-100.)

[11] BISHOP C M. Pattern recognition and machine learning[M]. New York:Springer, 2006.

[12] 王宝帅. 基于微多普勒效应的空中飞机目标分类研究[D]. 西安:西安电子科技大学, 2015. (WANG Baoshuai. Study on classification of airplane target based on micro-Doppler effect[D]. Xi'an,China:Xidian University, 2015.)

（下转第391页）